Journal of Controlled Release 378 (2025) 619–636

Contents lists available at ScienceDirect

Journal of Controlled Release

journal homepage: www.elsevier.com/locate/jconrel

AI-directed formulation strategy design initiates rational drug development

Nannan Wang [a], Jie Dong [b,**], Defang Ouyang [a,c,*]

[a] State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences (ICMS), University of Macau, Macau, China
[b] Xiangya School of Pharmaceutical Sciences, Central South University, Changsha, China
[c] Department of Public Health and Medicinal Administration, Faculty of Health Sciences (FHS), University of Macau, Macau, China

ARTICLE INFO

Keywords:
Formulation strategy design
Artificial intelligence
Rational drug development
Poorly water-soluble drugs
Positive-unlabeled learning
Partially supervised learning

ABSTRACT

Rational drug development would be impossible without selecting the appropriate formulation route. However, pharmaceutical scientists often rely on limited personal experiences to perform trial-and-error tests on diverse formulation strategies. Such an inefficient screening manner not only wastes research investments but also threatens the safety of clinical volunteers and patients. A design-oriented paradigm for formulation strategy determination is urgently needed to initiate rational drug development. Herein, we introduce FormulationDT, the first data-driven and knowledge-guided artificial intelligence (AI) platform for rational formulation strategy design. Learning from approved drug formulations, FormulationDT devised a comprehensive formulation strategy design system containing 12 decisions for both oral and injectable administration. Utilizing PU-Decide, our specialized partially supervised learning framework designed for positive-unlabeled (PU) scenarios, FormulationDT developed precise and interpretable classification models for each decision, achieving area under the receiver operating characteristic curve (ROC_AUC) scores ranging from 0.78 to 0.98, with an average above 0.90. Incorporating extensive domain knowledge, FormulationDT is now accessible through a user-friendly web platform (http://formulationdt.computpharm.org/). Moreover, FormulationDT demonstrates its value by showcasing its application in proteolysis targeting chimeras (PROTACs) and recent drug approvals. Overall, this study created the first approved drug formulation dataset and tailored the PU-Decide framework to develop a high-performance, interpretable, and user-friendly AI formulation strategy design platform, which holds promise for driving risk reduction and efficiency gains across the life cycle of drug discovery and development.

1. Introduction

Drug development remain endeavors characterized by high investment and substantial risks. The attrition rate of drug candidates entering Phase I clinical trial reaches as high as 90 % [1], which underscores the inadequacy in effectively predicting therapeutic and toxic responses at the preclinical stages. For one thing, the overemphasis of computer-aided drug design (CADD) and high-throughput drug screening on potent compounds leads to the "high affinity trap" [2], the insufficient attention given to developability has become a hurdle in the pathway to drug approval. For another thing, as low-hanging fruits are competitively pursued, the increasing complexity of drug candidates presents escalating formulation challenges [3]. This necessitates a higher level of involvement from formulation scientists in the upstream stages of drug development. In response, some Big Pharma have formed developability teams [4,5], where formulation scientists play the critical role in

Abbreviations: 2D, Two-dimensional; ADME, Absorption, distribution, metabolism, and elimination; AI, Artificial intelligence; AIDD, AI-driven drug discovery; APIs, Active pharmaceutical ingredients; BCS, Biopharmaceutics classification system; CADD, Computer-aided drug design; DCS, Developability classification system; DT, Decision Tree; FDA, Food and Drug Administration; FFP, Fit-for-purpose; IND, Investigational New Drug; KNN, k-Nearest Neighbors; LR, Logistic Regression; MCC, Matthews Correlation Coefficient; nBayes, Naive Bayes; NDA, New Drug Application; NMEs, New molecular entities; NN, Neural Networks; PROTACs, Proteolysis targeting chimeras; PU bagging, Positive-unlabeled bagging; PU learning, Positive-unlabeled learning; QED, Quantitative estimate of drug-likeness; R&D, Research and development; RF, Random Forest; ROC_AUC, Area under the receiver operating characteristic curve; SHAP, Shapley Additive exPlanations; SMILES, Simplified Molecular Input Line Entry System; SMOTE, Synthetic Minority Oversampling Technique; SVM, Support Vector Machine.

* Corresponding author at: State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences (ICMS), University of Macau, Macau, China.

** Corresponding author.

E-mail addresses: jiedong@csu.edu.cn (J. Dong), defangouyang@um.edu.mo (D. Ouyang).

https://doi.org/10.1016/j.jconrel.2024.12.043
Received 11 September 2024; Received in revised form 27 November 2024; Accepted 18 December 2024
Available online 26 December 2024

developability screening, non-clinical and commercial formulation design, and expediting clinical trial entry (Fig. 1).

For small molecule drugs, solubility crisis stands out as the foremost developability and formulation challenge [6]. Formulation scientists have devised non-conventional strategies to formulate poorly water-soluble molecules. These strategies are based on different solubilization principles, including scaling down the solute-solvent difference in solubility parameters [7], decreasing the lattice energy of the solute [8], and maintaining supersaturation of the drug (inhibiting precipitation rate) [8,9]. In contrast, conventional formulations refer to approaches without a specific solubilization purpose. These include widely used techniques to adjust dissolution or enhance manufacturability, such as the addition of wetting agents, micronization of active pharmaceutical ingredients (APIs), or conversion to salt forms [10]. Formulation strategies based on different principles are suitable for molecules with different structures. Unsuitable formulation strategies will result in wasted research investments, and even pose serious threats to the health of clinical volunteers or patients [11]. Therefore, as the early stage of drug development, rational developability evaluation and formulation strategy design, instead of trial-and-error tests, would be the crucial initiatives for initiating rational drug development.

By summarizing experience, researchers have compiled rules [12,13], classification systems [14–16], or expert systems [17–20] for formulation strategy decision-making for poorly water-soluble drugs. For example, following the Biopharmaceutics Classification System (BCS) [21], several extended applications and modifications based on the BCS concept have been suggested to help druggability assessment or formulation strategy decisions [14–16]. Drawing from 76 in-house development cases, Branchu et al. [10] explored tools like statistics, decision trees, and case-based reasoning to aid in formulation strategy selection. While the study presents valuable insights, there is potential for enhancement in data quantity and quality, modeling methods, and practical applicability. The aforementioned qualitative or semi-quantitative expert systematic studies on formulation strategy decision-making are the accumulation of valuable formulation development experience; however, the bias from individual developer's experience cannot be ignored. Another limitation is that such empirical decision-making schemes often require in-depth investigation on drug properties (gastrointestinal solubility, intestinal absorption characteristics, etc.), which heavily limits the applicability, especially in drug discovery stages.

Considering the limitations mentioned above and motivated by the scientific rationale that "structure determines nature and influences decision-making", we propose the machine learning solution for formulation strategy design, aiming to discern the correlation between the structure and the appropriate formulation routes. Machine learning research relies primarily on high-quality datasets. Compared to data derived from the literature or cases in the research and development (R&D) pipeline, marketed drug data are considered more convincing because they have been verified by clinical trials, drug regulatory agencies, and the markets [22]. Moreover, marketed drugs bring together the wisdom of outstanding pharmaceutical scientists across the globe, rather than the limited individual experience. As such, what lessons can we learn from approved drugs? By analyzing the available marketed drug information and incorporating our objectives, three points are concluded in Fig. 2. First, the necessity of solubilization can be learnt by comparing the conventional formulations with the non-conventional (solubilization) formulations, which is one of the criterions for the lead compound developability assessment and will be the first step in formulation strategy design [10]. Further, the feasibility of salt formation and specific non-conventional formulation strategies can be learned by generalizing the structural features of corresponding subsets of drugs. Based on these lessons, multiple classification models can be constructed to uncover each decision patterns within marketed drug data, then following by the established of an AI decision-making system.

A common issue with approved drug data is lacking reliable negative samples, which constitute positive-unlabeled tasks [23]. For instance, while we can infer that a drug marketed as a salt form should be suitable for salt formation, not all drugs marketed as prototypes are incapable of salt formation, highlighting the absence of reliable negative samples. Similarly, we cannot assume that a drug not employing a specific non-conventional strategy is technically unsuitable. For cost reasons, drug products using a non-conventional strategy can be considered as reliable positive samples, indicating a necessity for solubilization. However, new molecular entities (NMEs) are preferentially developed as conventional formulations to expedite time-to-market or prolong product lifecycle [24]. In other words, there remains potential to enhance certain conventional formulations with more complex formulation strategies. In data-driven formulation strategy decision-making, consideration of the confidence of negative samples is indispensable, i.e., what constitutes an "unsuitable formulation strategy", which aligns with Kuentz et al. [25] in their commentary on the rational selection of bio-enabling oral formulations. Indeed, positive-unlabeled problems of this nature are prevalent because, in numerous practical scenarios, acquiring reliable negative data is challenging or costly, or the definition of negative data remains ambiguous [26]. Positive-unlabeled (PU) learning, a class of partially supervised machine learning methodologies, has been devised to address scenarios where only reliable positive data and unlabeled data (or uncertain negative data) are available [23]. PU learning has been successfully applied to several biomedical tasks [27–30]. However, due to labelling deficiencies, most PU learning methods have the problems of lacking reliable validation and failing to make instantaneous predictions.

With the above motivations, the present study developed the first AI formulation strategy design platform integrating PU learning with domain knowledge. The main contributions are as follows. (1) The first AI formulation strategy design system was schemed by learning from the first collection of approved drug formulation data. (2) To achieve robust instantaneous prediction, the partially supervised learning framework PU-Decide was developed for PU learning tasks. (3) The AI formulation strategy design system was constructed through PU-Decide framework,

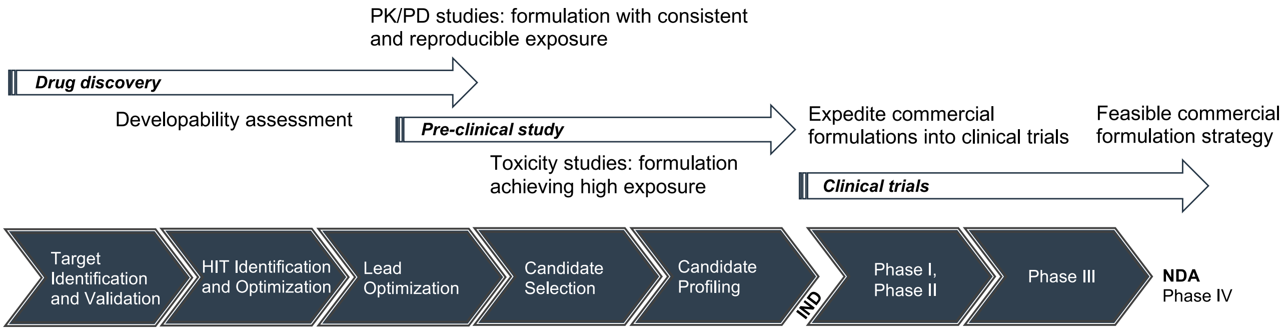

PK/PD studies: formulation with consistent and reproducible exposure
Drug discovery
Developability assessment
Pre-clinical study
Expedite commercial formulations into clinical trials
Feasible commercial formulation strategy
Toxicity studies: formulation achieving high exposure
Clinical trials
Target Identification and Validation
HIT Identification and Optimization
Lead Optimization
Candidate Selection
Candidate Profiling
IND
Phase I, Phase II
Phase III
NDA
Phase IV

Fig. 1. The typical drug development procedure and the key roles played by formulation scientists.

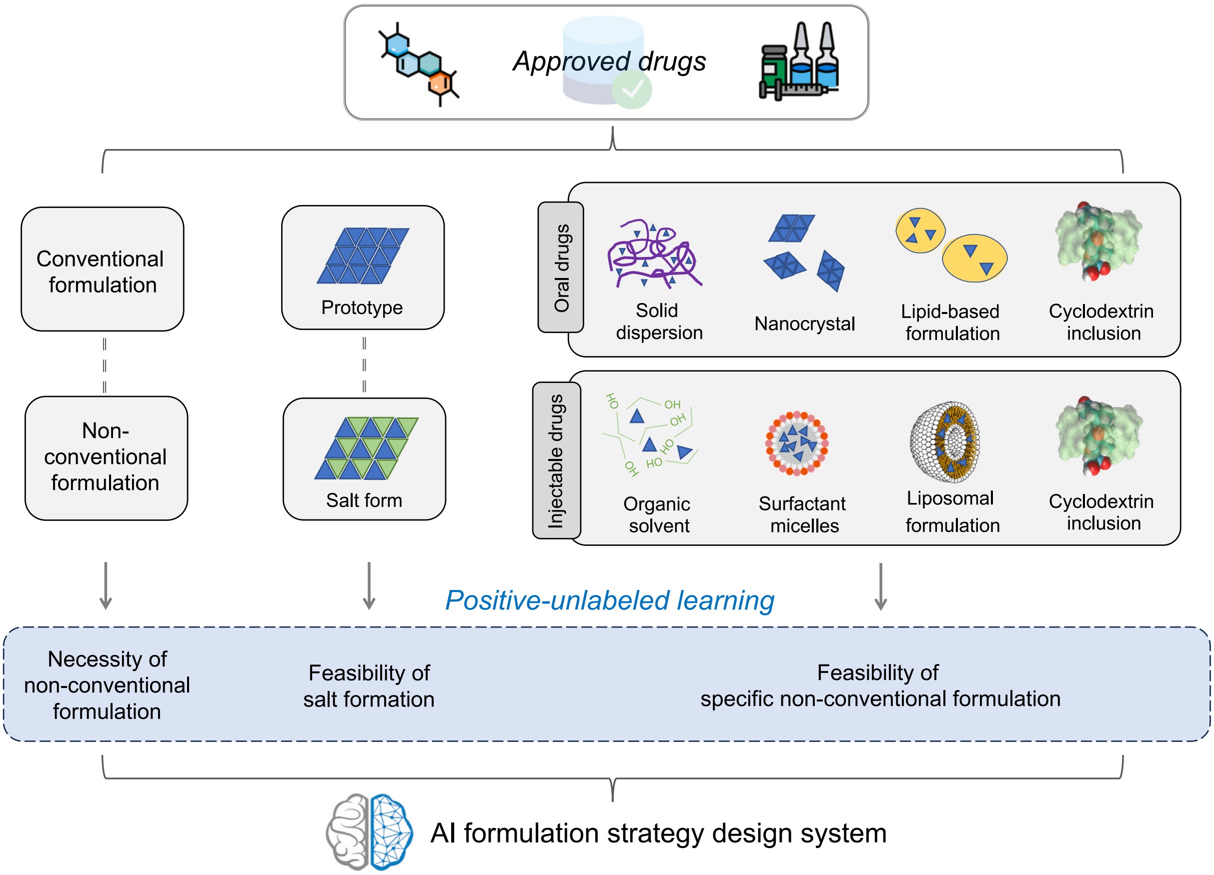


Fig. 2. Three types of lessons learned from approved drugs: the necessity of non-conventional formulation, the feasibility of salt formation, and the feasibility of the specific non-conventional formulation. For oral drugs, marketed non-conventional formulations include solid dispersions, nanocrystals, lipid-based formulations, and cyclodextrin inclusions. Common solubilization strategies applied to marketed injectable drugs include using organic solvents/co-solvents to adjust the solvent's solvation-related properties, dispersing the drug in the hydrophobic regions of surfactant micelles or liposomes, and the formation of cyclodextrin inclusions.

