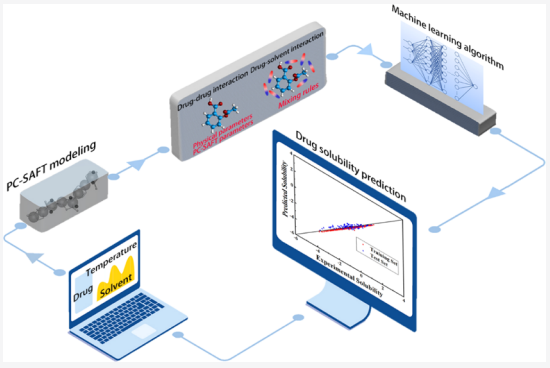
I&EC research
Industrial & Engineering Chemistry Research
pubs.acs.org/IECR
Article

Novel Computational Approach by Combining Machine Learning with Molecular Thermodynamics for Predicting Drug Solubility in Solvents

Kai Ge and Yuanhui Ji*

Cite This: *Ind. Eng. Chem. Res.* 2021, 60, 9259−9268

ABSTRACT: In this work, a novel strategy that combined molecular thermodynamic and machine learning was proposed to accurately predict the solubility of drugs in various solvents. The strategy was based on 16 molecular descriptors representing drug−drug interactions and drug−solvent interactions including physical parameters, pure perturbed-chain statistical associating fluid theory (PC-SAFT) parameters of drugs and solvents, and mixing rules. These molecular descriptors were inputted into five machine learning algorithms [multiple linear regression (MLR), artificial neural network (ANN), random forest (RF), extremely randomized trees (ET), and support vector machine (SVM)] to train the predictive model. A single-hidden-layer neural network was finally determined as the predictive model for predicting the solubility of drugs in various solvents. The drug solubility in the generalization evaluation set has also been successfully predicted, which indicates the good prediction performance of the model. Three directions for improving the model were summarized as adding molecular descriptors of drug−solvent interactions in the water system and drug−drug interactions in the organic solvent system and expanding the dataset to adequately obtain the features of multiple drugs. These findings show that the proposed model has the capability of solubility prediction, which is expected to provide important information for drug development and drug solvent screening.



1. INTRODUCTION

The solubility of drugs is essential for the discovery, preparation, and delivery of drugs and their formulations.[1] More and more drug candidates fail to be commercialized due to their poor aqueous solubility, which leads to a low dissolution rate and low bioavailability. In addition, the aqueous solubility of drugs can also significantly influence absorption, distribution, metabolism, excretion, toxicity, and pharmaceutical formulation.[2] Therefore, large numbers of methods, such as cosolvents, solubilizing agents, salt formation, micronization, nanocrystals, deep eutectic solvents, solid dispersion, etc., have been proposed to solve the problem of poor aqueous solubility.[3,4] Meanwhile, the solubility of drugs in organic solvents is also of vital importance, especially in aspects of drug separation, production and purification, and preparation of pharmaceutical formulation.[5,6] However, the determination of drug solubility often relies on experiments, which takes huge amounts of materials, manpower, and costs as most drugs are quite expensive. Accurate prediction of drug solubility in different solvents is the goal that is pursued by scientists and researchers.

To obtain the solubility of drugs in different solvents and at different temperatures more conveniently, the empirical models such as the Apelblat model and the λh model have been widely used in pharmaceutical research, but they have limited applicability in the absence of the empirical parameters regressed from a large number of necessary experimental data.[7] Solid−liquid equilibrium theory is another effective method to calculate the solubility as shown in eq 1, where the activity coefficient can be estimated by the activity coefficient model, such as the Wilson equation,[8] UNIFAC equation,[9] and UNIQUAC equation.[10] Recently, perturbed-chain statistical associating fluid theory (PC-SAFT) has also been successfully applied to describe the solid−liquid equilibrium of drug-related systems and calculate the thermodynamic properties of the systems accurately.[11−15] Furthermore, the Hansen solubility parameter,[7,16] molecular dynamics simulations,[17,18] COSMO-SAC,[19] and other methods are also proposed to achieve solubility prediction. Although many theoretical advances have been made, it is still difficult to achieve an accurate and

Received: March 14, 2021
Revised: May 12, 2021
Accepted: June 4, 2021
Published: June 18, 2021

https://doi.org/10.1021/acs.iecr.1c00998
Ind. Eng. Chem. Res. 2021, 60, 9259−9268

Table 1. Drug List of the Solubility Data Collected in the Literature

(+)/(−)-mandelic acid	aceclofenac	allisartan isoproxil	allopurinol
antipyrine	artemisinin	aspirin	azithromycin
benzamide	benzocaine	benzoic acid	bezafibrate
bifonazole	caffeine	candesartan cilexetil	capecitabine
captopril	carbamazepine	carvedilol	cefixime trihydrate
chlorpropamide	chlorzoxazone	cinnarizine	clotrimazole
dabigatran etexilate mesylate	dimethyl sulfone	dipyridamole	domperidone
doxofylline	edaravone	ethyl vanillin	febuxostat
felodipine	florfenicol	flurbiprofen	ganciclovir
gefitinib	glibenclamide	griseofulvin	hydrochlorothiazide
hydrocortisone	ibuprofen	indomethacin	isobutyramide
isoniazid	itraconazole	ketoconazole	ketoprofen
L-alanine	lamotrigine	lenalidomide	levetiracetam
L-glutamic acid	lidocaine	L-leucine	loratadine
L-proline	L-valine	maraviroc	mefenamic acid
mesalazine	milrinone	naproxen	nicotinamide
nicotinic acid	nifedipine	nitrendipine	osimertinib
oxcarbazepine	paclitaxel	*p*-aminobenzenesulfonamide	paracetamol
phenylbutazone	pindolol	piracetam	piroxicam
praziquantel	prednisolone	probenecid	propylthiouracil
pyrazinamide	quetiapine fumarate	rivaroxaban	roflumilast
saccharin	salicylamide	salicylic acid	simvastatin
spironolactone	succinic acid	sulfadimidine	sulfamerazine
sulfathiazole	sulpiride	terfenadine	theophylline
thiabendazole	triamterene	trimethoprim	xylitol
zaltoprofen	zonisamide	*β*-lapachone	

quantitative prediction of drug solubility as a function of temperature in the full absence of experimental data.[20]

$$x_{\mathrm{Drug}}^{\mathrm{L}} = \frac{1}{\gamma_{\mathrm{Drug}}^{\mathrm{L}}}\exp\left\{-\frac{\Delta h_{0\mathrm{Drug}}^{\mathrm{SL}}}{RT}\left(1-\frac{T}{T_{0\mathrm{Drug}}^{\mathrm{SL}}}\right) - \frac{\Delta c_{p,0\mathrm{Drug}}^{\mathrm{SL}}}{R}\left[\ln\left(\frac{T_{0\mathrm{Drug}}^{\mathrm{SL}}}{T}\right) - \frac{T_{0\mathrm{Drug}}^{\mathrm{SL}}}{T} + 1\right]\right\} \quad (1)$$

where $\Delta h_{0\mathrm{Drug}}^{\mathrm{SL}}$, $T_{0\mathrm{Drug}}^{\mathrm{SL}}$, and $\Delta c_{p,0\mathrm{Drug}}^{\mathrm{SL}}$ are the heat of fusion, the melting temperature, and the difference in solid and liquid heat capacities of the crystalline drug, respectively.

