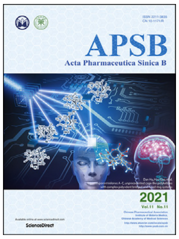
Acta Pharmaceutica Sinica B 2021;11(11):3585–3594

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com



ORIGINAL ARTICLE

Integrated *in silico* formulation design of self-emulsifying drug delivery systems

Haoshi Gao[a,b,†], Haoyue Jia[c,†], Jie Dong[a,†], Xinggang Yang[c,*], Haifeng Li[b,*], Defang Ouyang[a,*]

[a] *State Key Laboratory of Quality Research in Chinese Medicine, Institute of Chinese Medical Sciences (ICMS), University of Macau, Macau 999078, China*
[b] *Institute of Applied Physics and Materials Engineering, University of Macau, Macau 999078, China*
[c] *School of Pharmacy, Shenyang Pharmaceutical University, Shenyang 110016, China*

Received 22 January 2021; received in revised form 23 March 2021; accepted 31 March 2021

KEY WORDS

Self-emulsifying drug delivery system;
Random forest;
Central composite design;
Molecular dynamic simulation;
Meloxicam

Abstract The drug formulation design of self-emulsifying drug delivery systems (SEDDS) often requires numerous experiments, which are time- and money-consuming. This research aimed to rationally design the SEDDS formulation by the integrated computational and experimental approaches. 4495 SEDDS formulation datasets were collected to predict the pseudo-ternary phase diagram by the machine learning methods. Random forest (RF) showed the best prediction performance with 91.3% for accuracy, 92.0% for sensitivity and 90.7% for specificity in 5-fold cross-validation. The pseudo-ternary phase diagrams of meloxicam SEDDS were experimentally developed to validate the RF prediction model and achieved an excellent prediction accuracy (89.51%). The central composite design (CCD) was used to screen the best ratio of oil-surfactant-cosurfactant. Finally, molecular dynamic (MD) simulation was used to investigate the molecular interaction between excipients and drugs, which revealed the diffusion behavior in water and the role of cosurfactants. In conclusion, this research combined machine learning, central composite design, molecular modeling and experimental approaches for rational SEDDS formulation design. The integrated computer methodology can decrease traditional drug formulation design works and bring new ideas for future drug formulation design.

*Corresponding authors. Tel./fax: +853 88224514 (Defang Ouyang), +86 24 23986315 (Xinggang Yang), +853 88224035 (Haifeng Li).
E-mail addresses: yangxg123@163.com (Xinggang Yang), heifengli@um.edu.mo (Haifeng Li), defangouyang@umac.mo (Defang Ouyang).

[†] These authors made equal contributions to this work.

Peer review under responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2021.04.017

1. Introduction

In drug discovery, water-insoluble drugs face continuous hurdles in transforming into market medical products. As a simple administration route, oral administration is safe, convenient, underspend, and patient compliant[1]. An increased number of BCS class II drugs pose enormous challenges for oral formulation development. Many factors, including APIs' physicochemical properties and complicated internal environment of humans and animals, may affect their bioavailability[2]. For lipophilic drugs, especially the class II drug, the limitation of their absorption in the human body is the dissolution rate in the GI tract[3]. Pharmaceutical scientists have developed strategies to solve this issue, including solid dispersions[4], cyclodextrin inclusions[5], and nanoscale formulations[6]. In lipid-based formulation, drugs existed in liquid-state instead of in solid-state. The solubility of drugs increased by enhancing solvent−solvent interaction between formulations and the GI environment[2]. The self-emulsifying drug delivery system (SEDDS), a stable thermodynamic nanoformulation, consists of oil, surfactant, cosurfactant and APIs. For oral administration, drugs are dissolved in the SEDDS instead of solid-state, which benefits the absorption in the GI tract[7].

Since 40 years, US Food and Drug Administration (FDA) have approved about 12 SEDDS marketed formulations, such as Gengraf®, Norvir® and Depakene®, etc[8]. However, the modern SEDDS formulations design profoundly depends on the experimentation by the skills of independent researchers. The process of SEDDS formulation design includes three step as below: the determination of drug solubility in several oils, surfactants and cosurfactants; dissolve the mixture of oils, surfactants and cosurfactant into distilling water, and then draw the ternary phase diagram to identify the self-emulsion area; the evaluation of the SEDDS formulation by multiple characterization methods[9,10]. Therefore, current formulation development of SEDDS urgently need some effective methods to assist experimental design.

Machine learning is one part of artificial intelligence, which could learn from the experience and data by computer algorithms[11]. Presently, machine learning has been widely used in pharmaceutical science, such as drug discovery[12], quantitative structure−activity relationship (QSAR)[13], quantitative structure−property relationship (QSPR)[14], biomedicine[15], and drug formulation design[11]. In drug formulation design, machine learning can be an auxiliary tool to lighten pharmaceutical scientists' workload. It can be applied to predict the formulation performance in drug development by inputting the physicochemical properties of APIs and excipients and process parameters. For example, Zhao et al.[16] used molecular descriptors of drugs and cyclodextrins as input values to predict the binding energy of the drug cyclodextrin system, and the results confirmed that the model had good accuracy. Han et al.[17] applied random forest to predict solid dispersion stability, which completed a high accuracy and obtained parameters affecting the stability. He et al.[18] developed a lightGBM model to predict the size and PDI of nanocrystals prepared by three methods, which provided a new idea in industrial pharmaceutics. Gao et al.[19] constructed a drug/phospholipid complexation rate predicting model by the lightGBM algorithm.

Molecular dynamic (MD) simulation is another computational method that aided the pharmaceutical formulation design, which could mimic the physicochemical processes in the molecular scale[20]. In pharmaceutical science, the MD simulation has gradually become an increasingly vital tool to help scientists understand the drug delivery mechanism of dissolution, solubility, controlled release, and targeted delivery[21]. In the past ten years, our group had investigated numerous dosage forms, including the preparation and dissolution behavior of solid dispersion[22,23], the interaction between drug and cyclodextrin[24], liposome[25], drug−phospholipid complex[19], self-assembly platinum prodrug[26].

This study aimed to integrate machine learning, MD simulation and experimental approaches to rationally design SEDDS formulations. Firstly, 4495 SEDDS formulation datasets were collected to predict the pseudo-ternary phase diagram by the machine learning methods. Meloxicam (MLX) was chosen as a model drug to select the optimal oils, surfactants, and cosurfactants. Secondly, a pseudo-ternary phase diagram predicting model of MLX-SEDDS was experimentally constructed to validate the prediction model. Finally, MD simulation was utilized to mimic the molecular dissolving behavior of MLX-SEDDS in water.