with total 12 decision tasks achieving an average area under the receiver operating characteristic curve above 90 %. (4) Interpretable machine learning models for formulation strategy decisions provide new insights for pharmaceutical research. (5) The AI system was deployed into a user-friendly website to benefit pharmaceutical scientists. Moreover, the established platform was applied on proteolysis targeting chimeras (PROTACs) and recently drug approvals, showcasing its potential in enhancing efficiency, reducing costs, and elevating drug quality across the life cycle of drug discovery and development.

2. Method

2.1. Data preparation

Initially, marketed non-conventional formulations were compiled comprehensively and accurately from literature sources and public reports. Subsequently, drug molecules approved for administration via oral and injectable routes were summarized from the Orange Book database of the U.S. FDA. The non-conventional formulations were excluded from the approved drug dataset to form the conventional formulation data. Additionally, information regarding whether conventional formulations were marketed as prototypes or salts was identified through the Orange Book.

To enable end-to-end prediction, calculated RDKit descriptors [31] and predicted dissociation constants (pKa) [32] are employed to represent drug molecules. The RDKit program was employed to calculate the descriptors for characterizing drug molecule. To facilitate broader model applicability and enable end-to-end decision-making, we employed the methodology proposed by Pan et al. [32] to predict the pKa values of input molecules. The predicted pKa data is encoded into four features that describe whether the molecule possesses acidic or basic ionizable groups, as well as the dissociation constants of the strongest acidic and basic ionizable groups in the molecule.

2.2. Feature engineering

For each dataset, feature engineering was conducted to minimize redundant features and enhance model accuracy and efficiency. Given the absence of reliable negative labels in the original dataset, feature selection methods based on model performance are not applicable. So, filtering methods are employed. Specifically, the following steps were taken for feature engineering:

(1) Calculating the Pearson's correlation coefficient between features and retain only one of the features if its correlation coefficient exceeds 0.8.
(2) Eliminating features that do not show significant differences across target categories using the rank-sum test (Mann-Whitney *U* test) for binary targets, with a significance level of $\alpha = 0.01$.
(3) Standardizing the data using the StandardScaler method from Scikit-learn library [33]. This standardization aimed to ensure that data for different features had the same scale (following a standard normal distribution with a mean of 0 and a standard deviation of 1). This helps mitigate discrepancies in the scales of various features, improves model convergence, and enhances model stability, particularly for algorithms sensitive to feature scaling.

2.3. The structure of PU-Decide framework

To handle the positive-unlabeled data and provide robust online prediction, we designed a PU learning farmwork named PU-Decide on the basis of the bootstrap aggregating scheme proposed by Mordelet et al. [34]. PU-Decide consists of three components (Fig. 3). The first and core component PU bagging creates multiple sub-classifiers through bootstrap sampling to minimize the uncertainty introduced by unlabeled samples. As depicted in Fig. 3a, the PU bagging component includes the following 4 steps:

(1) Randomly selecting a proportion of unlabeled samples as negative samples, forming a balanced training subset with positive samples, and constructing a sub-classifier.
(2) Applying the constructed sub-classifier to each unlabeled sample out of the bootstrap samples and recording the probability of being classified as a positive sample.
(3) Repeating the above two steps, with the mean value of the probabilities obtained in each iteration serving as the final positive score.
(4) Labeling unlabeled samples according to a positive score threshold.

To broaden adaptability across diverse data structures, for the base classifier, four learning algorithms (decision tree, support vector machine, k-nearest neighbors, and logistic regression) employing distinct hypothesis functions will be compared.

To tackle the challenge of validating PU bagging caused by the absence of true labels, we devised the second component for base classifier selection and labeling threshold determination to ensure the reliability of label recovery. As illustrated in Fig. 3b, in each bagging round, several positive samples are randomly selected into the unlabeled data as validation points. After PU bagging, each validation point gets an average positive score. Moving the threshold from 1 to 0, the corresponding recall of the validation points can be obtained. Additionally, at each threshold, the proportion of unlabeled data marked as positive can be determined. To balance the true positive and false positive samples in labeling unlabeled samples, we define the RP value:

$$RP\ value = Recall\ of\ validation\ points - Positive\ rate\ of\ unlabeled\ data \tag{1}$$

Collectively, comparing the magnitude of the maximum RP values of various base classifiers can be utilized to select the best base classifier, and by identifying the threshold where the maximum RP value occurs, the optimal threshold can be determined. The detailed explanation was provided in **Supplementary material 1**.

Lastly, to provide robust and online prediction, binary classification models are trained for each task. Eight machine learning algorithms based on different hypothesis functions are employed for constructing classification models, including Decision Tree (DT), Random Forest (RF), k-Nearest Neighbors (KNN), Naive Bayes (nBayes), Light Gradient Boosting Machine (LightGBM), Logistic Regression (LR), Support Vector Machine (SVM), and Neural Networks (NN). These learning algorithms cover a spectrum of learning paradigms, including rule-based learning, ensemble learning, instance-based learning, probability models, linear models, kernel-based methods, and artificial neural networks.

2.4. Machine learning model development and evaluation

We used Scikit-learn library [33] in Python to build machine learning models, including DT, RF, KNN, nBayes, LightGBM, LR, SVM, and NN (Multilayer perceptron from Scikit-learn). To address data imbalance, the Synthetic Minority Oversampling Technique (SMOTE) [35] was applied in training set. Bayesian optimization was employed to explore the hyperparameter space [36], seeking the optimal hyperparameter combinations for each algorithm. In each Bayesian optimization, the initial number of sampling points is set to 10, and the total number of optimization iterations is set to 200. For each task, the 5-fold splitting strategy was used to acquire five non-overlapping subsets, each serving as a test subset for 5 independent model training, validation, and testing processes. Models were evaluated with multiple metrics, including accuracy, precision, recall, F1 score, and MCC as implemented in Scikit-learn. The mean and standard deviation of the models' performance on the test subsets are recorded for the evaluation and comparison of model generalization capability. The SHAP library [37] was used to interpret the features influencing each model's decisions. The rank-sum test (Mann-Whitney *U* test) was conducted with the "scipy.stats" module in Python.

2.5. The deployment of the artificial intelligence platform

FormulationDT is designed to offer a robust computing environment for real-time calculations to multiple users. The underlying hardware of the platform utilizes Alibaba Cloud's Elastic Compute Service, with Ubuntu as the server operating system. The platform is developed using the Python programming language, making full use of its mature community and rich AI and data processing libraries, such as Numpy [38], Pandas [39], and Scikit-learn [33]. For handling web requests and resource access, we employ Nginx (https://www.nginx.com/) to proxy requests and uWSGI (https://pypi.org/project/uWSGI/) to communicate between Nginx and the Python program. Built on the Django (https://www.djangoproject.com/) framework, the platform ensures a clear separation between business data (models) and user interface (views), facilitating convenient upgrades and maintenance. MySQL (https://www.mysql.com/) is employed as the widely used relational database engine for data storage. In terms of the user interface, we adopted asynchronous JavaScript (https://www.javascript.com/) and XML (AJAX) for asynchronous data retrieval, and leveraged CSS (cascading style sheets) and JavaScript to construct a cross-platform user interface. This technological combination enables FormulationDT to efficiently store and format models within the established framework, providing outstanding predictive services. The overall platform development aims to ensure high-performance computing, laying a solid foundation for future resource-based application programming interfaces.

3. Result

3.1. Approved drug data and AI formulation strategy decision system design

To learn from approved drugs, a formulation strategy dataset comprising 988 orally administered drugs and 448 injectable drugs approved by the U.S. FDA as of 2022 was compiled. The data distribution is depicted in Fig. 4b. In the conventional formulation category, oral prototype drugs constitute a slight majority (55.2 %), whereas in injectable drugs, 62.3 % are in salt form. The frequency of non-conventional techniques in oral (13.8 %) and injectable drugs (14.7 %) is close. Among the non-conventional formulations, oral drugs are predominantly formulated using lipid-based formulations or solid dispersion techniques. The more sophisticated nanocrystal techniques were used in fewer than 20 marketed products. There are 30 oral cyclodextrin inclusions marketed, nearly twice as many as those administered by injection. For injectable drugs, the number of marketed products for four non-conventional technologies ranged from 14 to 27. There are drugs that overlap in terms of non-conventional technologies. For instance, the hypolipidemic drug Fenofibrate has been successfully developed as a solid dispersion (*Lipanthyl Supra, Lipidil EZ*), nanocrystal (*Tricor, Lipidil Micro*), and lipid-based formulation (*Triglide, Antara*). The overlap information is indicated in the Venn plot in Fig. 4b.

Gleaning insights from approved drugs and incorporating the expertise of formulation scientists, the AI formulation strategy decision

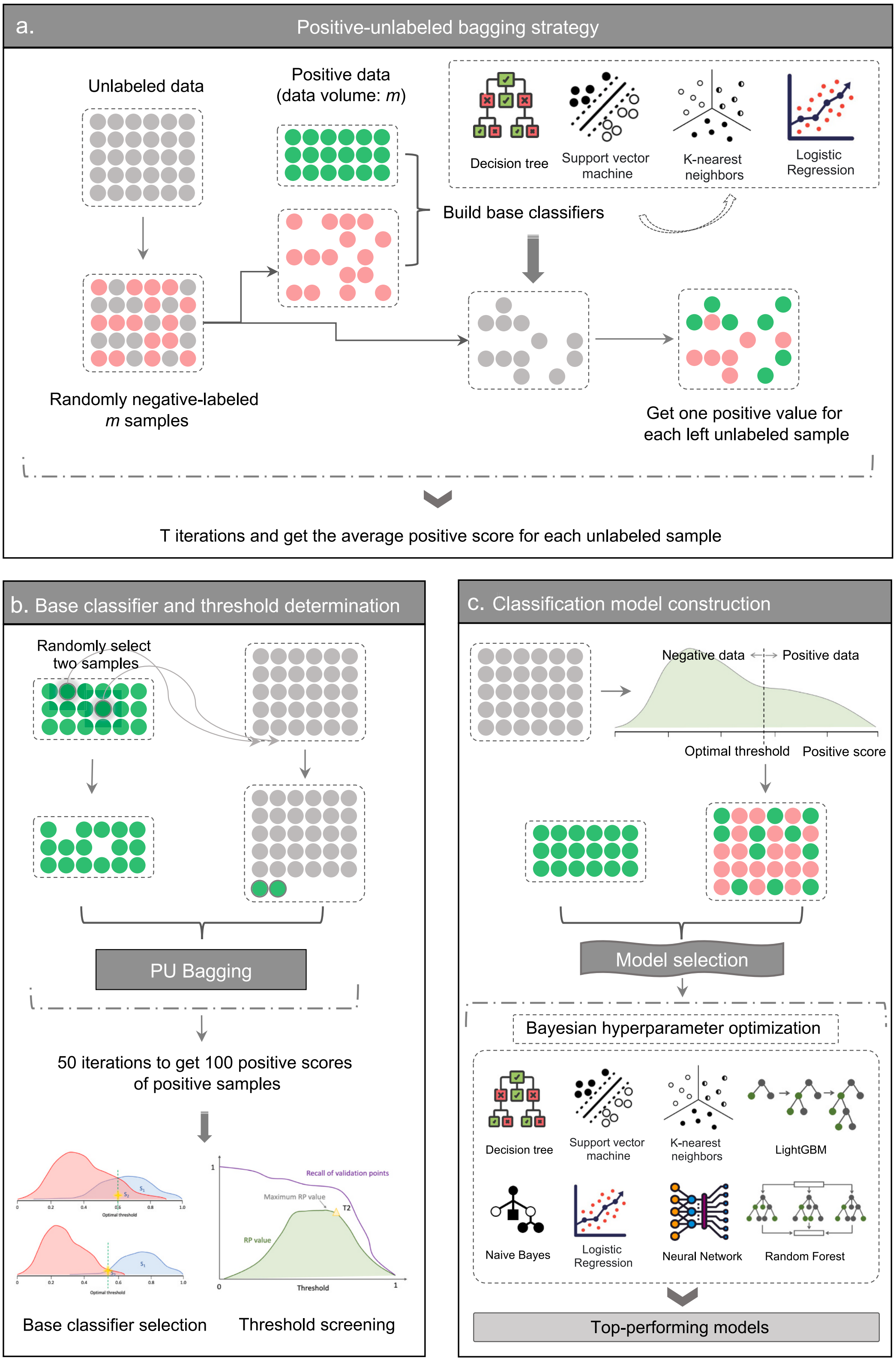


Fig. 3. The PU-Decide framework. The gray, green, and pink dots represent unlabeled, positive, and negative samples, respectively. **a.** Illustration of the PU bagging method. In this study, the number of iterations was set to 1000 based on preliminary experimental results. **b.** PU learning validation method for base classifier selection and threshold screening. Compare the maximum RP values of different base classifiers to select the best one. Identify the threshold corresponding to the maximum RP value allows determination of the optimal threshold. Fewer randomly selected positive samples per iteration is preferable. In this study, the number of randomly selected positive samples was set to 2 to ensure that even tasks with minimal data could provide 100 unique sample selections. **c.** Procedures of classification model construction on the label-recovered data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