In addition, driven by the ever-increasing computing power and advanced algorithms, machine learning is widely used in pharmaceutical research through the learning of big data,[21,22] such as preparation of drug nanocrystals,[23] solvate prediction of drugs,[24] and stability of solid dispersions.[25] The solubility datasets of many compounds have also been created,[26,27] and solubility predictions have been achieved to help with high-throughput calculation and screening via different machine learning algorithms.[2,3,28] Hansch et al., for the first time, proved the strong relationship between aqueous solubility and oil−water partition coefficients for an organic liquid.[29] For the aqueous solubility of organic crystals, a similar solubility relationship with oil−water partition coefficients has also been found.[30] Subsequently, Jain and Yalkowsky assessed the aqueous solubility of 580 organic nonelectrolytes to establish a modified general solubility equation that describes aqueous solubility as a function of oil−water partition coefficients and melting point.[31] McDonagh et al. combined cheminformatics and chemical theory that solution free energies and cheminformatics descriptors were inputted to partial least squares, random forest, and support vector machines to predict the intrinsic aqueous solubility.[20] Cui et al. pointed out that the deep neural network may have better performance and has successfully predicted the solubility of a series of new synthetic compounds.[2] According to previous literature, the solubility predictions of drugs are mostly limited to aqueous solubility at 298.15 K, which has greatly limited the scope of application of the trained models. It is necessary to establish a model that can accurately predict the solubility of drugs in various solvents and at different temperatures. Furthermore, although the above-mentioned machine learning models can achieve quite good quantitative accuracy with low root-mean-square error, these models still have considerable limitations that the developed machine learning model lacks understanding of the basic principles of thermodynamics.[32] Meanwhile, some models met the problem of the validity of applied descriptors and poor solubility data. Acree Jr et al. compared the reported melting point and melting enthalpy data of multiple drugs and found that data may be influenced by a number of factors, such as the experimental device, the method of experiment, and the purity of the experimental sample.[33]

Therefore, in this work, according to the current problems of available models, we focused on developing an explainable strategy that combines molecular thermodynamics and machine learning to predict the solubility of drugs in various solvents. PC-SAFT was utilized to model drugs and solvents as a representative theory of molecular thermodynamics. The melting point and pure PC-SAFT parameters (drug−drug interaction) as well as mixing rules of PC-SAFT (drug−solvent interaction) were used as the molecular descriptors. A comprehensive database containing 4567 solubility data of drugs in various solvents and at a wide range temperature was first established. The database was further cleaned and divided into five typical datasets that covered the entire polar range of solvents. Based on the proposed molecular descriptor strategy,

https://doi.org/10.1021/acs.iecr.1c00998
Ind. Eng. Chem. Res. 2021, 60, 9259−9268

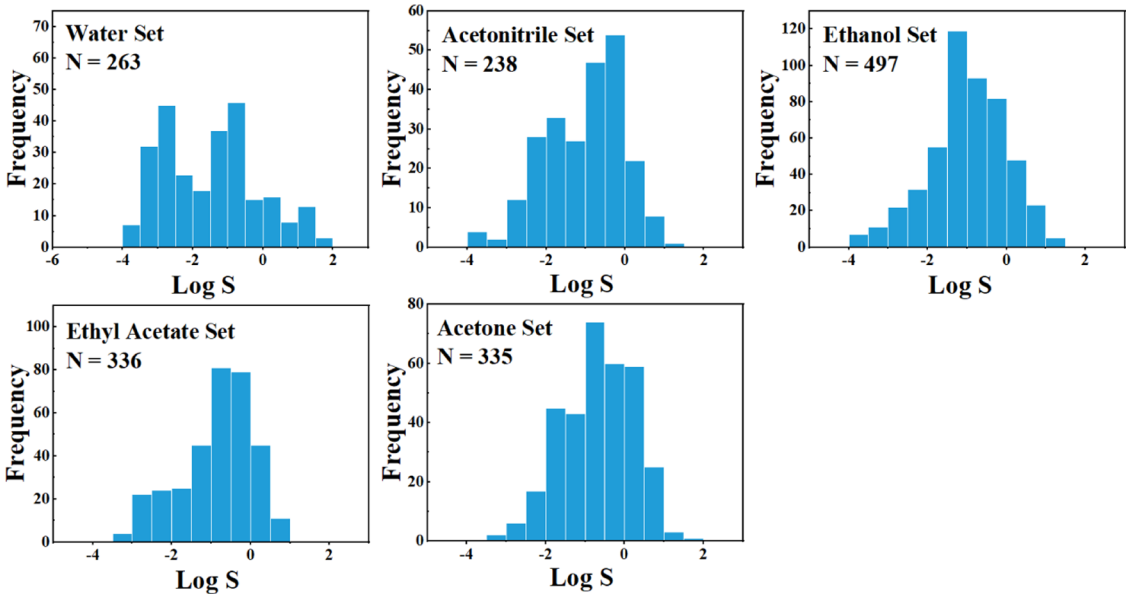


Figure 1. Solubility distributions of the water set, acetonitrile set, ethanol set, acetone set, and ethyl acetate set.

the solubility datasets were analyzed and these molecular descriptors were further inputted into the multiple linear regression (MLR), artificial neural network (ANN), random forest (RF), extremely randomized trees (ET), and support vector machine (SVM) for predicting the drug solubilities. Subsequently, the performance of five machine learning models combined with molecular thermodynamics was evaluated and compared. The interpretability of the model was studied. Finally, the generalization performance of the model was further evaluated and validated.

2. METHODOLOGY

2.1. Datasets of the Solubility of Drugs. Extensive research studies have been conducted on the solubility of drugs (Log *S*, the base 10 logarithm of molarity) to create a database consisting of 4567 solubility data, which has been provided in the Supporting Information. The drug list of the solubility data is summarized in Table 1. This database covers a variety of drugs and solvents (in total, 103 drugs and 49 solvents) as well as a fairly wide range of temperatures (from 253.15 to 364.5 K), which is also an important feature for influencing the solubility of drugs. However, due to the uneven distribution of data points, the total database was further cleaned and divided into five datasets to cover the entire polarity range of the solvents as far as possible, including water set, acetonitrile set, ethanol set, acetone set, and ethyl acetate set. The range for Log *S* in the organic solvent set was trimmed to a typical value between −4 and 2, and the solubility scope in the water set was also limited to the same range for the convenience of comparison. The solubility distributions of the five datasets are presented in Figure 1, which shows the normal distribution.