2. Materials and methods

2.1. *Materials*

MLX with 98% purity was purchased from Tianjin Heowns Biochemical Co., Ltd. (Tianjin, China). Labrafil M 1944 CS and Transcutol HP were obtained from Gattefossé (Saint-Priest Cedex, France). AEO-9, Cremophor RH40, and Cremophor EL were obtained from BASF (Ludwigshafen, Germany). Ethanol and isopropanol were purchased from Tianjin Fuyu Fine Chemical Co., Ltd. (Tianjin, China). Isopropyl myristate (IPM) was purchase from Shanghai CHUXING Chemical Co., Ltd. (Shanghai, China). Isopropyl palmitate (IPP) was purchase from Linyi Lusen Chemicals Co., Ltd. (Linyi, China). Caprylic/Capric Triglyceride (GTCC) was obtained from KLK OLEO (Petaling Jaya, Malaysia). Tween 80 was obtained from Tianjin Kemiou Chemical Reagent Co., Ltd. (Tianjin, China). PEG 400 was obtained from Tianjin Bodi Chemical Co., Ltd. (Tianjin, China).

Figures were plotted by OriginPro2018 SR1 (OriginLab Corporation, Northampton, MA, USA), Amber18 package (University of California, San Francisco, CA, USA), Design-Expert v10 (Stat-Ease, Inc., Minneapolis, MN, USA), BIOVIA Discovery Studio (BIOVIA corp., San Diego, CA, USA), PACKMOL package (University of Campinas, Campinas, Brazil).

2.2. *Machine learning*

2.2.1. *Pseudo-ternary phase diagram dataset*

The pseudo-ternary phase diagram dataset was obtained from 45 diagrams in 25 reported literature. Each data included the information of excipient and aqueous solution. Then, the relative molecular descriptors, which were the main physicochemical properties of oils, surfactant, and cosurfactant selected as Table 1 shown. The data collection principle in pseudo-ternary phase diagrams was showed as Fig. 1. In the self-emulsion area, the coordinate point inside the intersection point between the self-emulsifying region and the grid lines were selected as the data points of self-emulsion. In the non-self-emulsifying area, the vertex of the grid line was selected as the data point of non-self-emulsion. Each selected point in the pseudo-ternary phase diagram was a single data point. The total dataset of SEDDS formulations were 4495.

2.2.2. *Machine learning methods in pseudo-ternary phase diagram*

Seven machine learning methods were utilized in the research of pseudo-ternary diagram, include random forests (RF), K-nearestNeighbor (KNN), decision Tree (DT), naïve Bayes (NB), support vector machines (SVM), Light Gradient Boosting Machine (lightGBM) and XGBoost. Shortly, DT is one of the commonly used machine learning methods. In dealing with classification issues, it is used as the tree-like structure model to classify instances based on features. It also can be thought of as a conditional probability distribution on characteristics and classification[27]. Furthermore, RF is a more advanced algorithm based on the decision tree. It was developed from classification and regression trees and determined by the mode of the category output by an individual tree[28]. KNN algorithm is a famous statistical pattern recognition, which plays a critical role in machine learning classification algorithms. Generally, the algorithm calculates the Euclidean distance between a point and all other points and extracts the K points closest to the end. Then count the K points with the largest proportion of the category, which belongs to the classification[29]. NB is based on Bayes' principle and uses probability statistics to classify the sample data set. It is based on subjective judgments: it is equally possible to estimate a value without knowing all the objective facts and then continually revise it based on the actual results[30]. SVM is a supervised machine learning model that uses classification algorithms to solve two sets of classification problems. It works as classifying the next text when the SVM model with labeled training data for each category has been given. LightGBM is a novel algorithm presented by Ke et al[31]., and it has been applied in various types of data mining tasks, such as classification, regression and ranking. XGBoost, developed by Chen et al.[32] in 2016, is a machine learning algorithm under the gradient boosting framework. XGBoost provides a parallel tree lift that can quickly and accurately solve many data science problems.

The classifier was trained with the training set by using 5-fold cross validation, and then the model was validated with the validation set to record the final classification accuracy. In brief, the initial dataset was divided into five sub-datasets, a single subdataset is retained for testing, and the other four datasets were used for training. The cross-validation is repeated five times, and each sub-dataset was verified once. The results of five times are averaged, or other combination methods are used to obtain a sole estimation. By comparing model performances of different algorithms, a best model will be obtained.

Table 1 The selected descriptors of three excipient.

Excipient	Selected molecular descriptor
Oils	Molecular weight, log*P*, boiling point, melting point, density, viscosity, HLB, flash point, surface tension, saponification value
Surfactants	Molecular weight, log*P* melting point, density, viscosity, HLB, flash point, saponification value
Cosurfactants	Molecular weight, log*P* melting point, density, viscosity, flash point

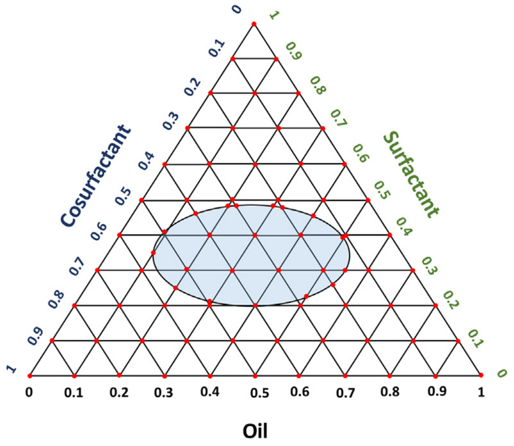


Figure 1 The pseudo-ternary phase diagram of SEDDS.

2.3. *Molecular dynamic simulation*

The structure of meloxicam, Labrafil M 1944 CS, Cremophor RH40, and Transcutol HP were constructed by BIOVIA Discovery Studio Visualizer 2016, as shown in Fig. 2. Next, the simulation box including drug molecules, excipients, and solvent water, was packed by PACKMOL[33]. The AMBER 18 software package performed the MD simulation process with the GAFF force field.

In the MD simulation, the process parameters were set concerning the experiment. Shortly, the simulation processes were divided into three steps. Firstly, the SEDDS system runs a sum of 2000 steps of energy minimization; 1000 steps of steep descent minimization, followed by 1000 steps of the conjugate gradient. Secondly, the whole system was heated from 0 to 310 K. Finally, the system was kept at 310 K for 200 ns to mimic the self-emulsification process for the formulation in water.

2.4. *Experimental validation*

2.4.1. *Solubility study*

An excess amount of MLX was added into 2 mL different excipients (oils, surfactants, cosurfactants) in a 5 mL microcentrifuge tube in triplicate. After vertexing, samples were shaken for 72 h in a constant temperature shaker at 37 ± 0.5 °C and then centrifuged at 12,000 rpm for 10 min (TG16MW centrifuge, Hunan Herexi Instrument & Equipment Co., Ltd.). Finally, the supernatant was diluted with methanol to the appropriate multiple and determined by UV-8000 UV−Vis spectrophotometer (Shanghai metash Instrument Co., Ltd., Shanghai, China).

2.4.2. *Pseudo-ternary phase diagram*

The surfactant and the cosurfactant were mixed in a certain mass ratio (4:1, 3:1, 2:1, 1:1, 1:2, and 1:3) to form a mixed emulsifier. Then weigh a certain proportion of the oil was added into the mixed emulsifier at the mass ratio of 1:9, 2:8, 3:7, 4:6, 5:5, 6:4, 7:3, 8:2, and 9:1. Subsequently, 1 g of the mixture was dropped into 100 mL of purified water, which was stirred magnetically at 37 °C. When the emulsion droplets can diffuse in water and form a homogeneous emulsion, the proportion point was marked as a self-emulsifying point.