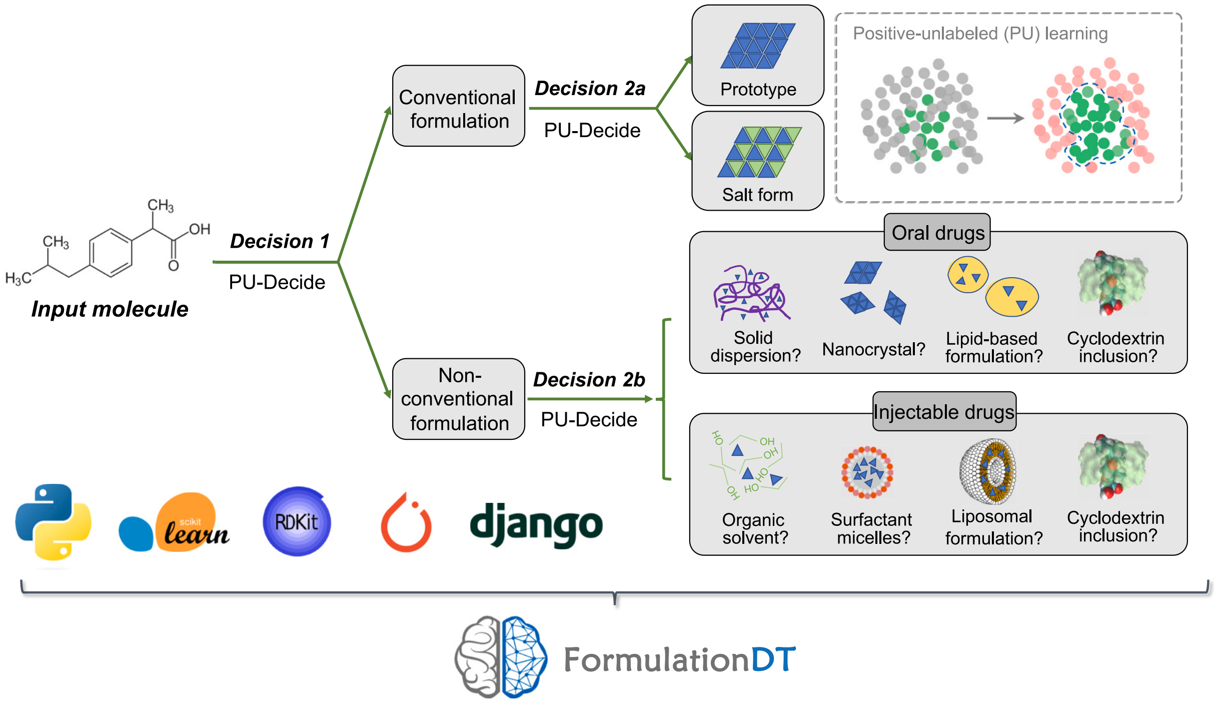


Fig. 4. Data profile. a. Data processing and flow for positive-unlabeled bagging and machine learning model evaluation; **b.** Formulation strategy distribution for marketed small molecule drugs as of 2022. The pie chart shows the proportion of oral versus injectable drugs. The donut chart illustrates the ratio of conventional formulations to non-conventional formulations, as well as the proportion of prototype drugs to salt form drugs within conventional formulations. The Venn diagram displays the overlap among specific non-conventional formulations. The bar chart presents the number of approved specific non-conventional formulation products.

system was designed as Fig. 5. Decision 1 pertains to the formulatability decision about whether the molecule necessitates adopting a non-conventional solubilization strategy. Utilizing Decision 1, targeted screening or designing lead compounds with high formulatability in drug discovery process will reduce R&D and production expenses. Simultaneously, accurately identifying the solubility challenges and promptly implementing solubilization strategies are crucial for mitigating drug development risks [40]. In conventional formulations, converting a prototype drug into a salt form is a widely employed and relatively low-cost chemical modification strategy. Salt-formation enhances the polarity of drug molecules and improves their interaction with the polar solvent, thereby increasing solubility. However, not all molecules can undergo salt formation, depending on factors such as ionizable groups, crystallizability, and stability of the target salts [41]. Assessing the feasibility of salt formation would be the subsequent practical decision (Decision 2a). Non-conventional formulation strategies, grounded in various solubilization principles, possess distinct advantages and disadvantages, with their applicability to different types of drug candidates varying accordingly. Hence, implementing Decision 2b will facilitate the recommendation of feasible non-conventional formulations, thereby enhancing the efficiency of both non-clinical and clinical formulation development. Further, the whole formulation strategy decision system can be decomposed into 12 machine learning tasks as shown in Table 1.

3.2. Processing positive-unlabeled datasets with PU-Decide

To select the base classifiers for each task, we first compared the maximum RP values obtained from PU bagging with 4 learning algorithms, DT, SVM, KNN, and LR, for the 12 tasks (Fig. 6a). Overall, the maximum RP values for all tasks lie between 0.35 and 0.68. For different tasks, three algorithms, except KNN, exhibit their respective strengths. DT and SVM have advantages in handling non-linear relationships in high-dimensional spaces, whereas LR is often suitable for linearly divisible or approximately linearly divisible problems. The optimal performance of DT, SVM, and LR in four, three, and five tasks, respectively, underscores the varied data characteristics and correlation patterns within different tasks. This emphasizes careful selections of base classifiers. KNN typically excels in local pattern recognition but may struggle with global generalization in high-dimensional spaces. Additionally, KNN is more sensitive to the choice of the hyperparameter k value [42], and as a base classifier without a tuning process, this characteristic somewhat limits its application.

To determine the optimal threshold, the recall curves of the validation points and the RP values for each task across different thresholds (the screening step is 0.01) were depicted (Fig. 6b), employing the optimal base classifiers. As the threshold transitions from 1 to 0, the RP values exhibit a trend of increasing and then decreasing, aligning well with the theoretical inference (**Fig. S1**). By locating the maximum RP value, the optimal thresholds for each task were determined. As shown in Fig. 6c, the optimal thresholds are distributed from 0.46 to 0.73. The threshold screening compensates for the lack of reliable negative data and inadequate hyperparameter optimization of the base classifiers. At the corresponding optimal thresholds, the recall distribution of the validation points ranges from 0.56 to 0.93 (Fig. 6d), averaging at 0.73. This highlights the effectiveness of the base classifier in identifying positive samples, even under data and training constraints.

The determined base classifiers and thresholds were chosen to perform PU bagging on the 12 datasets, respectively. The status of the original and relabeled datasets for all 12 tasks is shown in Table 2. Taking Task_o1 as an example, 21.36 % (182 of 852) of the unlabeled data were identified as positive samples in PU bagging using 0.56 as the decision threshold. This finding suggests that approximately 20 % of orally administered drugs, traditionally formulated as conventional formulations, offer potential for the development of modified new drugs using bio-enabling technology, thereby enhancing safety and efficacy. In salt formation decisions, the PU bagging results showed that for oral and injectable drugs, 22.13 % and 45.83 % of the prototypes were technically feasible to be developed into salt forms, respectively. Across the eight Decision 2b tasks concerning the feasibility of specific non-

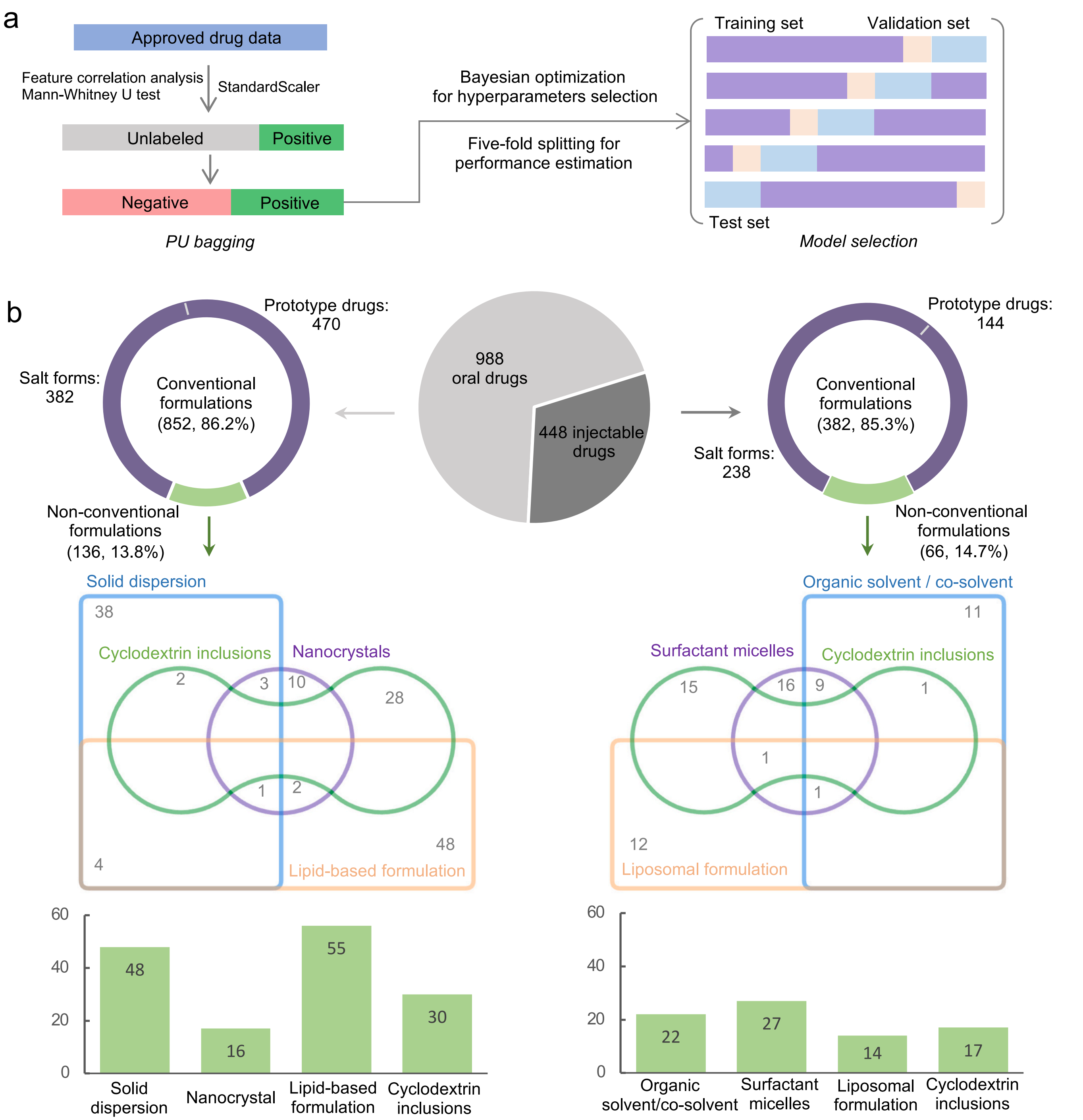


Fig. 5. Overview of the artificial intelligence formulation strategy design system. Decision 1 concerns determining whether the molecule requires a non-conventional solubilization strategy. Decision 2a involves evaluating the feasibility of salt formation. Decision 2b to recommend feasible non-conventional formulations. The positive-unlabeled learning framework, PU-Decide, will be utilized for each decision. Various computational frameworks, tools, and libraries will be utilized to build the user-friendly AI platform FormulationDT.

conventional formulation strategies, the positivity rate for unlabeled samples ranged from 3 % to 29 %. Such findings are promising to provide inspiration for the development of modified new drugs.

3.3. Model construction and evaluation

To more effectively assess machine learning model generalization ability, for each label-recovered dataset, a 5-fold data splitting approach was employed. Each of these folds was utilized as a test subset (20 %), while the remaining data was divided into training and validation subsets via random stratified sampling (Fig. 3a). The model generalization ability was evaluated by averaging the model performance obtained from five entirely independent model training, validation, and testing processes. Model performance comparison is shown in **Supplementary material 2.** Overall, SVM, NN, and the tree-based ensemble learning algorithms, RF and LightGBM, exhibit superior performance. Tree-based ensemble learning algorithms possess the capability for nonlinear modeling, resistance to overfitting, and relative robustness to noise and outliers. Additionally, these algorithms can handle imbalanced classes and mixed features, often exhibiting excellent

Table 1
The definition and description of the involving machine learning tasks.