2.2. Molecular Descriptor. PC-SAFT is a representative theory of molecular thermodynamics, which can describe the drug−drug interaction and the drug−solvent interaction via pure parameters and mixing rules, respectively. These molecular interactions are accordingly developed to act as the molecular descriptors in the solubility prediction.

In PC-SAFT, the molecule is treated as a chain with m_i^{seg} spherical segments whose diameter is regarded as σ_i.[34,35] For nonassociating molecules, in addition to the number of segments (m_i^{seg}) and the diameter of the segments (σ_i), the third pure parameter is the dispersion energy parameter (u_i/k_B, where k_B is Boltzmann's constant), which characterizes the chain-to-chain interactions. For the system containing associating molecules, the association-energy parameter ($\varepsilon_{hb}^{A_iB_i}/k_B$), and the association-volume parameter ($\kappa_{hb}^{A_iB_i}$) are also required to describe the molecule. Furthermore, the number of association sites (N^{assoc}) is determined by the molecular characteristics, which represent the number of proton donators and acceptors. Here, we provide a modeling example of rivaroxaban, as shown in Figure 2.

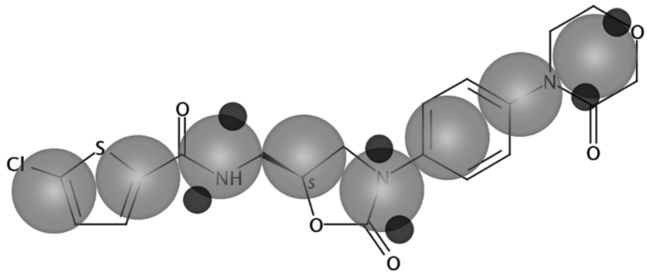


Figure 2. Modeling scheme of rivaroxaban by PC-SAFT described as a chain consisting of spherical segments (gray) and association sites N^{assoc} (black). In this example, N^{assoc} equals six: three electron acceptors and three electron donors.

Besides, the Berthelot−Lorentz combining rules are utilized to model the interactions between different components *i* and *j*.

$$\sigma_{ij} = \frac{1}{2}(\sigma_i + \sigma_j) \tag{2}$$

$$u_{ij} = (1 - k_{ij})\sqrt{u_i u_j} \tag{3}$$

where the binary interaction parameter k_{ij} is considered as a linear function of temperature.[34,35]

$$k_{ij} = k_{ij,T} \cdot T + k_{ij,0} \tag{4}$$

https://doi.org/10.1021/acs.iecr.1c00998
Ind. Eng. Chem. Res. 2021, 60, 9259−9268

The following combing rules (eqs 5 and 6) proposed by Wolbach and Sandler describe the cross-association interactions between the two associating components.[36]

$$\varepsilon_{hb}^{A_iB_j} = \frac{1}{2}(\varepsilon_{hb}^{A_iB_i} + \varepsilon_{hb}^{A_jB_j}) \quad (5)$$

$$\kappa_{hb}^{A_iB_j} = \sqrt{\kappa_{hb}^{A_iB_i}\kappa_{hb}^{A_jB_j}}\left[\frac{\sqrt{\sigma_{ii}\sigma_{jj}}}{(1/2)(\sigma_{ii}+\sigma_{jj})}\right]^3 \quad (6)$$

More detailed formulas are available in the literature.[34,35] All of the functions of PC-SAFT can be realized through programming. PC-SAFT pure-component parameters can be estimated by fitting to the liquid densities and vapor pressures. For modeling of drug-related systems, the pure-component parameters of a drug can also be achieved by fitting to drug solubility data in any solvent(s). More details on the PC-SAFT model can be found in the literature.[37]

For drugs, as presented in Table 1, whose PC-SAFT parameters are reported in the literature, their parameters are collected while the rest of the drugs with unreported parameters are modeled in this work. The pure PC-SAFT parameters of drugs and solvents are all summarized in Tables S1 and S2. The melting points and melting enthalpies of the collected drugs are presented in Table S3. In addition, this work also reports a considerable number of binary interaction parameters between drugs and multiple solvents in the database of the Supporting Information, which are determined by the modeling in this work.

Consequently, 16 molecular descriptors used in this work are presented in Table 2.

Table 2. List of the Molecular Descriptors

symbol	description	symbol	description
MP	melting point of drug	$Drug_{N_1}$	number of proton donators
ΔH	melting enthalpy of drug	$Drug_{N_2}$	number of proton acceptors
$Drug_M$	molecular weight of drug	seg_{ij}	σ_{ij}
$Drug_{mseg}$	number of drug segments	u_{ij}	u_{ij}
$Drug_{seg}$	diameter of drug segments	k_{ij}	binary interaction parameter
$Drug_u$	dispersion energy parameter of drug	e_{ij}	$\varepsilon_{hb}^{A_iB_j}$
$Drug_a$	association-energy parameter of drug	v_{ij}	$\kappa_{hb}^{A_iB_j}$
$Drug_v$	association-volume parameter of drug	T	temperature

2.3. Machine Learning Algorithm. The machine learning algorithm was implemented using Scikit-learn 0.21.3 and Python 3.7. A random 80% of each dataset was used to create a training set, and the remaining 20% of data was regarded as the test set to evaluate the predictive capability of the trained model. Before training and testing, the entire training set and test set were standardized to avoid that the prediction results will not be dominated by some feature values with too large dimensions. In this work, MLR, ANN, RF, ET, and SVM were used to establish the predictive models. In all machine learning models, except the default values used in the MLR model, the hyperparameters were optimized through 5-fold cross-validation and grid search to generate a robust model, where the accuracy of the regression was used as the score function. The number of neurons was optimized for the single-hidden-layer ANN model for the reason that deep neural networks were not adopted due to the small size of the dataset. Early stopping, when training was stopped to prevent overfitting when the score was not improved after 10 epochs, was also used. The number of trees was optimized for the RF model and the ET model. In addition, penalty parameters C and γ were optimized for the SVM model with radial basis function (RBF) kernel. All of the optimized hyperparameters are summarized in Table 3.

Table 3. Optimized Hyperparameters for MLR, ANN, RF, ET, and SVM Models

	MLR	ANN (the number of neurons)	RF (the number of trees)	ET (the number of trees)	SVM (penalty parameters C and γ)
water set	default	290	150	790	190, 0.008
acetonitrile set	default	820	580	570	2720, 0.005
ethanol set	default	620	720	210	150, 0.03
acetone set	default	500	280	20	270, 0.02
ethyl acetate set	default	840	300	140	950, 0.01

2.4. Evaluation Indicators. Some statistical indicators such as the root-mean-squared error (RMSE) and the coefficient of determination (R^2) are often employed to evaluate the model performance. However, as proposed by Boobier et al,[1] due to the typical experimental error for Log S ($\pm$0.5−0.7) caused by the experimental factors,[38] these statistical indicators usually cannot be a comprehensive and correct response for the results of drug solubility prediction. Consequently, two additional prediction indicators (Log S ($\pm$0.7)% and Log S ($\pm$1.0)%) were defined to represent the maximum accuracy and the ability to act as a product development guidance, respectively.