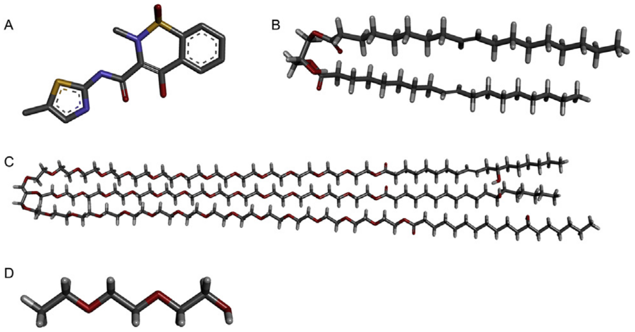


Figure 2 The molecular structure of (A) MLX, (B) Labrafil M 1944 CS, (C) Cremophor RH40 and (D) Transcutol HP.

2.4.3. *Formulation design of MLX-SEDDS*

The central composite design (CCD) was utilized to screen the optimal ratio for oil-surfactant-cosurfactant with two factors and five levels of optimization (shown in Table 2) by Design-Expert v10. X_1 (mass percent of oil) and X_2 (K_m = surfactant/cosurfactant) were set as independent variables, while Y_1 (droplet size), Y_2 (PDI), and Y_3 (drug loading) were assessed as test parameters. The determination method of droplet size and PDI was as followed: 0.5 mL of oil, surfactant, and cosurfactant mixed in a certain proportion was dissolved in 50 mL water in 37 °C and determined by Malvern Nano-ZS (Malvern Instruments, UK). The determination of drug loading was as followed. An excess amount of MLX was added into 2 mL mixture solution (oils, surfactants, cosurfactants) in a 5 mL microcentrifuge tube in triplicate. After vertexing, samples were shaken for 72 h in a constant temperature shaker at 37 ± 0.5 °C and then centrifuged (TG16MW, Hunan Herexi Instrument & Equipment Co., Ltd.) at 12,000 rpm for 10 min. Finally, the supernatant was diluted with methanol to the appropriate multiple and determined by UV-8000 UV−Vis spectrophotometer (Shanghai metash Instrument Co., Ltd. Shanghai, China).

3. Results and discussion

3.1. *Dataset distribution*

The distribution of oils, surfactants and cosurfactants in the dataset was exhibited in Fig. 3. The most used oil in the dataset was Capryol 90 occupied for 31.03% and the main content was propylene glycol caprylate. Three commonly used surfactant were Cremophor RH40, Cremophor EL and Tween 80, respectively. Among them, Cremophor RH40 and Cremophor EL were PEG-modified hydrogenated castor oil that was non-ionic surfactants considered non-ionic surfactants. In cosurfactant, almost half of the data containing Transcutol HP was diethylene glycol monoethyl ether.

Table 2 Levels of independent variable in the central composite design.

Factor	Level				
	−1.414	−1	0	+1	+1.414
X_1	9.64	20	45	70	80.36
X_2	0.38	1	2.5	4	4.62

X_1, mass percent of oil; X_2, K_m = Surfactant/Cosurfactant.

3.2. *Prediction model by different machine learning algorithms*

Table 3 shows classification performance for the evaluation of the self-emulsification in each oil, surfactant and cosurfactant combination. Both RF and XGBoost manifest the good performance while NB showed the worst performance in the self-emulsifying area predicting. However, the RF exhibited the optimal balance between sensitivity and specificity in test set. RF showed the best prediction performance with 91.3% for accuracy, 92.0% for sensitivity and 90.7% for specificity in 5-fold cross-validation and 93.0% for accuracy, 91.7% for sensitivity and 94.7% for specificity in test set. These results illustrated that the RF was the most suitable algorithm in self-emulsifying area prediction.

3.3. *Importance features ranking by random forest*

The important features were calculated by the optimal machine learning algorithm RF. These 27 important features could be divided into four categories: the properties of oil, surfactant, cosurfactant, and aqueous solution. As shown in Fig. 4, the top 10 most important features with a contribution value large than 0.5 were Dose_OIL, Dose_SUR, Molecular_weight_OIL, Flash_point_OIL, SEDDS concentration, Dose_COS, Saponification_OIL, HLB_OIL, Solution_pH, and HLB_SUR.

In the SEDDS formulation design, the selection of oils, surfactants and cosurfactants was on the basis of the pseudo ternary phase diagram. The proportion of oil, surfactant and cosurfactant was the most essential parameter in forming self-emulsion ranked the first, the second and the sixth. As the two most essential features, the oil and surfactant ratio played almost the same central role in this prediction model. The oils were the leading excipients in the SEDDS due to their excellent capability of lipophilic drugs and increased permeability for drugs[34]. Pouton[35] presented the lipid formulation classification system, which divided the lipid-based formulations into four types in 2000. There were two types of SEDDS: type IIIA and type IIIB. The difference between type IIIA and type IIIB was the amount of the oil phase. The type IIIA formulations were oily with 40%−80% oils, while the type IIIB formulations were water soluble with no higher than 20% oil. In general, the greater the proportion of the oil phase, the smaller the self-emulsification area on the pseudo ternary phase diagram. The surfactant functions in forming self-emulsion were promoted oil and water to form a homogeneous and stable mixture and prevent the precipitation of the drugs in the gastrointestinal tract with the dissolved state. As a result, the proportion of the surfactants exhibit almost the same feature importance value as the proportion of the oils. When increasing the ratio of emulsifiers, it benefited from forming a smaller droplet size formulation self-micro emulsifying drug delivery system.

The properties of oil had also displayed specific importance in SEDDS formulation design with the molecular weight (3rd), flash point (4th), saponification value (7th), and HLB value (8th). There are 15 kinds of oils in the dataset and the main content of these oils can be divided into two catalogs: glycerolipids and fatty acyls. Among glycerolipids, long (LCT) and medium-chain triglyceride (MCT) oils were the commonly used oils in SEDDS. The molecular weight and flash point were considered as the two most important properties in the prediction model. The molecular