Administration route	Tasks	Task description	Models
Oral	Task_o1	Necessity of being formulated as a non-conventional formulation	Model_o1
	Task_o2a	Feasibility of being formulated as salt form	Model_o2a
	Task_o2bs	Feasibility of being formulated as solid dispersion	Model_o2bs
	Task_o2bn	Feasibility of being formulated as nanocrystal	Model_o2bn
	Task_o2bl	Feasibility of being formulated as lipid-based formulations	Model_o2bl
	Task_o2bc	Feasibility of being formulated as cyclodextrin inclusions	Model_o2bc
Injectable	Task_i1	Necessity of being formulated as non-conventional formulation	Model_i1
	Task_i2a	Feasibility of being formulated as salt form	Model_i2a
	Task_i2bo	Feasibility of being formulated with organic solvent/co-solvent	Model_i2bo
	Task_i2bs	Feasibility of being formulated as surfactant micelles	Model_i2bs
	Task_i2bl	Feasibility of being formulated as liposomal formulation	Model_i2bl
	Task_i2bc	Feasibility of being formulated as cyclodextrin inclusion	Model_i2bc

performance in classification tasks with small datasets. In contrast, DT tends to overfit with limited samples and is susceptible to variations in input data [43]. The nBayes performed the poorest in this study. Naive Bayes algorithms are significantly influenced by prior probabilities and may struggle to accurately estimate the independence between features in the presence of limited data points, potentially leading to incorrect assumptions about the true data distribution [44]. The detailed performances of these models are presented in Table 3. The average performance metrics are shown in Fig. 7a. Overall, the average accuracy, recall, precision, and ROC_AUC of the 12 tasks are 86.4 %, 82.2 %, 85.9 %, and 91.2 %, respectively, which demonstrates a balanced and satisfactory classification performance of the developed models. Looking at the specifics, there is still room for improvement in the performance of feasibility predictions for three specific non-conventional formulations: lipid-based formulation, surfactant micelles, and liposomal formulation. This might be attributed to the fact that the purpose of these strategies is mainly, but not limited to, drug solubilization. For instance, some oral drugs formulated as lipid-based formulations aim to enhance the stability of APIs [45]. Such objective limitations may impact the quality of available data and subsequently influence model performance.

To demonstrate the necessity and effectiveness of employing a PU learning strategy to handle positive-unlabeled data, we conducted ablation experiments. Fig. 7b compares the MCC (Matthews Correlation Coefficient) metrics of classification models constructed using the same modeling process after three different data processing methods. For each task, the PU learning strategy adopted in this study significantly improved classification performance compared to modeling with raw data. Additionally, we randomly designated a certain number of samples (equal to the number of known positive samples) as positive samples and applied the same PU-Decide framework for PU learning on this artificially labeled dataset. The results indicate that such a random dataset does not lead to robust improvements in classification performance after PU learning. This also demonstrates that PU learning relies on the correct definition of the positive-unlabeled problem.

3.4. Model interpretation

To enhance user trust and gain insights from models, statistics and SHAP (Shapley Additive Explanations) analysis [37] were employed for model interpretation. To understand the overall impact of molecule features on formulation strategy design, we conducted rank-sum tests on 12 tasks, identifying top 20 molecule features with the most significant average impact. Subsequently, these key features were clustered according to their correlations, resulting in Fig. 8a, which shows that six main feature categories influence formulation strategy design: molecular complexity, aromaticity, basic ionization state, carboxyl group, lactones, and the proportion of tertiary carbons. For decision-making specific to each task, we performed SHAP analysis, which elucidates how the model generates predictions for each sample by assigning contribution values to individual features. Fig. 8b illustrates the feature importance ranking determining the necessity of non-conventional formulation for oral drugs. Prominent features include molecular ionization state, electronegativity, hydrophobicity, and substructural characteristics. A higher basic dissociation constant (indicating stronger basicity) negatively contributes to the necessity for solubilization, which is logical since basic molecules are more likely to ionize in the acidic or mildly basic gastrointestinal environment, enhancing their interaction with polar solvents and increasing apparent solubility [46]. Conversely, highly ionizable molecules can improve solubility through salt formation, reducing the need for bio-enabling formulations. Fig. 8c depicts the interaction effect of lipophilicity and basic groups on oral molecule solubilization necessity. The model quantified the cutoff value of MolLogP's solubilization necessity contribution as 2.3. For molecules without basic ionizable groups, lipophilicity will impose a relatively larger solubilization necessity contribution. Fig. 8d compares the decision basis for salt formation feasibility of oral and injectable drugs, using a heatmap to present clusters of samples influenced by different feature combinations. For oral drugs, the ionization constant of the basic groups significantly determines the feasibility of salt formation, as it directly influences the ionization state of drugs in the gastrointestinal tract, affecting absorption rate and extent. For injectable drugs, the factors influencing salt formation feasibility are more related to the combination of molecular electrostatic characteristics. Fig. 8e quantifies the impact of molecular mass on the choice between two solubilization strategies. For solid dispersions, a molecular mass above 430 provides a positive contribution. Notably, while solid dispersions are suitable for molecules with relatively large molecular mass, these molecules typically exhibit lower complexity (characterized by the lower HallKierAlpha index [47]), meaning they generally possess fewer ring structures, branches, and complex functional groups. For cyclodextrin inclusion complexes, a relative molecular mass above 420 and poor drug-likeness (characterized by the lower QED index [13]) contribute negatively because of the limited cavity volume of cyclodextrins. Fig. 8f illustrates the quantitative impact of FractionCSP3 (the proportion of tertiary carbons in the carbon skeleton) on the choice between injectable cyclodextrin inclusions and surfactant micelles. Poorly soluble molecules with low FractionCSP3, often termed “brick-dust” [48], can benefit from cyclodextrin inclusion to prevent aggregation and crystallization, while insoluble molecules with high FractionCSP3, referred to as “grease-ball”, can be stabilized within the hydrophobic regions of micelles. Conversely, molecules with high FractionCSP3 may have more three-dimensional structures that hinder stable binding with cyclodextrin cavities [49]. Fig. 8g presents the decision-making process of the solubilization strategy for Fenofibrate, a poorly soluble drug that has been successfully marketed as lipid-based formulations, nanocrystals, and solid dispersions. Model interpretability analysis clearly demonstrates the key molecular features that recommend or dissuade certain formulation strategies. For Fenofibrate, excessive lipophilicity is a primary reason for its unsuitability for cyclodextrin inclusion, as it would lead to overly strong binding with the cyclodextrin cavity, hindering dissociation and absorption at the absorption window. Different tasks reveal distinct molecular features impacting decisions, as shown in **Fig. S2**. Some results align with empirical knowledge, while others provide new insights for formulation scientists. By elucidating molecular features related to drug formulation performance, pharmaceutical

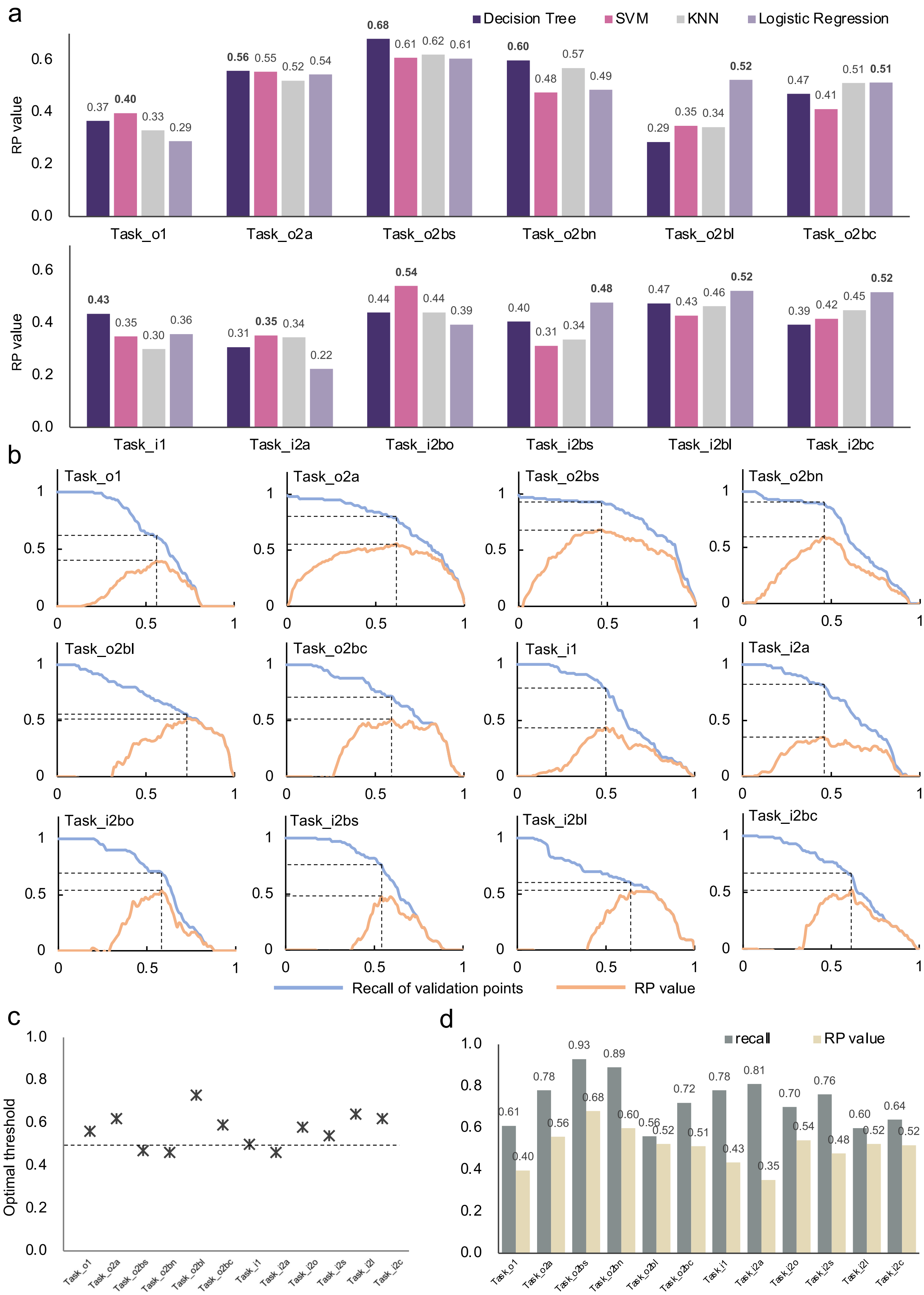


Fig. 6. Positive-unlabeled bagging results. **a.** The maximum RP value of 4 base classifiers for each task. The maximum RP values of the best base classifiers are bolded. **b.** The curve of the recall of validation points (blue lines) and the RP value (orange lines) with the best base classifier for each task. The dashed line indicates the optimal threshold, along with the corresponding recall of validation points and maximum RP values at that threshold. **c.** The distribution of optimal thresholds with the best classifiers. The dashed line indicates the threshold 0.5. **d.** The recall of validation points and the maximum RP value with the best base classifier and optimal threshold for each task. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Overview of the data volume before and after positive-unlabeled bagging for 12 tasks.

Task	Original dataset			Dataset after PU bagging			
	Positive samples	Unlabeled samples	Total samples	Recognized positive samples	Positive rate of unlabeled samples	Positive samples	Negative samples
Task_o1	136	852	988	182	21.36 %	318	670
Task_o2a	382	470	852	104	22.13 %	486	366
Task_o2bs	48	88	136	22	25.00 %	70	66
Task_o2bn	16	120	136	35	28.93 %	51	85
Task_o2bl	55	81	136	3	3.70 %	58	78
Task_o2bc	30	106	136	22	20.75 %	52	84
Task_i1	66	382	448	134	35.08 %	200	248
Task_i2a	238	144	382	66	45.83 %	304	78
Task_i2bo	22	44	66	9	20.45 %	31	35
Task_i2bs	27	39	66	11	28.21 %	38	28
Task_i2bl	14	52	66	4	7.69 %	18	48
Task_i2bc	17	49	66	6	12.24 %	23	43

Table 3
Detailed performance of the optimal models for 12 tasks (mean ± standard deviation, 5 independent tests).