3. RESULTS AND DISCUSSION

3.1. Model Performance. As described in Section 2.3, five machine learning algorithms with optimized hyperparameters were used to carry out the training and test stages, and the relevant results are summarized in Table S4. It was observed that the results of the linear model (MLP) were far inferior to those of the other four nonlinear models, suggesting that nonlinear models are more applicable in solubility prediction in various solvents in comparison with the linear model. In addition, a comparatively worse result predicted by the MLP model was found in the test set of ethanol set, where the value of R^2 was determined to be 0.68 and the value of RMSE was much larger than other predictions, as shown in Table S4. However, compared to these two statistical indicators, the newly defined prediction indicators are 84.00% for Log S ($\pm$0.7)% and 91.00% for Log S ($\pm$1.0)%, respectively, which shows that most of the experimental values matched the predicted values well and proved that statistical indicators cannot provide a comprehensive measure of model performance. In fact, on comparing the results predicted by the MLR model and the SVR model in the same test set of the ethanol set, it is found that despite the large difference in the results of R^2 and RMSE, the predictions of the two models achieve great results in Log S ($\pm$0.7)% and Log S ($\pm$1.0)%, demonstrating the poor reliability of R^2 and RMSE again.

https://doi.org/10.1021/acs.iecr.1c00998
Ind. Eng. Chem. Res. 2021, 60, 9259−9268

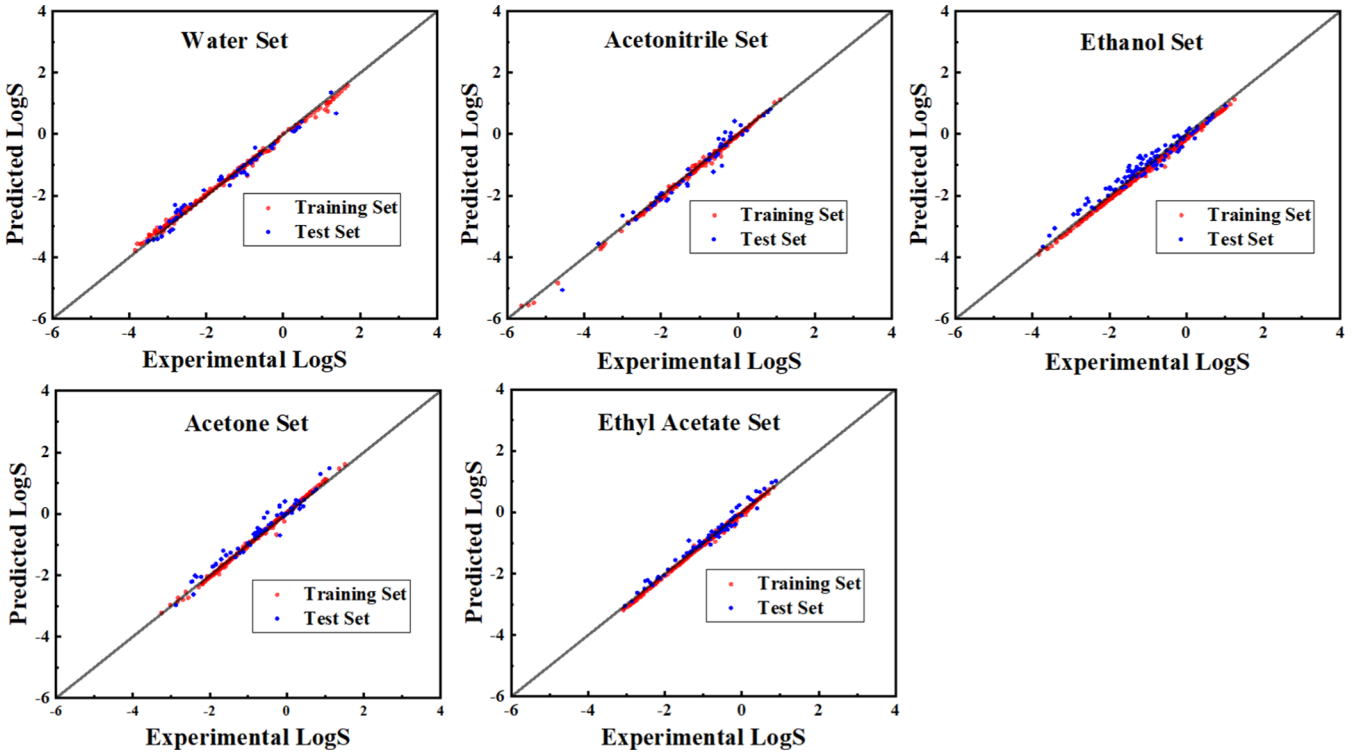


Figure 3. Prediction results of the ANN model for the water set, acetonitrile set, ethanol set, acetone set, and ethyl acetate set.

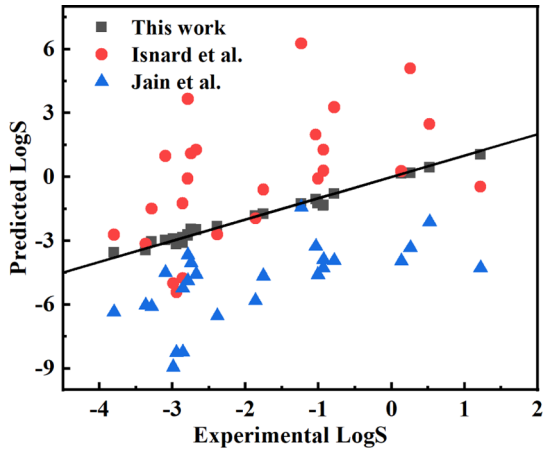


Figure 4. Comparison of the prediction performance of the proposed model in this work with the models in the literature for the aqueous solubility of the drug.

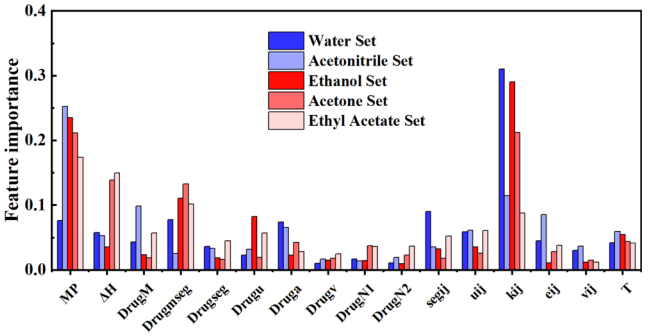


Figure 5. Feature importance of five datasets.