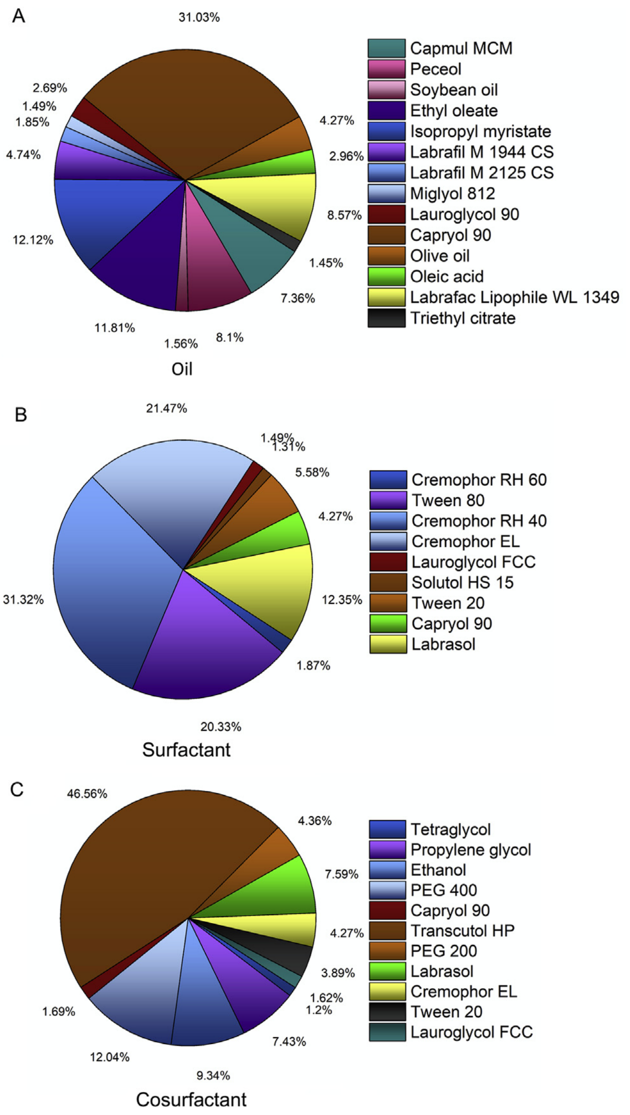


Figure 3 Distribution of excipients in the dataset: (A) oil, (B) surfactant, (C) cosurfactant.

weight of oils reflected the number of carbons and the chain's length in oils, which was the essential feature of oils. The flash point represented the lowest temperature at which fire ignites at a specific vapor pressure.

The saponification value (SV) represented the number of milligrams of KOH needed to saponify one gram of lipids, which can be calculated by Eq. (1):

$$\mathrm{SV} = N \times 1000 \times \frac{56.1}{\mathrm{MW}} \tag{1}$$

where N was the number of fatty acid residues in 1-mol lipid, 56.1 was the molecular weight of KOH, and MW was the molecular weight of the lipid.

Experimentally measured SV was relative to the fatty acids residues numbers and MW of the lipids. Hence, the SV of MCT was often higher than that of LCT when they had the same number of fatty acid residues. Similarly, triglycerides had higher SV than contain fatty acids with the same carbon number of diacylglycerols and monoacylglycerols. Also, SV could help evaluate the main lipid in the mixed oils. In reported researches of SEDDS, the HLB value of the surfactants was more concerned than that of the oils. In fact, HLB value of the oils was more critical than that of the surfactants. The HLB value of mixed compounds can be represented by Eq. (2):

$$\mathrm{HLB} = \frac{(C_1 \times \mathrm{HLB}_1) + (C_2 \times \mathrm{HLB}_2) + (C_3 \times \mathrm{HLB}_3)\cdots}{C_{\mathrm{total}}} \tag{2}$$

where C_1, C_2, and C_3 were the proportion of each component, respectively.

The HLB value of SEDDS was a significant factor in the evaluation of self-emulsification[35].

Furthermore, the properties of the aqueous solution were also had a particular impact on the SEDDS. Some researchers determined the area of SEDDS by the titration method. The pseudo-ternary phase diagram indicated a bigger self-emulsion region at the lower concentration of SEDDS in water[36–38]. Another way to draw a three-phase diagram is to fix the concentration of SEDDS and plot the three phases of oil, surfactant and cosurfactant, respectively[39–41].

3.4. *The solubility of MLX in excipient*

The solubility of MLX in different oils, surfactants, cosurfactants was determined to ensure optimal drug loading. The solubilities of MLX in excipients were shown in Table 4. Among the excipient of oils, Labrafil M 1944 CS exhibited significantly higher solubility than IPM, IPP, and GTCC. Thus, Labrafil M 1944 CS was chosen as the appropriate oily excipient. Similarly, Tween 80 showed the highest soluble capacity among the surfactants. Besides, AEO-9 and Cremophor RH40 also displayed a suitable soluble ability of MLX, so these two surfactants were selected as the candidate surfactant. The three cosurfactants with the highest solubility of MLX were Transcutol HP, ethanol, and PEG 400, respectively. Since ethanol is readily volatile, MLX may be at risk of precipitation in SEDDS. Thus, PEG 400 and Transcutol HP were chosen as the alternative cosurfactant.

3.5. *The comparison between experimental and predicting SEDDS region*

The prediction points of the three-phase diagram are consistent with the experiment in 2.3.2, and the data was predicted by RF algorithms as an external test. Subsequently, to evaluate the accuracy of the prediction, we used experimental methods to verify. The predicting SEDDS region was showed in Table 5, resulting in 89.51% accuracy. The sensitivity was defined as the sensitivity of the method to positive, while the specificity was defined as the ability of this method to estimate negative. The imbalance between sensitivity (67.91%) and specificity (99.23%) was most likely due to more negative results than positive results in the external experimental data set. Fig. 5 shows the predicting SEDDS region for each oil–surfactant–cosurfactant combination. In Fig. 5A, E and B, F, the predicting SEDDS region was the same. However, the results show that the predicting accuracy of Labrafil M 1944 CS and Tween 80 system was slightly higher than that of Labrafil M 1944 CS and AEO-9 system. In the system of Labrafil M 1944 CS and Cremophor RH40, the accuracy was 75.93% and 85.19% with Transcutol HP and PEG 400, respectively. The optimal ternary phase combination was Labrafil M 1944 CS, Cremophor RH40 and Transcutol HP (Fig. 5C), which exhibited the largest SEDDS region.

Table 3 The classification performance results for the evaluation of the self-emulsification.

Method	5-Fold cross-validation				Test			
	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC (%)
RF	91.3	92.0	90.7	96.4	93.0	91.7	94.0	97.4
KNN	88.6	93.1	85.0	94.4	88.4	95.9	82.7	94.8
DT	90.2	91.0	89.6	93.5	90.4	89.3	90.9	92.6
NB	61.8	68.5	56.5	71.3	65.4	71.7	60.7	72.7
SVM	82.3	81.7	82.8	90.8	86.3	87.5	85.4	93.6
LightGBM	80.5	62.4	95.1	92.6	81.9	65.1	94.8	93.8
XGBoost	92.0	88.1	95.1	97.9	91.5	86.1	95.5	98.1

RF, random forests; KNN, K-nearestNeighbor; DT, decision Tree; NB, naïve Bayes; SVM, support vector machines; LightGBM, Light Gradient Boosting Machine.

Fig. 6A–F exhibits the pseudo-ternary phase diagram constructed by the experimental method. The largest self-emulsifying region was observed when Cremophor RH40 and Transcutol HP were surfactant and cosurfactant (Fig. 6C). The experimental result showed the same tendency as the prediction result. Thus, experiment results have extensively validated SEDDS region predicting results.