Model	Accuracy	Precision	Recall	ROC_AUC	F1 score	MCC
Model_o1	0.8927 ± 0.0108	0.8429 ± 0.0223	0.8208 ± 0.0447	0.9348 ± 0.0161	0.8309 ± 0.0204	0.7533 ± 0.0261
Model_o2a	0.8686 ± 0.0306	0.9035 ± 0.0426	0.8643 ± 0.0537	0.9432 ± 0.0303	0.8821 ± 0.0286	0.7375 ± 0.0619
Model_o2bs	0.9193 ± 0.0399	0.9846 ± 0.0344	0.8571 ± 0.0714	0.9768 ± 0.0288	0.9150 ± 0.0440	0.8482 ± 0.0740
Model_o2bn	0.9259 ± 0.0454	0.9022 ± 0.0606	0.9000 ± 0.1000	0.9665 ± 0.0218	0.8987 ± 0.0648	0.8434 ± 0.0986
Model_o2bl	0.7352 ± 0.0961	0.7115 ± 0.1156	0.6394 ± 0.2160	0.7789 ± 0.1040	0.6608 ± 0.1423	0.4605 ± 0.2085
Model_o2bc	0.8754 ± 0.0659	0.8402 ± 0.1320	0.8691 ± 0.1368	0.9191 ± 0.0584	0.8423 ± 0.0851	0.7564 ± 0.1233
Model_i1	0.8528 ± 0.0392	0.8444 ± 0.0577	0.8250 ± 0.0586	0.9245 ± 0.0341	0.8333 ± 0.0439	0.7035 ± 0.0799
Model_i2a	0.8899 ± 0.0306	0.9278 ± 0.0438	0.9375 ± 0.0244	0.9228 ± 0.0522	0.9317 ± 0.0163	0.6495 ± 0.0140
Model_i2bo	0.8945 ± 0.0671	0.9500 ± 0.1118	0.8429 ± 0.1558	0.9075 ± 0.0610	0.8799 ± 0.0780	0.8103 ± 0.1187
Model_i2bs	0.7857 ± 0.1022	0.8556 ± 0.0843	0.7679 ± 0.1651	0.8674 ± 0.1697	0.8012 ± 0.0988	0.5926 ± 0.1977
Model_i2bl	0.8615 ± 0.0843	0.7833 ± 0.2173	0.7167 ± 0.1826	0.8478 ± 0.1021	0.7310 ± 0.1631	0.6542 ± 0.2120
Model_i2bc	0.8505 ± 0.0853	0.7928 ± 0.1308	0.7800 ± 0.2465	0.9544 ± 0.0435	0.7699 ± 0.1614	0.6767 ± 0.2074

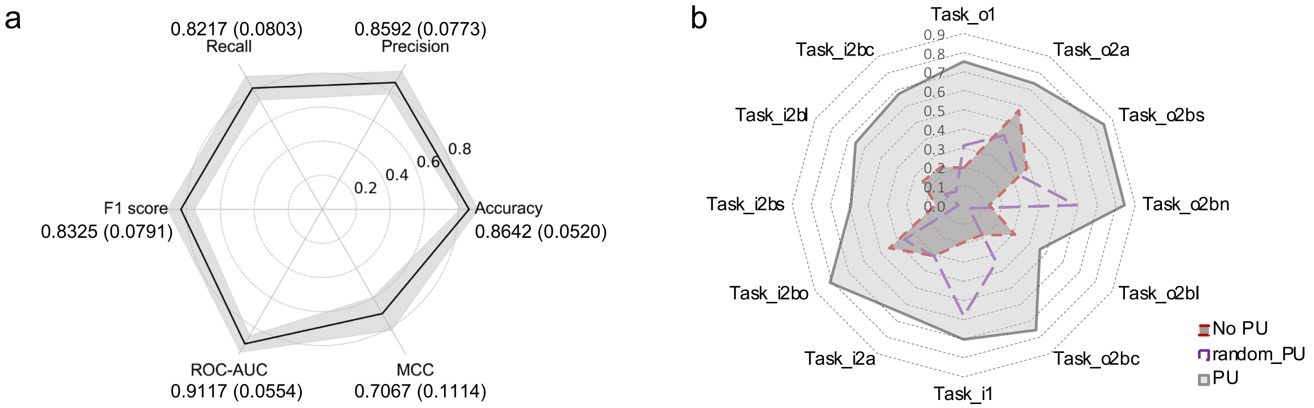


Fig. 7. Average model performance and model comparison. **a.** Average performance of the top models across 12 tasks. The shading areas and the values in parentheses denote the standard deviation. **b.** Spider plot comparing MCC metrics across 12 tasks using three different data processing methods. Gray: PU learning method adopted in this research; orange: no PU learning; purple: PU learning with randomly selected positive samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

scientists can also optimize drug design by adjusting molecular structures during early development stages to mitigate certain feature impacts. Furthermore, SHAP analysis can assess interactions between different descriptors that might produce nonlinear effects in complex biological systems. These profound insights will facilitate more accurate predictions for the in vivo fate of drugs.

3.5. Web-platform construction and function display

To broaden the application scenarios and reduce the application barrier of the developed AI system, we deployed it as a user-friendly web platform named FormulationDT. With FormulationDT, users only need to input drug names or the structure of candidate drugs to promptly receive recommendations for suitable formulation strategies. The user interface of FormulationDT comprises two primary modules. The “Webserver” module serves as the portal for accessing and utilizing FormulationDT, while the “Documentation” module offers information pertaining to the dataset and model performance. Users simply need to type the name or input the structure of the query molecule using either Simplified Molecular Input Line Entry System (SMILES) [50] notation or by directly drawing it. FormulationDT will then intelligently perform the prediction and analysis process. The format conversion function of *ChemDes* [51] is provided to facilitate the user to obtain the SMILES of the molecule. Fig. 9 displays the snapshots of the FormulationDT user

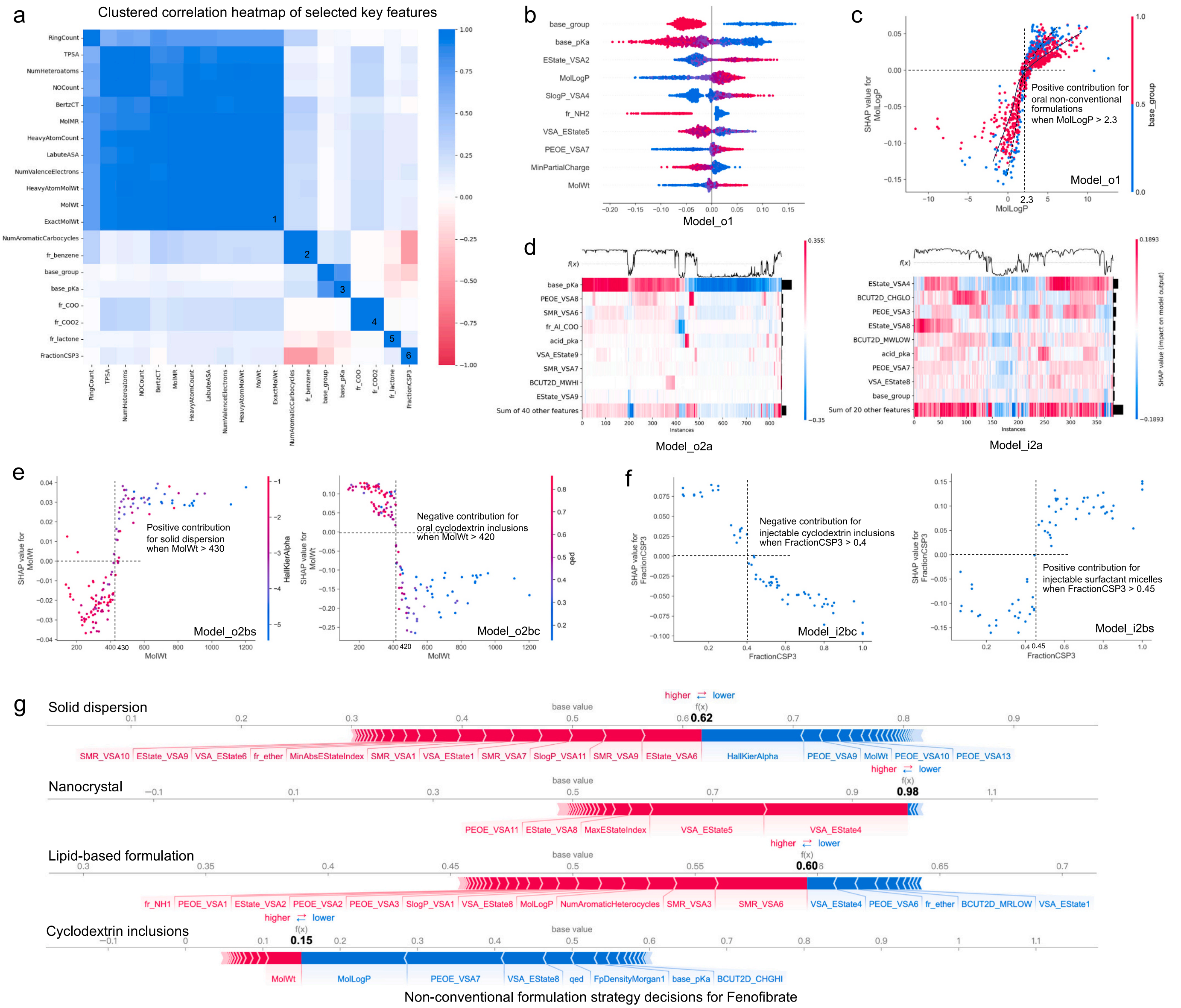
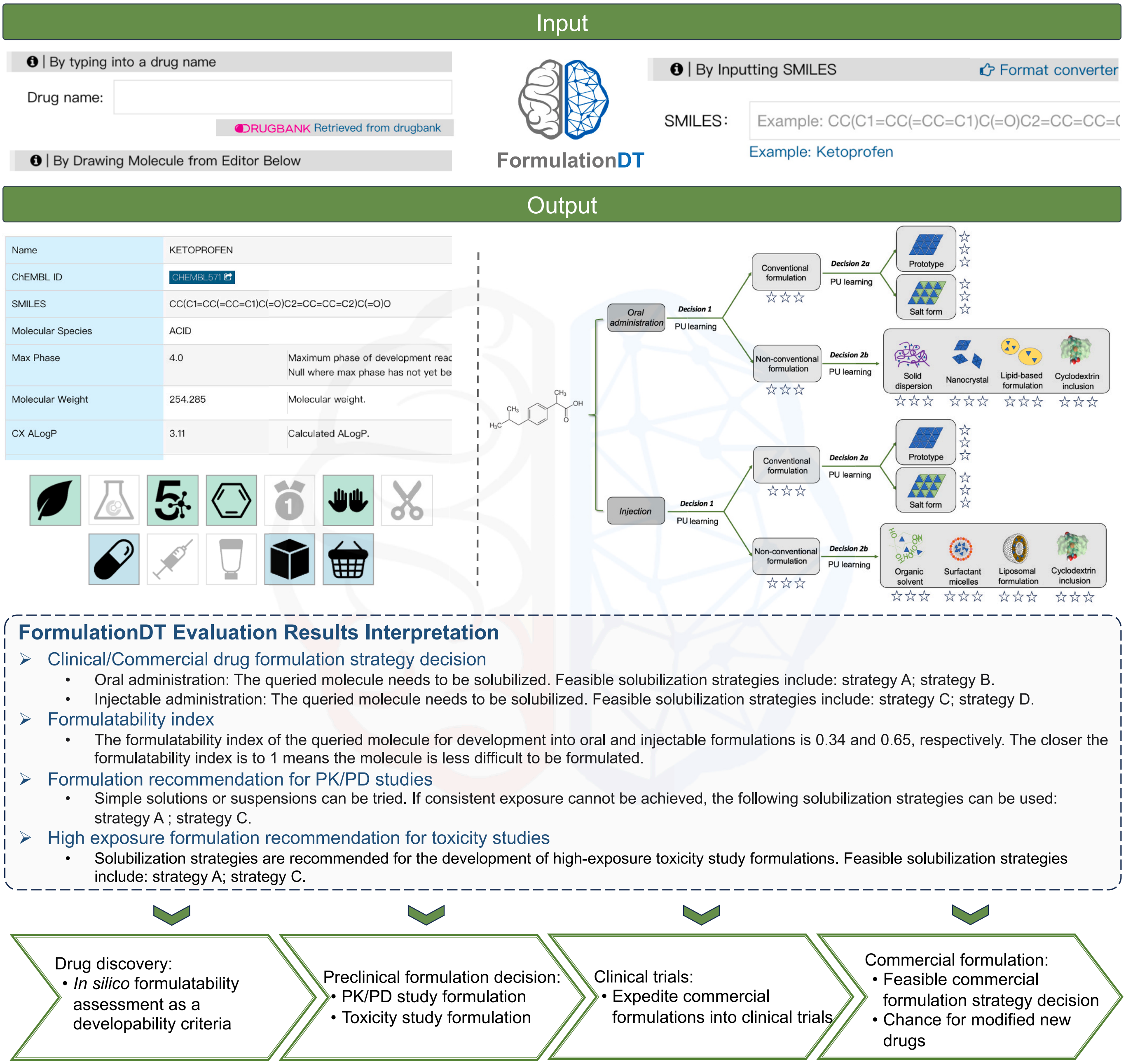


Fig. 8. Model interpretation and analysis. **a.** Clustered correlation heatmap of top 20 key features with the most significant average impact across the 12 tasks. The green and red colour respectively indicates positive and negative correlation. **b.** Beeswarm plot of the contributions of top 10 important features for Model_o1. Blue dots indicate instances with lower feature values, while red dots represent higher values. The horizontal coordinate for each instance reflects the influence of the feature value on model decision, with positive and negative values indicating positive and negative contribution to the positive model decision, respectively. The absolute magnitude of the coordinate indicates the quantitative contribution. **c.** Interaction dependence plot of the influence of MolLogP and basic group existance on Model_o1. Each point represents a sample. The colour of the point indicates the presence (red) or absence (blue) of basic ionizable groups. The horizontal coordinate represents the actual MolLogP value of the sample, while the y-axis represents the contribution of this sample's MolLogP value to the model's positive prediction. The value greater or less than 0 respectively indicates a positive or negative contribution. **d.** Heatmap plots for Model_o2a and Model_i2a. The top section is the model's output, clustered and arranged based on similar output values obtained from different feature combinations. The heatmap below represents the influence of feature combinations, with red and blue respectively indicating positive and negative contributions of the feature to the model's positive output. **e.** Interaction dependance plot showing the interaction of MolWt with HallKierAlpha and QED index to influence the prediction of Model_o2bs and Model_o2bc, respectively. **f.** Dependance plot showing how FractionCSP3 influence the prediction of Model_i2bc and Model_i2bs. The interpretation of subplot e and i are similar to subplot c. **g.** The force plots of four oral bio-enabling formulation feasibility analysis for Fonofibrate. The length of the colour blocks quantifies the contribution of the features to the model's prediction, with red and blue representing positive and negative contributions to the model's positive output, respectively. The bold numbers indicate the models' predicted positive outcome probability. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

interface. The output content consists of three components. First, it presents the calculated information and characteristics of the input molecule. Second, it visualizes the decision-making process for formulation strategy. In this visualization, the formulation strategy recommended by the AI system is depicted in colour, with the number of pentagrams representing the level of recommendation which is derived from the probability belonging to a specific class predicted by machine learning models. Note that Task 2b consists of multiple binary classification models, so it is possible for no viable strategies to be selected. In this case, alternative solubilization strategies not included in this study can be considered for the queried molecule. Lastly, FormulationDT integrates the outputs of the AI system with domian knowledge to provide interpretation of the decision results, which encompasses various application scenarios.