Table 4. Evaluation Indicators of ANN Model Predictions with and without Melting Point Consideration in the Ethanol Set

	with melting point		without melting point	
evaluation indicators	training set	test set	training set	test set
R^2	0.98	0.95	0.98	0.83
RMSE	0.02	0.04	0.02	0.16
Log S (±0.7)%	100.00	100.00	99.75	96.00
Log S (±1.0)%	100.00	100.00	99.75	96.00

Therefore, the performances of ANN, RF, ET, and SVM models in different datasets were further evaluated to determine the most suitable prediction model. In terms of the newly defined indicators, the performance of the ANN model is slightly inferior to the SVR model in the water set. The ANN model was chosen as the best model in the acetone set due to the excellent results of evaluation indicators in both the training set and the test set. In the acetonitrile set, although the ANN and ET models have achieved similar excellent performance, the ANN model has a slight advantage in the newly defined indicators. Additionally, the same result was also found in the ethanol set and the ethyl acetate set. For consistency and practical application convenience, the ANN model was used for all solvent sets. The prediction results of the ANN model for the organic solvent set and the water set are also displayed in Figure 3. It can be found that almost all predicted values are well matched to the experimental values and no data points that deviate significantly from the diagonal line are observed, revealing the excellent performance of the model.

Meanwhile, the proposed aqueous solubility model was used to further compare with some simple and straightforward models in the literature in the prediction of aqueous solubility of drugs. As shown in Figure 4, it can be found that the

https://doi.org/10.1021/acs.iecr.1c00998
Ind. Eng. Chem. Res. 2021, 60, 9259−9268

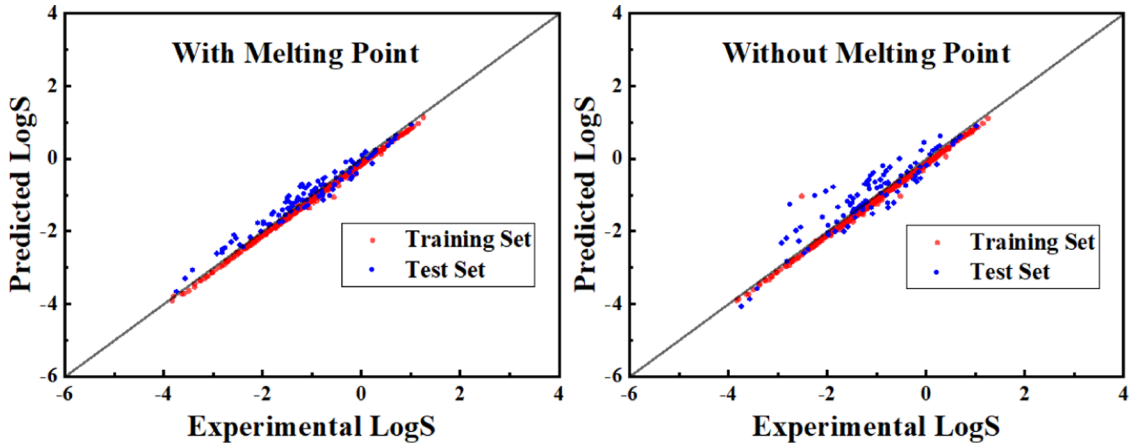


Figure 6. Comparison between the predicted and experimental solubility with and without melting point consideration in ethanol set.

Table 5. Drug List of the Generalization Evaluation Set

generalization evaluation set	types of drugs
water evaluation set	rivaroxaban, pyrazinamide, edaravone
acetonitrile evaluation set	itraconazole, hydrochlorothiazide, pyrazinamide, lamotrigine
ethanol evaluation set	rivaroxaban, indomethacin, itraconazole, hydrochlorothiazide, hydrocortisone, prednisolone, pyrazinamide, lamotrigine
acetone evaluation set	saccharin, carbamazepine, succinic acid, mefenamic acid
ethyl acetate evaluation set	rivaroxaban, indomethacin, itraconazole, hydrochlorothiazide, pyrazinamide, capecitabine

Table 6. Evaluation Indicators of the Binary Interaction Parameter Prediction Model with Five Datasets

	evaluation indicators	training set	test set
water set	R^2	0.99	0.99
	RMSE	0.0001	0.0001
acetonitrile set	R^2	0.98	0.97
	RMSE	0.0001	0.0001
ethanol set	R^2	0.99	0.98
	RMSE	0.00003	0.0001
acetone set	R^2	0.98	0.94
	RMSE	0.0001	0.0003
ethyl acetate set	R^2	1.00	0.98
	RMSE	0.00003	0.0002

proposed model is more accurate than the model in the literature. The predicted value from the model proposed by Isnard et al. is generally higher than the experimental data, while the model proposed by Jain et al. predicts generally smaller values considering the influence of melting point.[30,31]

3.2. Feature Importance. Interpretability has always been a critical aspect of machine learning models because explainable machine learning models can help researchers better understand the output and improve the generalization performance. There are many methods to calculate the feature importance, some of which are based on specific models such as the regression coefficients in linear regression models. The rest are the universal approaches, where permutation importance is selected for the calculation in this work.

The method of permutation importance can determine the feature importance via calculating the change in the scoring function when the molecular descriptor values are replaced by randomly arranged values in turn. Figure 5 shows the feature importance of various molecular descriptors used in this work. Binary interaction parameters play an important role in both organic solvent and water systems, and it was also found that the importance of binary interaction parameters increases with solvent polarity, except for acetonitrile. The reason for this is that drug molecules and solvent molecules often have different sizes and ionization potentials, causing the nonideal behavior of the mixture, while the binary interaction parameters can quantitatively amend the interactions between different molecules that affect the properties of the mixture. In particular, its influence is greater in the presence of polar molecules. Since the difference in size and shape between water molecules and drug molecules is more obvious than that of organic solvent molecules, the correction of intermolecular interactions of drug−water systems is very important for the excess properties of mixtures, which further significantly affects the calculated aqueous solubility of the drug. Binary interaction parameters in acetonitrile do not show sufficient importance, possibly due to the limited size of datasets. The melting point is observed to be of vital importance in organic solvents, especially in acetonitrile and ethanol, while the water system shows less dependence on the melting point conversely. The obtained results in drug−water systems are in agreement with some reported results in the literature. Salahinejad et al. and Emami et al. also found the same low values of feature importance for enthalpy and melting point,[39,40] despite in some reports, such as the GSE equation,[31] aqueous solubility has a significant correlation with the melting point and the oil−water partition coefficient. Furthermore, another finding is that the feature importance of each molecular descriptor except the binary interaction parameter is uniformly distributed in the water set. The reason for the above phenomena can be summarized as that the solubility of drugs in water is dominated by the solvation energy and the drug−solvent interaction, while for systems with organic solvents, the drug−drug interaction is very important.[1,41] This shows that the interpretation of the machine learning model is consistent with the understanding of solubility in aspects of physics and chemistry.