3.6. *Formulation optimization of MLX-SEDDS by central composite design*

By predicting the pseudo three-phase diagram, we found the optimal combination (oil: Labrafil M 1944 CS, surfactant: Cremophor RH40, cosurfactant: Transcutol HP). To further explore the three-phase ratio, we used the central composition design to ensure the parameters of MLX-SEDDS. The result of the MLX-SEDDS formulation constructed by the central composite design is showed in Table 6. The relationship between independent variables (X_1: oil content, X_2: K_m = Surfactant/Cosurfactant, m/m) and test parameters (Y_1: droplet size, Y_2: PDI, Y_3: drug loading) were exceptionally fitted by the third-order polynomial model with the coefficient of determination (R^2) for 0.9692, 0.9420, 0.9582 and all of the significant fitting values (P) were less than 0.05. Therefore, the contour plots (shown in Fig. 7) were drawn in accordance with the followed fitting Eqs. (3)–(5):

$$Y_1 = -202.4716 + 16.2372X_1 + 77.8642X_2 - 6.9546X_1X_2 - 0.1489X_1^2 + 1.1204X_2^2 + 0.05945X_1^2X_2 + 0.3161X_1X_2^2 \tag{3}$$

$$Y_2 = 0.2486 - 0.003861X_1 - 0.1493X_2 + 0.001027X_1X_2 + 0.000037X_1^2 + 0.04086X_2^2 + 0.000043X_1^2X_2 - 0.000959X_1X_2^2 \tag{4}$$

$$Y_3 = 1.7659 - 0.01395X_1 + 0.1457X_2 - 0.0015X_1X_2 - 0.000385X_1^2 - 0.001811X_2^2 + 0.000048X_1^2X_2 - 0.000674X_1X_2^2 \tag{5}$$

As Fig. 7A shown, the droplet size was more extensive with the increase of the oil amount. Furthermore, K_m also has a certain effect on the droplet size. As the experimental result in Table 6 (Nos. 4—5), the particle size increases significantly with the increase in the concentration of cosurfactant. The cosurfactant effect could reduce the interfacial tension and change the curvature of the emulsion droplet[42]. In Fig. 7B, the relationship between oil content, K_m, and PDI showed the same tendency. When the oil content in the system increased, the PDI value also increased. That mean the droplet size of the system was more homogeneous, under the condition of less oil phase content. The result indicated that the surfactant disperses the mixed-phase into tiny droplets and reduces the interface's free energy to maintain a smaller surface area. However, there was a contrary tendency between oil content,

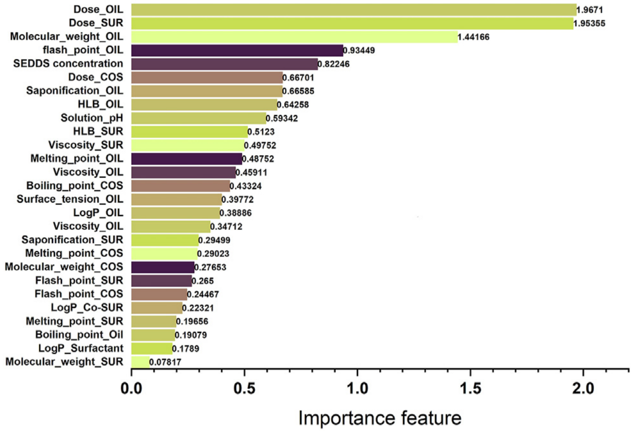


Figure 4 The importance feature of classification in SEDDS ranking by RF.

Table 4 Solubility of MLX in excipients.

Category	Excipient	Solubility (mg/mL)
Oils	IPM	0.151
	IPP	0.172
	GTCC	0.244
	Labrafil M 1944 CS	1.072
Surfactants	Tween 80	11.902
	Cremophor EL	4.423
	AEO-9	7.194
	Cremophor RH40	7.188
Co-surfactants	PEG400	1.765
	Transcutol HP	3.701
	Ethanol	3.351
	Isopropanol	0.278

IPM, isopropyl myristate; IPP, isopropyl palmitate, GTCC, caprylic/capric triglyceride.

Table 5 The comparison in accuracy, sensitivity and specificity of six SEDDS combination.

Diagram	Accuracy (%)	Sensitive (%)	Specificity (%)
A	92.59	71.42	100
B	98.15	83.33	100
C	75.93	50	100
D	85.19	60	100
E	90.74	72.73	95.35
F	94.44	70	100
Overall	89.51	67.91	99.23

K_m, and drug loading than droplet size and PDI. The loading capacity of SEDDS was calculated by Eq. (6) [43]:

$$C_{SEDDS} = \sum W_e C_e \tag{6}$$

where C_{SEDDS} was the solubility of the SEDDS system, W_e was the content of each excipient, and C_e was the solubility of each excipient. According to Eq. (6), the solubility of MLX in the oil phase was relatively lower than in surfactant and cosurfactant, so that the drug loading decreased when the oil content increased. Overall, among 13 formulations, No. 7 had the most extensive loading, relatively small droplet size and suitable PDI so that it

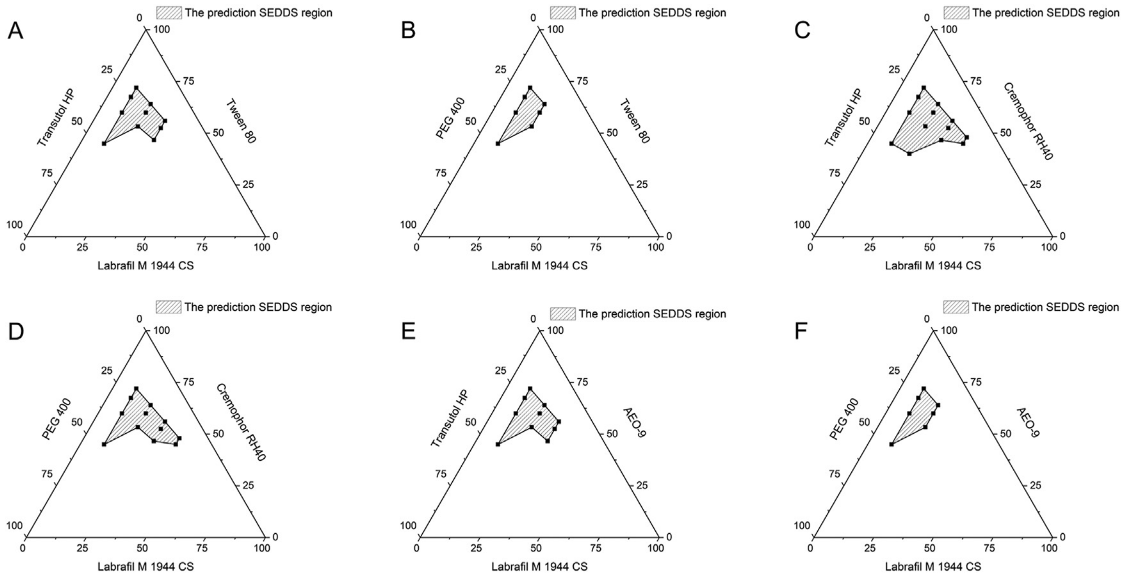


Figure 5 The predicting pseudo-ternary phase diagram with different oil: surfactant: co-surfactant combination: (A) Labrafil M 1944 CS: Tween 80: Transcutol HP, (B) Labrafil M 1944 CS: Tween 80: PEG 400, (C) Labrafil M 1944 CS: Cremophor RH40: Transcutol HP, (D) Labrafil M 1944 CS: Cremophor RH40: PEG 400, (E) Labrafil M 1944 CS: AEO-9: Transcutol HP, (F) Labrafil M 1944 CS: AEO-9: PEG 400.