The core functionality of FormulationDT is to decide on the appropriate commercial formulation strategy according to the input



Name	KETOPROFEN	
ChEMBL ID	CHEMBL571	
SMILES	CC(C1=CC(=CC=C1)C(=O)C2=CC=CC=C2)C(=O)O	
Molecular Species	ACID	
Max Phase	4.0	Maximum phase of development reac Null where max phase has not yet be
Molecular Weight	254.285	Molecular weight.
CX ALogP	3.11	Calculated ALogP.

FormulationDT Evaluation Results Interpretation

➤ Clinical/Commercial drug formulation strategy decision

- Oral administration: The queried molecule needs to be solubilized. Feasible solubilization strategies include: strategy A; strategy B.
- Injectable administration: The queried molecule needs to be solubilized. Feasible solubilization strategies include: strategy C; strategy D.

➤ Formulatability index

- The formulatability index of the queried molecule for development into oral and injectable formulations is 0.34 and 0.65, respectively. The closer the formulatability index is to 1 means the molecule is less difficult to be formulated.

➤ Formulation recommendation for PK/PD studies

- Simple solutions or suspensions can be tried. If consistent exposure cannot be achieved, the following solubilization strategies can be used: strategy A ; strategy C.

➤ High exposure formulation recommendation for toxicity studies

- Solubilization strategies are recommended for the development of high-exposure toxicity study formulations. Feasible solubilization strategies include: strategy A; strategy C.

Drug discovery:
- *In silico* formulatability assessment as a developability criteria

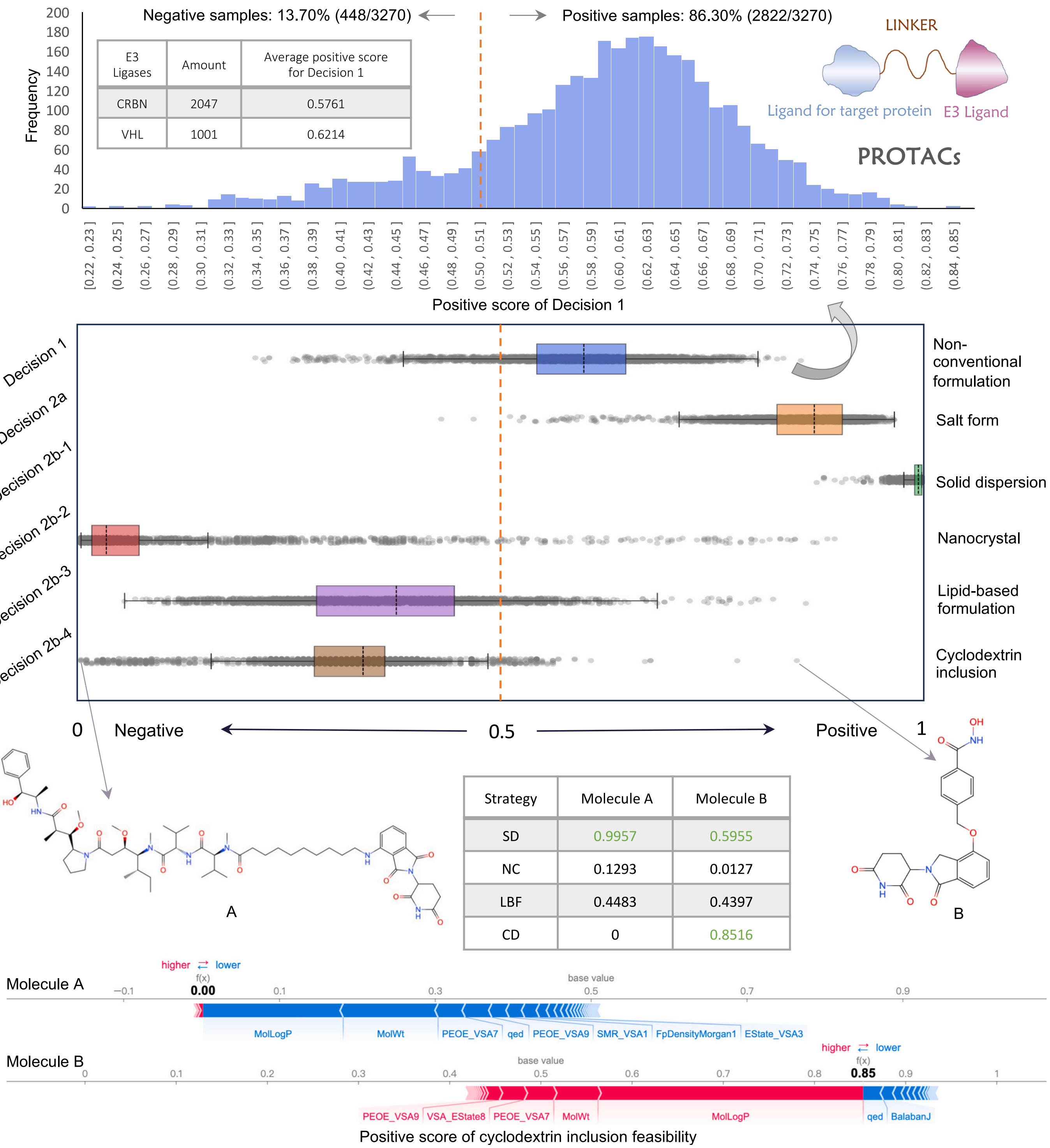
Preclinical formulation decision:
- PK/PD study formulation
- Toxicity study formulation

Clinical trials:
- Expedite commercial formulations into clinical trials

Commercial formulation:
- Feasible commercial formulation strategy decision
- Chance for modified new drugs

Fig. 9. Snapshots of the user interface and functionality display of FormulationDT. Users can type the name or input the structure of the query molecule using either Simplified Molecular Input Line Entry System (SMILES) notation or by directly drawing it. The output of FormulationDT will include the basic information of the query molecule, the formulation strategy design results for both oral and injectable administration. Moreover, the detailed interpretation for the design result will be provided to enable the application of FormulationDT in drug discovery, preclinical formulation design, expediting commercial formulations into clinical trials, feasible commercial formulation strategy decision, and identifying chance for modified new drugs.

molecules. Based on such core function and the well-developed result interpretation, FormulationDT can serve as a formulation expert to exert influence across various stages of drug discovery. First, FormulationDT provides a formulatability index for drug discovery. For oral and injectable administration, the formulatability index equals the probability that the input molecule is negative (indicating that non-conventional formulation strategies are not required) predicted by Model_o1 or Model_i1, respectively. The formulatability index ranges from 0 to 1, with values closer to 1 indicating lower difficulty in formulating the molecule. The formulatability index can be initially divided into three equal ranges indicating low, medium, and high formulatability. This can serve as one of the developability metrics for screening, generating, or designing drug molecules to control the cost and risk of subsequent drug development stages. Second, in the preclinical stage, if a simple solution or suspension fails to meet the formulation requirements, combining the core functionality of FormulationDT and the special requirements of preclinical formulations, the established AI system can also recommend feasible non-conventional formulations, thus facilitating the design of formulations with consistent and reproducible exposure for PK/PD studies, as well as formulations achieving high exposure for toxicity studies. Third, FormulationDT can assist formulation scientists in advancing suitable commercial formulations into clinical trials with minimal investment, thereby decreasing reliance on Fit-for-purpose (FFP) formulations. Such endeavors have been noted to significantly abbreviate the duration from Investigational New Drug (IND) to New Drug Application (NDA) [16]. In early clinical studies such as First in human study, Proof of concept study, etc., the FFP approach may be adopted to fulfill requirements by employing simple formulations, thereby avoiding excessive investment in formulation development during the high-risk clinical trial phase. However, the drawbacks of FFP approaches are evident; formulations unsuitable for large-scale production or failing to meet market and commercial demands necessitate formulation modifications later in development. These changes can potentially impact the pharmacokinetics of the initial FFP formulation, resulting in clinical downtime and necessitating bridging in vivo studies, thereby incurring additional risk and investment. Forth, FormulationDT can be employed for the retrospective analysis of marketed molecular entities to identify



E3 Ligases	Amount	Average positive score for Decision 1
CRBN	2047	0.5761
VHL	1001	0.6214

Strategy	Molecule A	Molecule B
SD	0.9957	0.5955
NC	0.1293	0.0127
LBF	0.4483	0.4397
CD	0	0.8516

Fig. 10. The application of FormulationDT on PROTACs. The central boxplot shows the oral prediction results of FormulationDT for 3270 PROTACs, with each gray dot representing a sample. Above is a histogram of the frequency distribution of the necessity for oral solubilization of PROTACs. Below are the solubilization technique design results and cyclodextrin inclusion feasibility analysis for two example molecules, shown as force plots. The length of the colour blocks quantifies the contribution of features to the model's prediction, with red and blue representing positive and negative contributions to the model's positive output, respectively. The bold numbers indicate the probability of a positive result predicted by the model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

opportunities for developing formulation optimization-based modified new drugs. This is essential for improving drug efficacy, safety, and adherence, while also addressing unmet clinical and market needs.

3.6. Applications of FormulationDT

3.6.1. Prospective study facilitating PROTACs development

To demonstrate the prospective guidance provided by FormulationDT for new molecule formulation development, we applied FormulationDT to oral formulation strategy design for 3270 PROTACs curated in PROTAC-DB [52,53]. PROTACs induce selective degradation of target proteins through the ubiquitin-proteasome system, representing an innovative drug discovery strategy that has garnered widespread attention. Despite significant progress in the past decade, designing ideal PROTACs remains a substantial challenge. To date, several PROTACs are in clinical stages, but none have been approved. In Decision 1, FormulationDT predicts that non-conventional formulation is necessary for over 85 % PROTACs. PROTACs are heterobifunctional molecules consisting of a small molecule targeting the protein of interest, a small molecule recruiting an E3 ligase, and a linker connecting these two moieties. The high molecular weight limits their solubility, permeability, and other drug-like properties. Notably, differences in the average positive scores of PROTACs composed of different E3 ligase ligands were observed. As depicted at the top of Fig. 10, a total of 2047 CRBN-targeted PROTACs exhibited lower average positive scores compared to VHL-targeted PROTACs, indicating that the CRBN-targeted PROTACs tend to have better formulatability. Currently most of the PROTACs entering clinical trials are CRBN-targeted PROTACs, partly due to the enhanced druggability attributed to the relatively smaller molecular weight of CRBN E3 ligase ligands [54]. Regarding Decision 2a, the models deemed the majority of PROTAC molecules feasible for salt formation. This suggests to pharmaceutical scientists that, for molecules with moderate solubility enhancement needs, prioritizing salt formation could be advantageous. All molecules were considered suitable for development as solid dispersions which is exactly the commonly used bio-enabling strategy for PROTACs [55]. Nanocrystals and lipid-based formulations exhibit structural preferences in PROTAC applications. With FormulationDT, the feasibility of developing PROTACs into nanocrystals or lipid-based formulations can be quantitatively assessed. Regarding cyclodextrin inclusion technology, the model indicates that most PROTACs are unsuitable. This is because PROTACs consist of three linked components, resulting in elongated molecular shapes (e.g., molecule A) that do not fit well within the cavities of commonly used cyclodextrins [56]. However, there are exceptions; molecules with relatively smaller molecular weight and length, such as molecule B, are considered to have potential for development as cyclodextrin complexes. FormulationDT showcases its ability to make comprehensive formulation decisions in bulk and swiftly for any drug candidate with known structure, which will greatly facilitate drug development through rational formulation strategy selection and has the potential to guide the screening and design of drug molecules.