Subsequently, the predictions were compared by the ANN model using the ethanol set in the presence and absence of the melting point due to the vital feature importance of the melting point. The optimized number of neurons was determined as 670 in the ANN model without melting point. Table 4 shows

https://doi.org/10.1021/acs.iecr.1c00998
Ind. Eng. Chem. Res. 2021, 60, 9259−9268

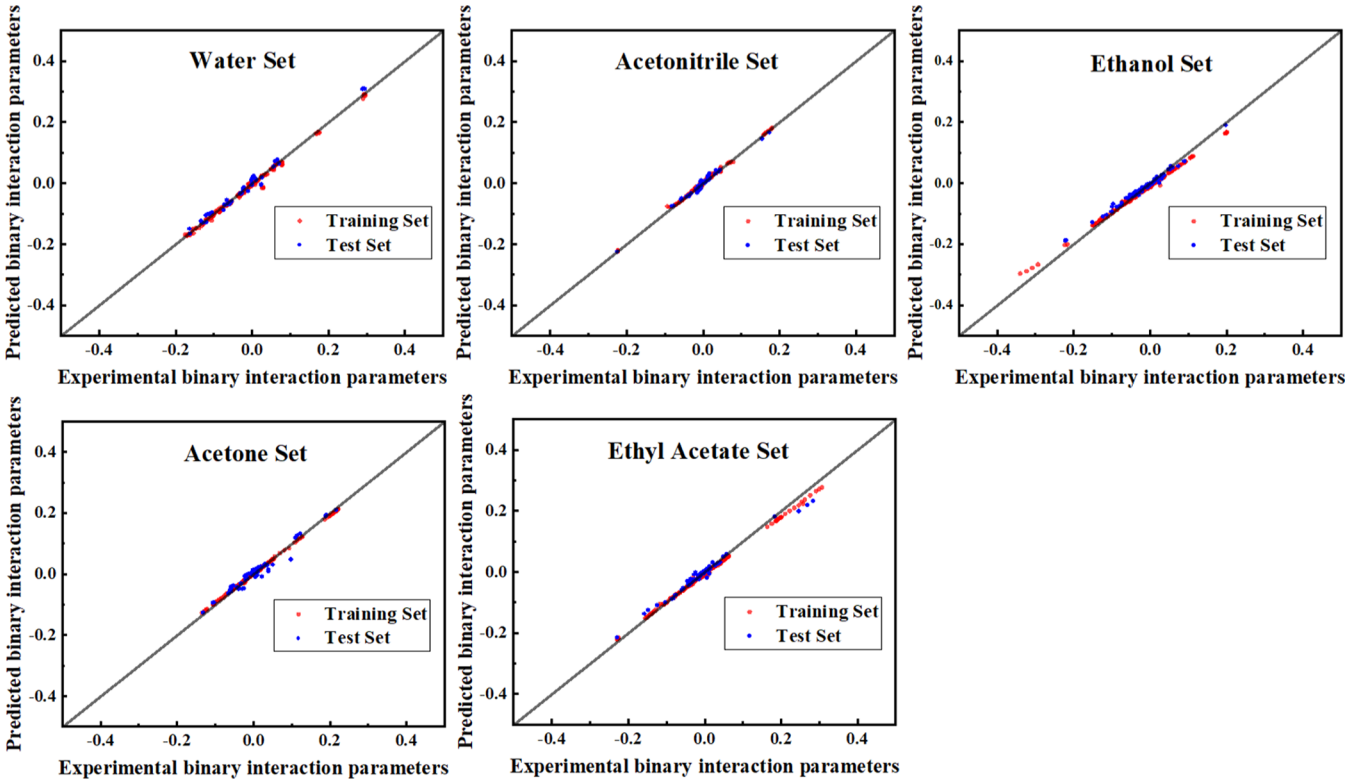


Figure 7. Comparison between the predicted and experimental binary interaction parameters.

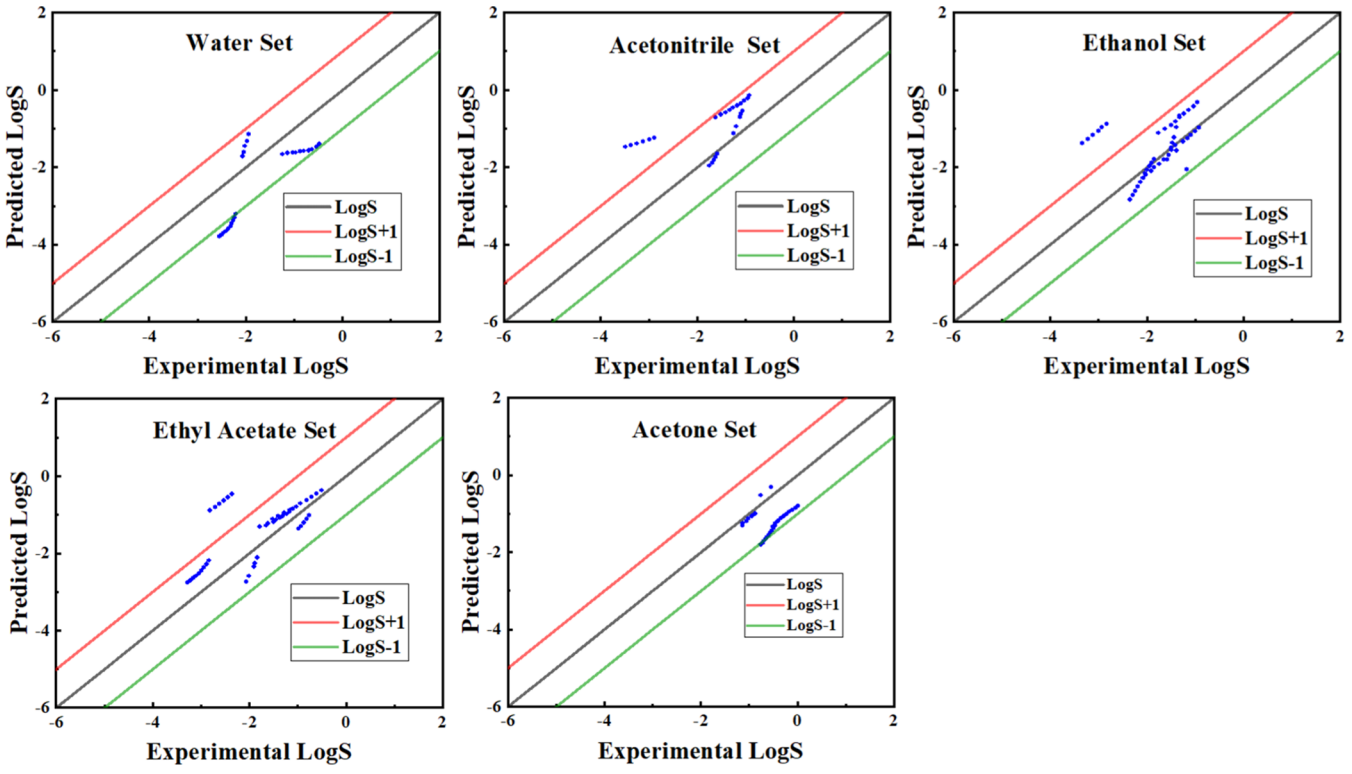


Figure 8. Comparison between the predicted and experimental solubility in the generalization evaluation set.

the evaluation indicators with and without melting point consideration, showing that the performance of the model becomes worse in both training and test sets in the absence of the melting point. Figure 6 further shows the comparison with and without considering the melting point. It is obvious that the model shows better performance and the data points fit more closely on the diagonal when the melting point is taken into account. The results of drug solubility in other organic solvent sets that do not consider the melting point have also been shown in Figures S1−S3 and Tables S6−S8, and the

https://doi.org/10.1021/acs.iecr.1c00998
Ind. Eng. Chem. Res. 2021, 60, 9259−9268

same results as the ethanol set have been found. Overall, the importance of the melting point as the molecular descriptor and the interpretability of the model are further demonstrated. In other words, those results can provide guidance for improving model performance, namely, adding more molecular descriptors of drug−solvent interactions in the water system and drug−drug interactions in the organic solvent systems.