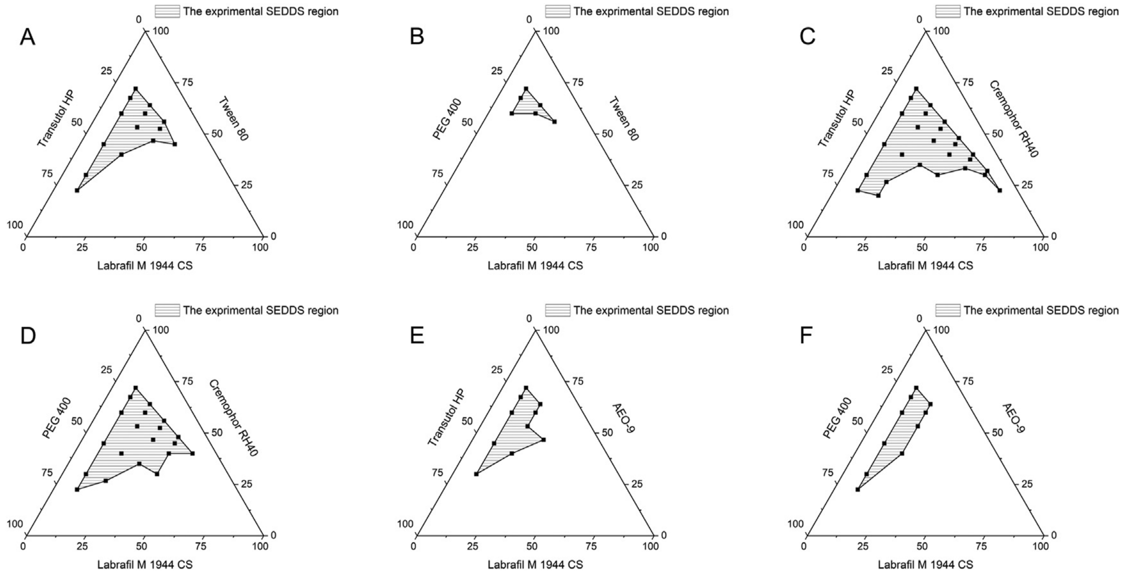


Figure 6 The experimental pseudo-ternary phase diagram with different oil: surfactant: co-surfactant combination: (A) Labrafil M 1944 CS: Tween 80: Transcutol HP, (B). Labrafil M 1944 CS: tween 80: PEG 400, (C) Labrafil M 1944 CS: Cremophor RH40: Transcutol HP, (D) Labrafil M 1944 CS: Cremophor RH40: PEG 400; (E) Labrafil M 1944 CS: AEO-9: Transcutol HP, (F) Labrafil M 1944 CS: AEO-9: PEG 400.

Table 6 The central composition design of MLX-SEDDS.

No.	X_1 (%)	X_2	Y_1 (nm)	Y_2	Y_3 (mg/mL)
1	80.36	2.50	74.64	0.51	0.77
2	9.64	2.50	17.83	0.11	4.58
3	70.00	4.00	95.28	0.23	1.41
4	45.00	0.38	196.17	0.18	2.77
5	45.00	4.62	35.03	0.06	2.98
6	45.00	2.50	35.80	0.10	2.99
7	20.00	4.00	21.31	0.05	4.83
8	45.00	2.50	35.36	0.10	2.90
9	20.00	1.00	22.02	0.06	4.09
10	70.00	1.00	99.51	0.23	1.20
11	45.00	2.50	35.53	0.08	3.97
12	45.00	2.50	35.75	0.09	4.06
13	45.00	2.50	36.85	0.10	3.31

X_1, mass percent of oil; X_2, K_m = Surfactant/Cosurfactant. Y_1, droplet size; Y_2, PDI; Y_3, drug loading.

was considered the optimal formulation. Thus, the optimal content of Labrafil M 1944 CS, Cremophor RH40, and Transcutol HP was 20%, 64% and 16%, respectively.

3.7. *Molecular modeling for MLX-SEDDS*

The diffusion process of MLX-SEDDS in water was mimicked by the MD simulation and showed in Fig. 8. After 200 ns simulation, Oils (white molecules) and surfactants (green molecules) were formed the droplet skeleton of MLX-SEDDS. And the cosurfactants (blue molecules) were dispersed around the droplet. To further explore the cosurfactant role in the system, we had constructed a self-emulsifying system without cosurfactant for comparison. Fig. 9A presentes the root-mean-square deviation (RMSD) of MLX-SEDDS and MLX-SEDDS without cosurfactant for 200 ns? Obviously, the MLX-SEDDS system produced for about 3 Å fluctuation while the MLX-SEDDS without cosurfactant system had about 18 Å fluctuation. This result illustrated that cosurfactant could make the whole system more stable. The cosurfactant could decrease the interface tension between oil and water to replace part of surfactants to reduce the side effects caused by surfactants. The mass-weight radius of gyration (R_g) and solvent accessible surface areas (SASA) analyzed the molecule movement of two systems in an aqueous solution, shown in Fig. 9B and C. R_g represented the distribution of molecules in aqueous solution over time while SASA was defined as the surface area accessibility of molecules to the solvent. Both R_g and SASA curves of the system with cosurfactant were higher than those of the system without cosurfactant, indicating the former system easier to self-emulsify than the later system in water. This result illustrated that cosurfactant could improve the emulsion of SEDDS and was well consistent with the observation of the experiment. Therefore, the cosurfactant effect may result in a larger droplet size of SEDDS.

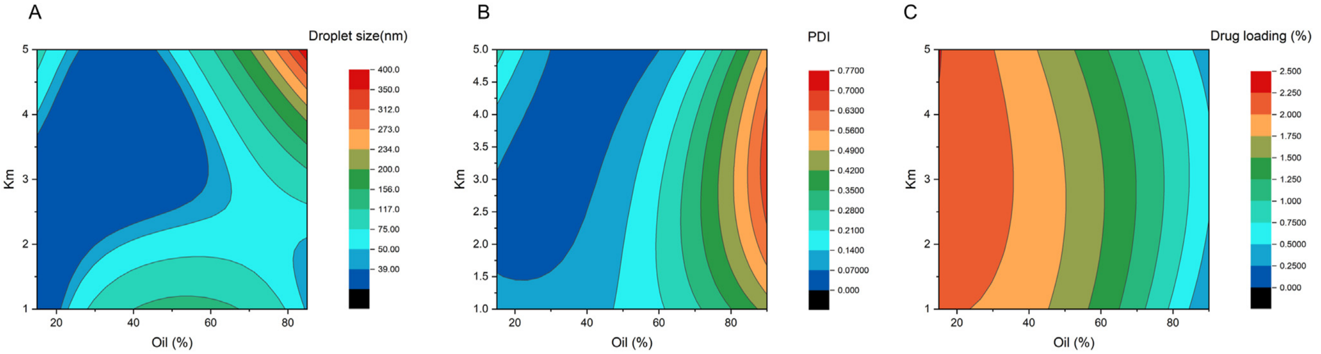


Figure 7 The color fills of (A) droplet size, (B) PDI, (C) drug loading.