3.6.2. Identifying modification opportunities for approved drugs

Through retrospectively evaluating marketed molecules, FormulationDT can also identify opportunities for modified new drugs, which are of vital importance to enhance the efficacy, safety, and adherence of drugs and to fill unmet clinical gaps [57]. In 2023, the U.S. FDA approved 29 small molecule NMEs, with 23 of them being orally administered. Among these oral molecules, 7 were marketed as salt forms. The results of FormulationDT's batch predictions on such orally administered NMEs are presented in Fig. 11. For the formulatability assessment in Decision 1, FormulationDT predicted that 10 out of the 23 orally administered NMEs required improvements in delivery efficiency through non-conventional formulation strategies (the left side of Fig. 11). Suitable solubilization strategies were also provided by FormulationDT, with different technically recommended priorities. With the intelligent decisions of FormulationDT, users can quickly follow up and efficiently conduct the development of modified new drugs in conjunction with their respective non-technical considerations. For the remaining 13 molecules with high formulatability, FormulationDT's salt formation feasibility decision concluded that 8 of them could be formulated as salt forms (the right side of Fig. 11). All 7 molecules currently marketed in salt form were successfully identified by FormulationDT as feasible for formulation as salt, highlighting FormulationDT's high recall in predicting salt formation feasibility. Capivasertib, currently marketed as a prototype, was assigned a salt-forming feasibility score of up to 0.98 by FormulationDT. This suggests that drug developers may explore the clinical or manufacturing advantages of Capivasertib's salt form to uncover opportunities for modified new drugs. For the remaining 5 molecules determined by FormulationDT as not requiring being formulated as non-conventional formulations and lacking salt formation feasibility, efforts for modified new drugs should not be wasted in the above manner. Possible alternative approaches can be explored, such as optimizing manufacturing processes or making slight modifications to the molecular structure to enhance stability or reduce toxicity.

4. Discussion

In the era of *Pharma 4.0*, significant progress has been made in AI-driven drug design [58,59] and rational formulation design [60–64]. However, as the early stage of drug development, formulation strategy decisions still rely on costly trial-and-error tests or limited experience. Experience is helpful, provided it is explicit and robust. Over the past century of modern pharmacy, thousands of approved drugs have accumulated the wisdom from countless scientists worldwide. Quantifying, instrumenting, and organically integrating these valuable lessons into an AI decision-making platform is the motivation and goals of the present work. To begin with, the formulation information of small molecule drugs approved by the U.S. FDA was manually collected, constituting what is, to our knowledge, the first systematic dataset on this topic. Following this, based on the scientific principle that structure determines nature and influences decision-making, we developed the PU-Decide framework to address the problem of missing reliable negative samples in marketed drug data to establish correlations between molecule structure and formulation decisions. The average ROC_AUC score of the best models exceeded 0.91 for total 12 classification tasks, ranging from 0.78 to 0.98. Lastly, integrating data-driven machine learning models with domain knowledge, the first AI formulation strategy decision-making platform was successfully developed, which is presented in a user-friendly website and freely available for drug discovery and development scientists.

Distinct from the expert system-type formulation strategy decision studies, the data-driven FormulationDT summarizes and quantifies the successes of approved drugs, enabling better generalization and providing clearer decision guidance. To facilitate end-to-end decision-making, calculated and predicted descriptors, rather than costly experimental properties, are utilized as input. Such design not only lowers the barrier of applying FormulationDT, but also enables high-throughput molecule assessment, which promises to expand the application scenarios of FormulationDT to different stages of drug discovery and development, such as formulatability assessment and rational drug design. The intentional development of PU-Decide framework addresses the structural deficiencies of the available data and, more importantly, demonstrates the utility of semi-supervised learning for localizing and exploring specific chemical spaces in drug development scenarios. The design and deployment of the online website enables the organic integration of machine learning models and domain knowledge, which further lowers the application barrier of FormulationDT, enhances its transparency, illuminates, and more importantly, empowers diverse application scenarios. As demonstrated in Section 3.5, beyond commercial drug formulation strategy decisions, FormulationDT assumes a

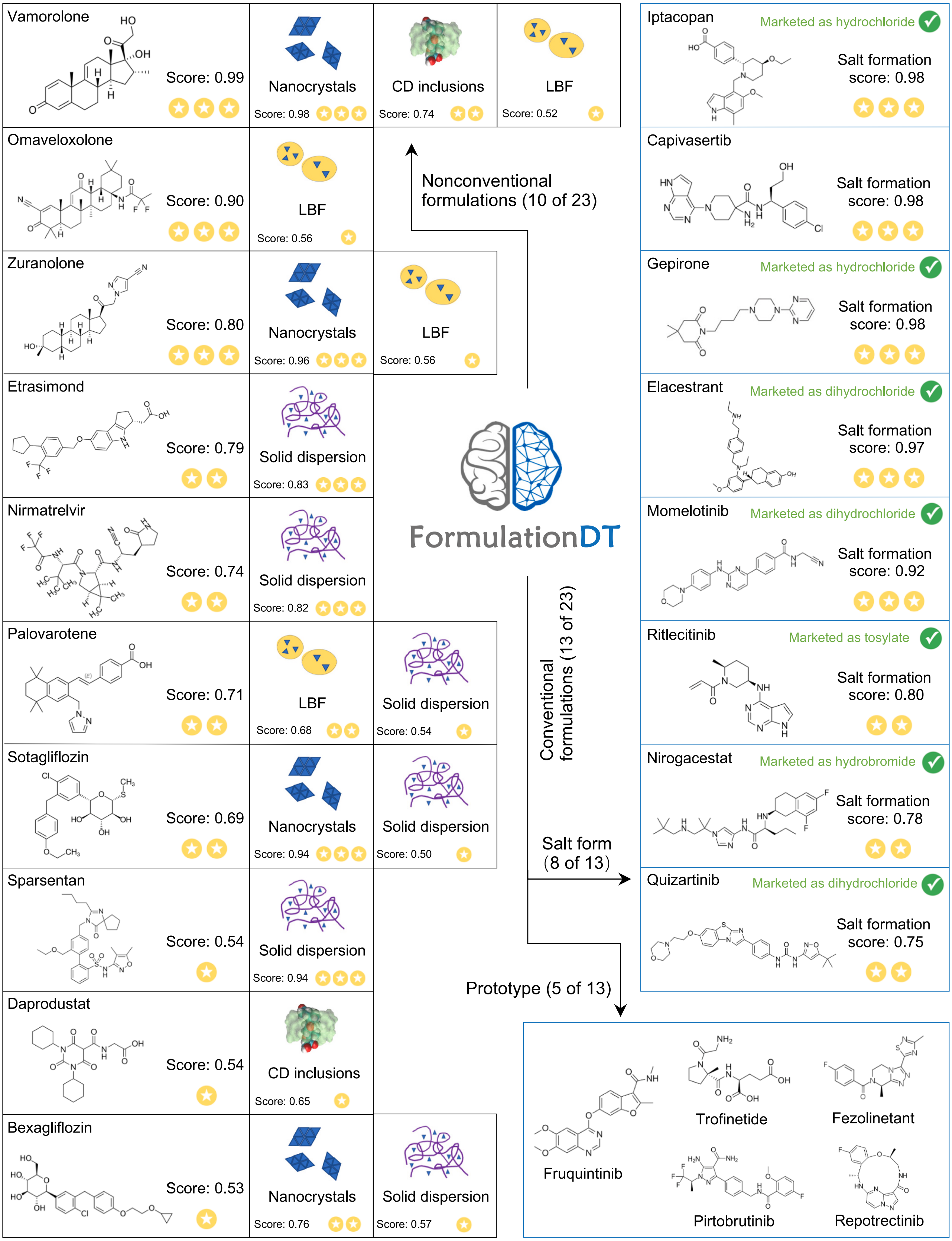


Fig. 11. The formulation strategy design by FormulationDT for the 23 oral NMEs approved by the U.S. FDA in the year 2023. The scores and the number of yellow stars indicate the degree of necessity for solubilization, feasibility for solubilization strategies, or feasibility for salt formation. Among the 13 molecules that do not require solubilization, 7 molecules that are marketed in salt form (with specific salt types noted in green text) were all correctly predicted (green check marks) with high salt formation feasibility. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

designer's role at multiple stages, from drug discovery, preclinical, clinical, and marketed formulation development, to the development of modified new drugs. Through its multi-scenario applications, FormulationDT promotes design-driven drug lifecycle development, aligning with the philosophy of "Quality by Design" [65]. It is anticipated to reduce the risk and cost of drug development, enhance development efficiency, and contribute to drug quality improvements. Furthermore, we believe that the current study and the corresponding dataset, as a pioneering attempt of PU learning in drug development, will contribute to the computational pharmacy community by advancing the semi-supervised learning paradigm for prediction and design tasks of drug development.

As the drug formulation strategy dataset becomes more comprehensive, the performance of the classification models in this study is expected to improve further. Currently, across a total of 12 classification tasks, the best model achieves an average ROC_AUC score exceeding 0.91, ranging from 0.78 to 0.98. Specifically, 3 models have ROC_AUC scores above 0.95, 9 above 0.90, and 11 above 0.84. Only the final model for Task_o2bl shows relatively lower predictive performance, with an ROC_AUC of 0.7789, which, although close to 0.8, still shows significant improvement over random guessing. At present, data-related issues—including quantity, quality, accessibility, and representativeness—remain the primary limitations on model performance. First, although the number of approved drugs has reached thousands, considering the complexity of the task, the expansion of data volume will lead to a more detailed portrayal of the chemical space of drugs, which would greatly benefit the performance of FormulationDT. Second, despite our efforts to collate formulation routes of listed drugs as accurately and comprehensively as possible, limitations such as transparency and the degree of information disclosure may lead to individual data omissions or mislabeling. Fortunately, the implementation of our PU learning framework PU-Decide somewhat attenuates the interference of outliers and ensures the robustness of the models. Third, data accessibility constrains what tasks can be accomplished. In the present work, we designed each step of decision according to available data. Decision 1 was for the necessity of non-conventional formulations, while Decision 2 determined the feasibility of salt formation or specific non-conventional strategies. However, recommending an "optimal formulation strategy" based solely on marketed drug data is not feasible. This is because determining the so-called "optimal" requires numerous comparative studies, leading to limited data availability. Additionally, for a particular R&D entity, the optimal formulation selection is also influenced by factors such as available production conditions, commercial and clinical needs, and intellectual property considerations [66]. Therefore, it is reasonable for FormulationDT to determine the technical feasibility of specific formulation strategies. The fourth aspect concerns data representation. This study adopts calculated descriptors to represent drug molecules, enabling end-to-end prediction. Most of these computed descriptors are used to depict the structural and microscopic underlying properties within atoms or molecules, offering insights for research into solubilization mechanisms and formulation principles. However, drug development scientists typically make formulation strategy decisions based on macroscopic molecular properties, such as solubility, permeability, and melting point [10,20,67,68], which provide more intuitive interpretability. In future work, by integrating findings from existing studies on formulation strategy decision-making, the PU-Decide framework could be adapted to handle property-based drug representations, thereby establishing a formulation strategy design platform more aligned with drug development intuition.

Despite data limitations, through the organic integration of the PU-Decide framework and domain knowledge, FormulationDT has showcased its ability to offer expert formulation strategy decisions, providing valuable assistance and inspiration for drug discovery and development. As the core of machine learning, data presents one of the most common challenges for current machine learning applications [61]. We anticipate that more effective data sharing between pharmaceutical academia and industry will enhance the performance and functionality of FormulationDT. Concurrently, innovations in data sharing approaches and data regulatory science are necessary to further break down data silos and promote the completion of digital drug development frameworks.

From a computational pharmaceutics standpoint, the successful establishment of FormulationDT adds a vital piece to the new computer-driven drug development paradigm we proposed in 2023 [61]. As illustrated in Fig. 12, distinct from the conventional inefficient "screen-validate-rescreen" formulation development procedure, the new "understand-design-validate-optimize" paradigm emphasizes the application of computational modeling to comprehend the in vivo fate of drugs and to guide the rational drug formulation design through an integrated computer-driven framework. FormulationDT will serve as the pivotal module of the in silico formulation design session. User-entered molecules will initially receive recommendations from FormulationDT for suitable formulation strategies. Subsequently, these molecules will progress to the next step into the FormulationAI [69] and PharmDE [70] modules. FormulationAI is an AI prediction platform for 16 essential formulation properties across six formulation types (cyclodextrin inclusions, solid dispersions, phospholipid complexes, nanocrystals, self-emulsifying system, and liposomal formulations). Efficient in silico excipient selection and formulation & process parameter design can be accomplished by simply entering basic information of the drug and excipients. PharmDE, for its part, was developed to complete drug-excipient compatibility assessments as part of excipient selections. Both the online webserver for FormulationAI (https://formulationai.computpharm.org/) and PharmDE (https://pharmde.computpharm.org/) are freely accessible. As a component of the in silico developability assessment, the preformulation properties prediction module is currently under development, which will further improve the predictive performance and interpretability of FormulationDT. It's expected that the development of FormulationDT and subsequent modules will propel the realization of an efficient computer-driven drug development paradigm.

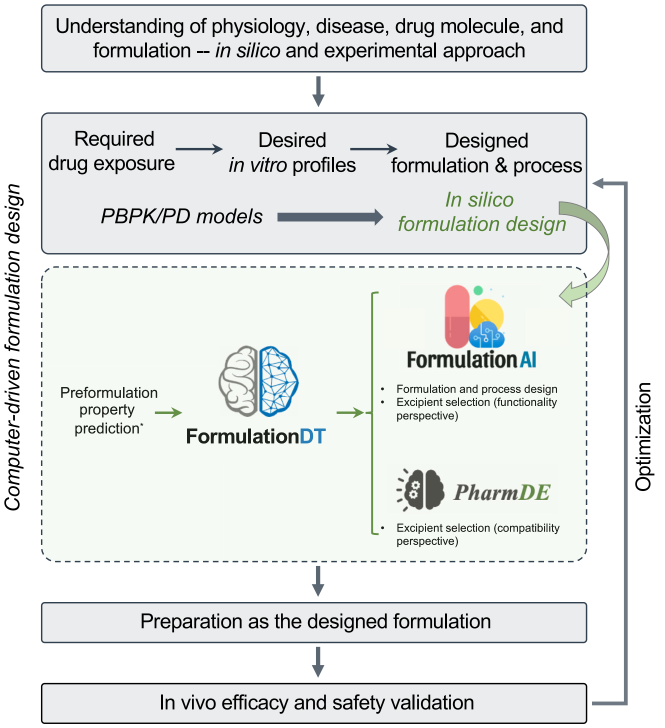

Understanding of physiology, disease, drug molecule, and formulation -- in silico and experimental approach
Required drug exposure
Desired in vitro profiles
Designed formulation & process
PBPK/PD models
In silico formulation design
Computer-driven formulation design
Preformulation property prediction*
FormulationDT
FormulationAI
Formulation and process design
Excipient selection (functionality perspective)
PharmDE
Excipient selection (compatibility perspective)
Optimization
Preparation as the designed formulation
In vivo efficacy and safety validation

Fig. 12. Future perspectives on the role of FormulationDT in computer-driven drug development framework.