3.3. Evaluation of Generalization Performance. To better evaluate the generalization performance of the established model, the model should be compared with the reported models or applied to a new unrelated test set. However, it is difficult to fairly compare models of this work with reported models due to the various datasets adopted by each model. Therefore, a generalization evaluation set was created consisting of some new data points that never appeared in the previous database, and the drug list of these data is shown in Table 5.

Since binary interaction parameters are often obtained by regressing the experimental data or obtained by the group contribution method and are often unknown as well, a binary interaction parameter prediction model is proposed using a single-hidden-layer ANN model with the previous solvent sets, where 15 parameters other than binary interaction parameters in Table 2 are used as inputs and the binary interaction parameters as outputs. The optimized hyperparameters are 500 neurons for the water set, 670 neurons for the acetonitrile set, 960 neurons for the ethanol set, 870 neurons for the acetone set, and 730 neurons for the ethyl acetate set, respectively. The results are shown in Table 6 and Figure 7. It can be found that the predicted values and the experimental data are well matched and all of the binary interaction parameters are well predicted without overfitting, demonstrating excellent predictive performance. Researchers can also try other approaches to retrain the solubility prediction models based on the conclusions of Section 3.2. Due to the importance of the drug−water interaction in water systems and that of drug−drug interaction in organic solvent systems, the binary interaction parameters in the water system can be replaced by other parameters that characterize the interaction between molecules such as average interaction energy obtained from the molecular simulation, while that in the organic solvent systems can be assumed as 0.

Consequently, the binary interaction parameters of the generalization evaluation set were predicted by the binary interaction parameter prediction model and were then taken into the solubility prediction model. The solubility prediction of the generalization evaluation set is displayed in Figure 8. Although the accuracy of the predictions has declined relative to the test set, it is obvious that most of the drug solubilities are well predicted and are between plus and minus 1 for Log S, revealing that the developed model in this work has good generalization performance. In addition, the outliers for each generalization evaluation set were further analyzed, and the results are shown in Table S5. The appearance of these outliers was probably due to the different data distributions of the training set and the generalization evaluation set. Drugs with outliers tend to have higher molecular weight and more acidic functional groups, which are not clearly expressed in the training set. In addition, it may also be influenced by the quality of the collected data, such as the validity of applied melting point and melting enthalpy data.[33] Accordingly, these results also indicate the requirement of further expanding the dataset to adequately obtain the features of multiple drugs so that the solubility of the drug in various solvents can be more accurately predicted.

4. CONCLUSIONS

In summary, a novel strategy that combined molecular thermodynamic and machine learning was proposed to predict the solubility of drugs in various solvents accurately. The strategy was based on 16 molecular descriptors representing drug−drug interactions and drug−solvent interactions. A single-hidden-layer neural network was finally determined as the hybrid predictive model for the solubility of drugs in various solvents.

On the basis of five datasets created by the established solubility database, the predictive model was trained and the experimental values of the five datasets were successfully predicted. The feature importance of each molecular descriptor was also studied, which has improved the interpretability and transparency of the model. Meanwhile, the model was applied to a generalization evaluation set consisting of some new data points that never appeared in the previous database, showing that the model has good generalization performance.

Additionally, according to the above results, three directions for improving the model were summarized as adding molecular descriptors of drug−solvent interactions in the water system and drug−drug interactions in the organic solvent system, as well as expanding the dataset to adequately obtain the features of multiple drugs. Overall, the proposed model in this work requires fewer molecular descriptors, can achieve excellent results based on small datasets, and has good interpretability. These findings show that the proposed model has the capability of solubility prediction, which can provide important information for drug development and drug solvent screening.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.iecr.1c00998.

Pure PC-SAFT parameters of drugs and solvents; melting point and melting enthalpy of drugs; table of the evaluation indicators of machine learning predictions with five datasets; table of the outliers for the generalization evaluation set; and comparison of the proposed model performance with or without melting point (PDF)

Database; water set; acetonitrile set; ethanol set; acetone set; ethyl acetate set; and generalization evaluation set (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

Yuanhui Ji − *Jiangsu Province Hi-Tech Key Laboratory for Biomedical Research, School of Chemistry and Chemical Engineering, Southeast University, Nanjing 211189, People's Republic of China;* orcid.org/0000-0002-3039-4368; Phone: +86-13951907361; Email: yuanhui.ji@seu.edu.cn, yuanhuijinj@163.com

Author

Kai Ge − *Jiangsu Province Hi-Tech Key Laboratory for Biomedical Research, School of Chemistry and Chemical Engineering, Southeast University, Nanjing 211189, People's Republic of China*

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.iecr.1c00998

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This research received funding from the National Natural Science Foundation of China (Grant nos. 21776046 and 21978047), the Fundamental Research Funds for the Central Universities (Grant no. 2242020K40033), and the Six Talent Peaks Project in Jiangsu Province (Grant no. XCL-079).