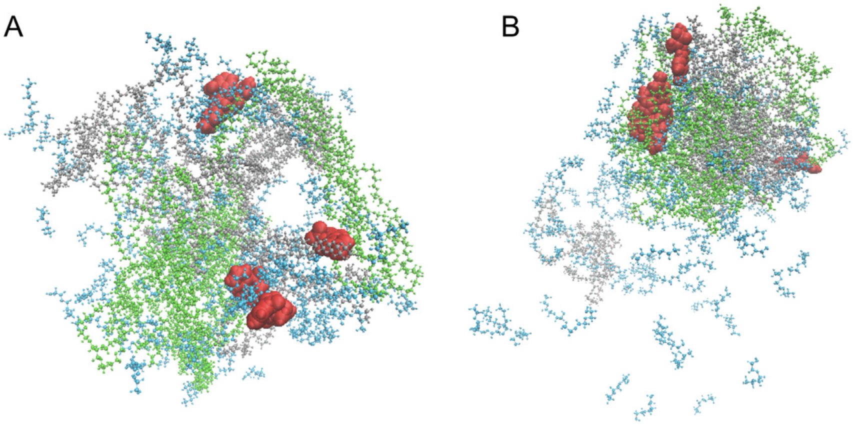


Figure 8 Snapshots of MLX-SEDDS in water (A) 0 and (B) 200 ns (Red molecules: MLX; grey molecules: oils; green molecules: surfactants; blue molecules: cosurfactants).

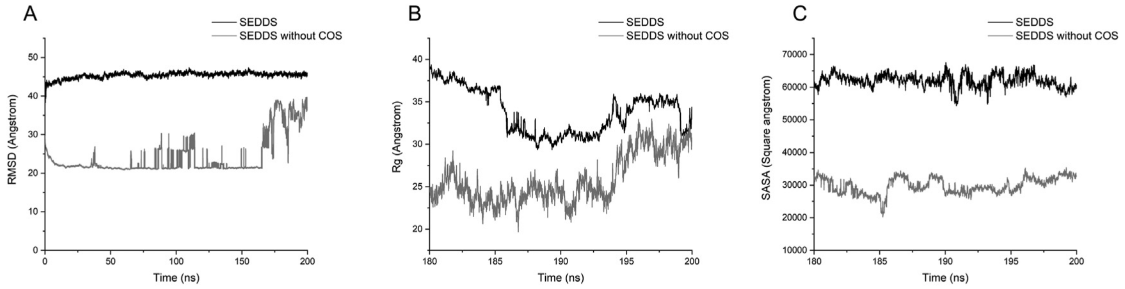


Figure 9 CPPTRAJ analysis results of MLX-SEDDS. (A) Root-mean-square deviation of MLX-SEDDS and MLX-SEDDS without cosurfactant. (B) Mass-weight radius of gyration of MLX-SEDDS and MLX-SEDDS without cosurfactant. (C) Solvent accessible surface areas of MLX-SEDDS and MLX-SEDDS without cosurfactant.

4. Conclusions

In this research, the pseudo-ternary phase diagram prediction model was successfully constructed by the RF method, which also revealed the important features in SEDDS. The MLX pseudo-ternary phase diagram by experiments validated the prediction model with 89.51% accuracy. The CCD experimental design helped find the optimal MLX-SEDDS and revealed the relationship between excipient content and SEDDS' s properties. Finally, the MD simulation provided us the molecular interaction between drug and excipient and the role of cosurfactant. The integrated *in silico* and experimental methodology are well applied in rational formulation design of SEDDS, which also brings new ideas for future drug formulation design.

Acknowledgments

This work was financially supported by the Science and Technology Development Fund (FDCT) of Macau (0029/2018/A1, China) and the University of Macau Research Grants (MYRG2019-00041-ICMS, China). This work was performed in part at the High-Performance Computing Cluster (HPCC) which is supported by Information and Communication Technology Office (ICTO) of the University of Macau, China.

Author contributions

Haoshi Gao contributed to the molecular dynamic simulation, data processing and drafting of the manuscript. Haoyue Jia contributed to the experiments. Jie Dong contributed to machine learning and manuscript revision. Haifeng Li, Xinggang Yang and Defang Ouyang contributed to conception of the work and manuscript revision.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Munzone E, Colleoni M. Clinical overview of metronomic chemotherapy in breast cancer. *Nat Rev Clin Oncol* 2015;**12**:631−44.
2. Porter CJH, Trevaskis NL, Charman WN. Lipids and lipid-based formulations: optimizing the oral delivery of lipophilic drugs. *Nat Rev Drug Discov* 2007;**6**:231−48.
3. Ghadi R, Dand N. BCS class IV drugs: highly notorious candidates for formulation development. *J Control Release* 2017;**248**:71−95.
4. Danda LJdA, Batista LdM, Melo VCS, Soares Sobrinho JL, Soares MFdLR. Combining amorphous solid dispersions for improved kinetic solubility of posaconazole simultaneously released from soluble PVP/VA64 and an insoluble ammonio methacrylate copolymer. *Eur J Pharmaceut Sci* 2019;**133**:79−85.
5. Celebioglu A, Uyar T. Metronidazole/hydroxypropyl-β-cyclodextrin inclusion complex nanofibrous webs as fast-dissolving oral drug delivery system. *Int J Pharm* 2019;**572**:118828.
6. Jeevanandam J, Chan YS, Danquah MK. Nano-formulations of drugs: recent developments, impact and challenges. *Biochimie* 2016;**128−129**:99−112.
7. Chatterjee B, Hamed Almurisi S, Ahmed Mahdi Dukhan A, Mandal UK, Sengupta P. Controversies with self-emulsifying drug delivery system from pharmacokinetic point of view. *Drug Deliv* 2016;**23**:3639−52.
8. Mishra V, Nayak P, Yadav N, Singh M, Tambuwala MM, Aljabali AAA. Orally administered self-emulsifying drug delivery system in disease management: advancement and patents. *Expet Opin Drug Deliv* 2021;**18**:315−32.
9. Kommuru TR, Gurley B, Khan MA, Reddy IK. Self-emulsifying drug delivery systems (SEDDS) of coenzyme Q10: formulation development and bioavailability assessment. *Int J Pharm* 2001;**212**:233−46.
10. Balakrishnan P, Lee BJ, Oh DH, Kim JO, Hong MJ, Jee JP, et al. Enhanced oral bioavailability of dexibuprofen by a novel solid Self-emulsifying drug delivery system (SEDDS). *Eur J Pharm Biopharm* 2009;**72**:539−45.
11. Damiati SA. Digital pharmaceutical sciences. *AAPS PharmSciTech* 2020;**21**:206.
12. Natalie S, Emily S, Jessica C, Jason R, David R, Nicola J, et al. Survey of machine learning techniques in drug discovery. *Curr Drug Metabol* 2019;**20**:185−93.
13. Wang T, Wu MB, Lin JP, Yang LR. Quantitative structure−activity relationship: promising advances in drug discovery platforms. *Expet Opin Drug Deliv* 2015;**10**:1283−300.
14. Shayanfar A, Fakhree MAA, Jouyban A. A simple QSPR model to predict aqueous solubility of drugs. *J Drug Deliv Sci Technol* 2010;**20**:467−76.
15. Tian XY, Chen DZ, Gao J. An overview on protein fold classification *via* machine learning approach. *Curr Proteonomics* 2018;**15**:85−98.
16. Zhao QQ, Ye ZYF, Su Y, Ouyang DF. Predicting complexation performance between cyclodextrins and guest molecules by integrated machine learning and molecular modeling techniques. *Acta Pharm Sin B* 2019;**9**:1241−52.
17. Han R, Xiong H, Ye ZYF, Yang YL, Huang TH, Jing QF, et al. Predicting physical stability of solid dispersions by machine learning techniques. *J Control Release* 2019;**311−312**:16−25.