5. Conclusion

In summary, learning from a compiled dataset of approved drug products, the current study successfully designed and developed FormulationDT, the first data-driven and knowledge-guided AI formulation strategy design platform for small molecules. Utilizing the PU-Decide framework, the efficient data representation, and the user-friendly webserver, the resulting AI platform can efficiently accomplish tasks at multiple stages of drug development, such as formulatability assessment, preclinical and clinical formulation strategy decisions. Bridging the gap in conventional formulation strategy decision-making, FormulationDT emerges as a key puzzle piece in the new paradigm of computer-driven drug development. Promising to enhance drug development efficiency and improve drug quality, FormulationDT is poised for continual refinement through user feedback, ultimately showcasing its value in the *Pharma 4.0* era.

CRediT authorship contribution statement

Nannan Wang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jie Dong:** Writing – review & editing, Visualization, Validation, Supervision, Software, Project administration, Investigation. **Defang Ouyang:** Writing – review & editing, Validation, Supervision, Project administration, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and code for implementing this study are available on the GitHub repository at https://github.com/NamanWang/FormulationDT.

Acknowledgements

This work was financially supported by the Macau Science and Technology Development Fund (0071/2024/RIA1) and University of Macau Multi-Year Research Grant (MYRG-GRG2023-00077-ICMS-UMDF) and Dr. Stanley Ho Medical Development Foundation “Set Sail for New Horizons, Create the Future” Grant 2024 (SHMDF-VSEP/2024/001).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jconrel.2024.12.043.

References

[1] D. Sun, et al., Why 90% of clinical drug development fails and how to improve it? Acta Pharm. Sin. B 12 (7) (2022) 3049–3062.

[2] C.A. Lipinski, Drug-like properties and the causes of poor solubility and poor permeability, J. Pharmacol. Toxicol. Methods 44 (1) (2000) 235–249.

[3] S.R. Munnangi, et al., Drug complexes: perspective from academic research and pharmaceutical market, Pharm. Res. 40 (6) (2023) 1519–1540.

[4] V. Saxena, et al., Developability assessment in pharmaceutical industry: an integrated group approach for selecting developable candidates, J. Pharm. Sci. 98 (6) (2009) 1962–1979.

[5] M. Palucki, et al., Strategies at the interface of drug discovery and development: early optimization of the solid state phase and preclinical toxicology formulation for potential drug candidates, J. Med. Chem. 53 (16) (2010) 5897–5905.

[6] A.B. Benowitz, P.T. Scott-Stevens, J.D. Harling, Challenges and opportunities for in vivo PROTAC delivery, Future Med. Chem. 14 (3) (2022) 119–121.

[7] S. Jankovic, et al., Application of the solubility parameter concept to assist with oral delivery of poorly water-soluble drugs – a PEARRL review, J. Pharm. Pharmacol. 71 (4) (2019) 441–463.

[8] Y. Kawabata, et al., Formulation design for poorly water-soluble drugs based on biopharmaceutics classification system: basic approaches and practical applications, Int. J. Pharm. 420 (1) (2011) 1–10.

[9] B. Xie, et al., Solubilization techniques used for poorly water-soluble drugs, Acta Pharm. Sin. B. 14 (11) (2024) 4683–4716.

[10] S. Branchu, et al., A decision-support tool for the formulation of orally active, poorly soluble compounds, Eur. J. Pharm. Sci. 32 (2) (2007) 128–139.

[11] C. Grudzinskas, et al., Impact of formulation on the abuse liability, safety and regulation of medications: the expert panel report, Drug Alcohol Depend. 83 (Suppl. 1) (2006) S77–S82.

[12] C.A. Lipinski, et al., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 23 (1–3) (1997) 3–25.

[13] G.R. Bickerton, et al., Quantifying the chemical beauty of drugs, Nat. Chem. 4 (2) (2012) 90–98.

[14] J.M. Butler, J.B. Dressman, The developability classification system: application of biopharmaceutics concepts to formulation development, J. Pharm. Sci. 99 (12) (2010) 4940–4954.

[15] J. Rosenberger, J. Butler, J. Dressman, A refined developability classification system, J. Pharm. Sci. 107 (8) (2018) 2020–2032.

[16] M.S. Ku, W. Dulin, A biopharmaceutical classification-based right-first-time formulation approach to reduce human pharmacokinetic variability and project cycle time from first-in-human to clinical proof-of-concept, Pharm. Dev. Technol. 17 (3) (2012) 285–302.

[17] A. Van den Bergh, et al., Preclinical bioavailability strategy for decisions on clinical drug formulation development: an in depth analysis, Mol. Pharm. 15 (7) (2018) 2633–2645.

[18] P. Zane, et al., In vivo models and decision trees for formulation development in early drug development: a review of current practices and recommendations for biopharmaceutical development, Eur. J. Pharm. Biopharm. 142 (2019) 222–231.

[19] W. Zheng, et al., Selection of oral bioavailability enhancing formulations during drug discovery, Drug Dev. Ind. Pharm. 38 (2) (2012) 235–247.

[20] G.A. Fridgeirsdottir, et al., Support tools in formulation development for poorly soluble drugs, J. Pharm. Sci. 105 (8) (2016) 2260–2269.

[21] G.L. Amidon, et al., A theoretical basis for a biopharmaceutic drug classification: the correlation of in vitro drug product dissolution and in vivo bioavailability, Pharm. Res. 12 (1995) 413–420.

[22] T. Peryea, et al., Global substance registration system: consistent scientific descriptions for substances related to health, Nucleic Acids Res. 49 (D1) (2021) D1179–D1185.

[23] J. Bekker, J. Davis, Learning from positive and unlabeled data: a survey, Mach. Learn. 109 (4) (2020) 719–760.

[24] K. Daidoji, S. Yasukawa, S. Kano, Effects of new formulation strategy on life cycle management in the US pharmaceutical industry, J. Gener. Med. 10 (3–4) (2013) 172–179.

[25] M. Kuentz, et al., Rational selection of bio-enabling Oral drug formulations – a PEARRL commentary, J. Pharm. Sci. 110 (5) (2021) 1921–1930.

[26] B. Žunkovič, Positive unlabeled learning with tensor networks, Neurocomputing 552 (2023).

[27] Y. Zheng, et al., DDI-PULearn: a positive-unlabeled learning method for large-scale prediction of drug-drug interactions, BMC Bioinformatics 20 (2019) 1–12.

[28] W. Lan, et al., Predicting drug–target interaction using positive-unlabeled learning, Neurocomputing 206 (2016) 50–57.

[29] P. Yang, et al., Positive-unlabeled learning for disease gene identification, Bioinformatics 28 (20) (2012) 2640–2647.

[30] P. Yang, et al., Ensemble positive unlabeled learning for disease gene identification, PLoS ONE 9 (5) (2014) e97079.

[31] G. Landrum, RDKit: a software suite for cheminformatics, computational chemistry, and predictive modeling, Greg Landrum (2013) 8.

[32] X. Pan, et al., MolGpka: a web server for small molecule pKa prediction using a graph-convolutional neural network, J. Chem. Inf. Model. 61 (7) (2021) 3159–3165.

[33] F. Pedregosa, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[34] F. Mordelet, J.-P. Vert, A bagging SVM to learn from positive and unlabeled examples, Pattern Recogn. Lett. 37 (2014) 201–209.

[35] N.V. Chawla, et al., SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[36] J. Snoek, H. Larochelle, R.P. Adams, Practical bayesian optimization of machine learning algorithms, Adv. Neural Inf. Proces. Syst. 25 (2012).

[37] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf. Proces. Syst. 30 (2017).

[38] T.E. Oliphant, Guide to Numpy Vol. 1, Trelgol Publishing USA, 2006.

[39] J. Reback, et al., Pandas-Dev/Pandas: Pandas 1.0. 5. Zenodo, 2020.

[40] M.S. Landis, et al., Commentary: why pharmaceutical scientists in early drug discovery are critical for influencing the design and selection of optimal drug candidates, AAPS PharmSciTech 19 (1) (2018) 1–10.

[41] A.T.M. Serajuddin, Salt formation to improve drug solubility, Adv. Drug Deliv. Rev. 59 (7) (2007) 603–616.

[42] T. Cover, P. Hart, Nearest neighbor pattern classification, IEEE Trans. Inf. Theory 13 (1) (1967) 21–27.

[43] J.R. Quinlan, Induction of decision trees, Mach. Learn. 1 (1) (1986) 81–106.

[44] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (2–3) (1997) 131–163.

[45] O.M. Feeney, et al., 50 years of oral lipid-based formulations: provenance, progress and future perspectives, Adv. Drug Deliv. Rev. 101 (2016) 167–194.

[46] W. Zhang, et al., Evaluation of accuracy of amorphous solubility advantage calculation by comparison with experimental solubility measurement in buffer and biorelevant media, Mol. Pharm. 15 (4) (2018) 1714–1723.

[47] L.H. Hall, L.B. Kier, The molecular connectivity chi indexes and kappa shape indexes in structure-property modeling, Rev. Comput. Chem. (1991) 367–422.

[48] P. Agarwal, et al., Trends in small molecule drug properties: a developability molecule assessment perspective, Drug Discov. Today 27 (12) (2022) 103366.

[49] W. Wei, et al., Fsp3: a new parameter for drug-likeness, Drug Discov. Today 25 (10) (2020) 1839–1845.

[50] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36.

[51] J. Dong, et al., ChemDes: an integrated web-based platform for molecular descriptor and fingerprint computation, J. Chemother. 7 (1) (2015) 60.

[52] G. Weng, et al., PROTAC-DB: an online database of PROTACs, Nucleic Acids Res. 49 (D1) (2020) D1381–D1387.

[53] G. Weng, et al., PROTAC-DB 2.0: an updated database of PROTACs, Nucleic Acids Res. 51 (D1) (2022) D1367–D1372.

[54] H. Jiang, et al., E3 ligase ligand optimization of clinical PROTACs, Front. Chem. 11 (2023) 1098331.

[55] F. Pöstges, et al., Solubility enhanced formulation approaches to overcome oral delivery obstacles of PROTACs, Pharmaceutics 15 (1) (2023).

[56] Q. Zhao, et al., Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques, Acta Pharm. Sin. B 9 (6) (2019) 1241–1252.

[57] W.F. Salminen, M.E. Wiles, R.E. Stevens, Streamlining nonclinical drug development using the FDA 505 (b)(2) new drug application regulatory pathway, Drug Discov. Today 24 (1) (2019) 46–56.

[58] J. Vamathevan, et al., Applications of machine learning in drug discovery and development, Nat. Rev. Drug Discov. 18 (6) (2019) 463–477.

[59] A.V. Sadybekov, V. Katritch, Computational approaches streamlining drug discovery, Nature 616 (7958) (2023) 673–685.

[60] W. Wang, et al., Computational pharmaceutics-a new paradigm of drug delivery, J. Control. Release 338 (2021) 119–136.

[61] N. Wang, et al., How can machine learning and multiscale modeling benefit ocular drug development? Adv. Drug Deliv. Rev. 196 (2023) 114772.

[62] P. Bannigan, et al., Machine learning directed drug formulation development, Adv. Drug Deliv. Rev. 175 (2021) 113806.

[63] Z. Bao, et al., Revolutionizing drug formulation development: the increasing impact of machine learning, Adv. Drug Deliv. Rev. 202 (2023) 115108.

[64] Wang, N., et al., Introduction to computational pharmaceutics. Explor. Comput. Pharm. AI Model. Pharma 4.0, 2024: p. 1–9.

[65] L.X. Yu, et al., Understanding pharmaceutical quality by design, AAPS J. 16 (2014) 771–783.

[66] P. van Hoogevest, X. Liu, A. Fahr, Drug delivery strategies for poorly water-soluble drugs: the industrial perspective, Expert Opin. Drug Deliv. 8 (11) (2011) 1481–1500.

[67] M. Kuentz, R. Holm, D.P. Elder, Methodology of oral formulation selection in the pharmaceutical industry, Eur. J. Pharm. Sci. 87 (2016) 136–163.

[68] Y.-C. Lee, P.D. Zocharski, B. Samas, An intravenous formulation decision tree for discovery compound formulation development, Int. J. Pharm. 253 (1–2) (2003) 111–119.

[69] J. Dong, et al., FormulationAI: a novel web-based platform for drug formulation design driven by artificial intelligence, Brief. Bioinform. 25 (1) (2023).

[70] N. Wang, et al., PharmDE: a new expert system for drug-excipient compatibility evaluation, Int. J. Pharm. 607 (2021) 120962.