18. He Y, Ye ZYF, Liu XY, Wei ZJ, Qiu F, Li HF, et al. Can machine learning predict drug nanocrystals?. *J Control Release* 2020;**322**:274−85.
19. Gao HS, Ye ZYF, Dong J, Gao HL, Yu H, Li HF, et al. Predicting drug/phospholipid complexation by the lightGBM method. *Chem Phys Lett* 2020;**747**:137354.
20. Ouyang DF, Smith SC. *Introduction to computational pharmaceutics. Computational pharmaceutics.*. New Jersey: John Wiley & Sons, Ltd.; 2015. p. 1−5.
21. Bunker A, Róg T. Mechanistic understanding from molecular dynamics simulation in pharmaceutical research 1: drug delivery. *Front Mol Biosci* 2020;**7**:604770.
22. Ouyang DF. Investigating the molecular structures of solid dispersions by the simulated annealing method. *Chem Phys Lett* 2012;**554**:177−84.
23. Gao HL, Wang W, Dong J, Ye ZYF, Ouyang DF. An integrated computational methodology with data-driven machine learning, molecular modeling and PBPK modeling to accelerate solid dispersion formulation design. *Eur J Pharm Biopharm* 2021;**158**:336−46.
24. Zhao QQ, Miriyala N, Su Y, Chen WJ, Gao XJ, Shao L, et al. Computer-aided formulation design for a highly soluble lutein−cyclodextrin multiple-component delivery system. *Mol Pharm* 2018;**15**:1664−73.
25. Wilkhu JS, Ouyang DF, Kirchmeier MJ, Anderson DE, Perrie Y. Investigating the role of cholesterol in the formation of non-ionic surfactant based bilayer vesicles: thermal analysis and molecular dynamics. *Int J Pharm* 2014;**461**:331−41.
26. Yang CL, Tu K, Gao HL, Zhang L, Sun Y, Yang T, et al. The novel platinum(IV) prodrug with self-assembly property and structure-transformable character against triple-negative breast cancer. *Biomaterials* 2020;**232**:119751.
27. Yu Z, Haghighat F, Fung BCM, Yoshino H. A decision tree method for building energy demand modeling. *Energy Build* 2010;**42**:1637−46.
28. Burgette LF, Reiter JP. Multiple imputation for missing data *via* sequential regression trees. *Am J Epidemiol* 2010;**172**:1070−6.
29. Du MJ, Ding SF, Jia HJ. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowl-Based Syst* 2016;**99**:135−45.
30. Li T, Li J, Liu ZL, Li P, Jia CF. Differentially private Naive Bayes learning over multiple data sources. *Inf Sci* 2018;**444**:89−104.
31. Ke GL, Meng Q, Finley T, Wang TF, Chen W, Ma WD, et al. LightGBM: a highly efficient gradient boosting decision tree. Neural Imformation Processing Systems 2017; 2017 Dec 4−7; Long Beach. In: *Proceedings of the 31st international conference on neural information processing systems of downloaded from*. California, USA: Curran Associates Inc.; 2017. https://dl.acm.org/at Long Beach Convention Center on December 4.
32. Chen TQ, Guestrin C. XGBoost: a scalable tree boosting system. San Francisco Union Square on August 13, Knowledge Discovery and Data Mining 2016; 2016 Aug 13−17. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining of download from*. San Francisco, California, USA: Association for Computing Machinery; 2016. https://dl.acm.org/at Hilton.
33. Martínez JM, Martínez L. Packing optimization for automated generation of complex system's initial configurations for molecular dynamics and docking. *J Comput Chem* 2003;**24**:819−25.
34. Neslihan Gursoy R, Benita S. Self-emulsifying drug delivery systems (SEDDS) for improved oral delivery of lipophilic drugs. *Biomed Pharmacother* 2004;**58**:173−82.
35. Pouton CW. Lipid formulations for oral administration of drugs: non-emulsifying, self-emulsifying and 'self-microemulsifying' drug delivery systems. *Eur J Pharmaceut Sci* 2000;**11**:S93−8.
36. Lee JH, Kim HH, Cho YH, Koo TS, Lee GW. Development and evaluation of raloxifene-hydrochloride-loaded supersaturatable SMEDDS containing an acidifier. *Pharmaceutics* 2018;**10**:78.
37. Czajkowska-Kośnik A, Szekalska M, Amelian A, Szymańska E, Winnicka K. Development and evaluation of liquid and solid self-emulsifying drug delivery systems for atorvastatin. *Molecules* 2015;**20**:21010−22.
38. Tung NT, Tran CS, Pham TMH, Nguyen HA, Nguyen TL, Chi SC, et al. Development of solidified self-microemulsifying drug delivery systems containing L-tetrahydropalmatine: design of experiment approach and bioavailability comparison. *Int J Pharm* 2018;**537**:9−21.
39. Yeom DW, Song YS, Kim SR, Lee SG, Kang MH, Lee S, et al. Development and optimization of a self-microemulsifying drug delivery system for atorvastatin calcium by using D-optimal mixture design. *Int J Nanomed* 2015;**10**:3865−77.
40. Kim DW, Kwon MS, Yousaf AM, Balakrishnan P, Park JH, Kim DS, et al. Comparison of a solid SMEDDS and solid dispersion for enhanced stability and bioavailability of clopidogrel napadisilate. *Carbohydr Polym* 2014;**114**:365−74.
41. Truong DH, Tran TH, Ramasamy T, Choi JY, Lee HH, Moon C, et al. Development of solid self-emulsifying formulation for improving the oral bioavailability of erlotinib. *AAPS PharmSciTech* 2016;**17**:466−73.
42. Rahman MA, Hussain A, Hussain MS, Mirza MA, Iqbal Z. Role of excipients in successful development of self-emulsifying/microemulsifying drug delivery system (SEDDS/SMEDDS). *Drug Dev Ind Pharm* 2013;**39**:1−19.
43. Alskär LC, Porter CJH, Bergström CAS. Tools for early prediction of drug loading in lipid-based formulations. *Mol Pharm* 2016;**13**:251−61.