■ REFERENCES

(1) Boobier, S.; Hose, D. R. J.; Blacker, A. J.; Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **2020**, *11*, No. 5753.
(2) Cui, Q. J.; Lu, S.; Ni, B. W.; Zeng, X.; Tan, Y.; Chen, Y. D.; Zhao, H. P. Improved prediction of aqueous solubility of novel compounds by going deeper with deep learning. *Front. Oncol.* **2020**, *10*, No. 121.
(3) Zhao, J. J.; Yang, J.; Xie, Y. Improvement strategies for the oral bioavailability of poorly water-soluble flavonoids: An overview. *Int. J. Pharm.* **2019**, *570*, No. 118642.
(4) Fernandes, G. J.; Kumar, L.; Sharma, K.; Tunge, R.; Rathnanand, M. A review on solubility enhancement of carvedilol—a BCS Class II drug. *J. Pharm. Innov.* **2018**, *13*, 197−212.
(5) Loschen, C.; Klamt, A. Solubility prediction, solvate and cocrystal screening as tools for rational crystal engineering. *J. Pharm. Pharmacol.* **2015**, *67*, 803−811.
(6) Sheikholeslamzadeh, E.; Rohani, S. Solubility prediction of pharmaceutical and chemical compounds in pure and mixed solvents using predictive models. *Ind. Eng. Chem. Res.* **2012**, *51*, 464−473.
(7) Huang, Z.; Sha, J.; Chang, Y.; Cao, Z.; Hu, X.; Li, Y.; Li, T.; Ren, B. Solubility measurement, model evaluation and Hansen solubility parameter of ipriflavone in three binary solvents. *J. Chem. Thermodyn.* **2021**, *152*, No. 106285.
(8) Wilson, G. M. Vapor-liquid equilibrium. XI. a new expression for the excess free energy of mixing. *J. Am. Chem. Soc.* **1964**, *86*, 127−130.
(9) Fredenslund, A.; Jones, R. L.; Prausnitz, J. M. Group-contribution estimation of activity-coefficients in nonideal liquid-mixtures. *AIChE J.* **1975**, *21*, 1086−1099.
(10) Abrams, D. S.; Prausnitz, J. M. Statistical thermodynamics of liquid-mixtures - new expression for excess gibbs energy of partly or completely miscible systems. *AIChE J.* **1975**, *21*, 116−128.
(11) Prudic, A.; Ji, Y.; Sadowski, G. Thermodynamic phase behavior of API/polymer solid dispersions. *Mol. Pharmaceutics* **2014**, *11*, 2294−2304.
(12) Prudic, A.; Kleetz, T.; Korf, M.; Ji, Y.; Sadowski, G. Influence of copolymer composition on the phase behavior of solid dispersions. *Mol. Pharmaceutics* **2014**, *11*, 4189−4198.
(13) Prudic, A.; Ji, Y.; Luebbert, C.; Sadowski, G. Influence of humidity on the phase behavior of API/polymer formulations. *Eur. J. Pharm. Biopharm.* **2015**, *94*, 352−362.
(14) Forte, E.; Burger, J.; Langenbach, K.; Hasse, H.; Bortz, M. Multi-criteria optimization for parameterization of SAFT-type equations of state for water. *AIChE J.* **2018**, *64*, 226−237.
(15) Aigner, M.; Echtermeyer, A.; Kaminski, S.; Viell, J.; Leonhard, K.; Mitsos, A.; Jupke, A. Ternary system Co2/2-mthf/water—experimental study and thermodynamic modeling. *J. Chem. Eng. Data* **2020**, *65*, 993−1004.
(16) Wang, N.; Huang, X.; Gong, H.; Zhou, Y.; Li, X.; Li, F.; Bao, Y.; Xie, C.; Wang, Z.; Yin, Q.; Hao, H. Thermodynamic mechanism of selective cocrystallization explored by MD simulation and phase diagram analysis. *AIChE J.* **2019**, *65*, No. e16570.
(17) Boothroyd, S.; Anwar, J. Solubility prediction for a soluble organic molecule via chemical potentials from density of states. *J. Chem. Phys.* **2019**, *151*, No. 184113.
(18) Paluch, A. S.; Maginn, E. J. Predicting the solubility of solid phenanthrene: a combined molecular simulation and group contribution approach. *AIChE J.* **2013**, *59*, 2647−2661.
(19) Lee, B. S.; Lin, S. T. Prediction and screening of solubility of pharmaceuticals in single- and mixed-ionic liquids using COSMO-SAC model. *AIChE J.* **2017**, *63*, 3096−3104.
(20) McDonagh, J. L.; Nath, N.; De Ferrari, L.; van Mourik, T.; Mitchell, J. B. O. Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules. *J. Chem. Inf. Model.* **2014**, *54*, 844−856.
(21) Tabora, J. E.; Lora Gonzalez, F.; Tom, J. W. Bayesian probabilistic modeling in pharmaceutical process development. *AIChE J.* **2019**, *65*, No. e16744.
(22) Liu, Q.; Zhang, L.; Tang, K.; Liu, L.; Du, J.; Meng, Q.; Gani, R. Machine learning-based atom contribution method for the prediction of surface charge density profiles and solvent design. *AIChE J.* **2021**, *67*, No. e17110.
(23) He, Y.; Ye, Z.; Liu, X.; Wei, Z.; Qiu, F.; Li, H.-F.; Zheng, Y.; Ouyang, D. Can machine learning predict drug nanocrystals? *J. Controlled Release* **2020**, *322*, 274−285.
(24) Xin, D.; Gonnella, N. C.; He, X.; Horspool, K. Solvate prediction for pharmaceutical organic molecules with machine learning. *Cryst. Growth Des.* **2019**, *19*, 1903−1911.
(25) Han, R.; Xiong, H.; Ye, Z.; Yang, Y.; Huang, T.; Jing, Q.; Lu, J.; Pan, H.; Ren, F.; Ouyang, D. Predicting physical stability of solid dispersions by machine learning techniques. *J. Controlled Release* **2019**, *311*−*312*, 16−25.
(26) Livingstone, D. J.; Ford, M. G.; Huuskonen, J. J.; Salt, D. W. Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 741−752.
(27) Perryman, A. L.; Inoyama, D.; Patel, J. S.; Ekins, S.; Freundlich, J. S. Pruned machine learning models to predict aqueous solubility. *ACS Omega* **2020**, *5*, 16562−16567.
(28) Bergström, C. A. S.; Larsson, P. Computational prediction of drug solubility in water-based systems: Qualitative and quantitative approaches used in the current drug discovery and development setting. *Int. J. Pharm.* **2018**, *540*, 185−193.
(29) Hansch, C.; Quinlan, J. E.; Lawrence, G. L. Linear free-energy relationship between partition coefficients and the aqueous solubility of organic liquids. *J. Org. Chem.* **1968**, *33*, 347−350.
(30) Isnard, P.; Lambert, S. Aqueous solubility and n-octanol/water partition coefficient correlations. *Chemosphere* **1989**, *18*, 1837−1853.
(31) Jain, N.; Yalkowsky, S. H. Estimation of the aqueous solubility I: Application to organic nonelectrolytes. *J. Pharm. Sci.* **2001**, *90*, 234−252.
(32) Song, Z.; Shi, H.; Zhang, X.; Zhou, T. Prediction of CO2 solubility in ionic liquids using machine learning methods. *Chem. Eng. Sci.* **2020**, *223*, No. 115752.
(33) Acree, W. E., Jr; Jouyban, A. Comments on ″what if cocrystallization fails for neutral molecules? screening offered eutectics as alternate pharmaceutical materials: leflunomide-a case study″. *Pharm. Sci.* **2019**, *25*, 369−372.
(34) Gross, J.; Sadowski, G. Perturbed-Chain SAFT: An equation of state based on a perturbation theory for chain molecules. *Ind. Eng. Chem. Res.* **2001**, *40*, 1244−1260.
(35) Gross, J.; Sadowski, G. Application of the Perturbed-Chain SAFT equation of state to associating systems. *Ind. Eng. Chem. Res.* **2002**, *41*, 5510−5515.
(36) Wolbach, J. P.; Sandler, S. I. Using molecular orbital calculations to describe the phase behavior of cross-associating mixtures. *Ind. Eng. Chem. Res.* **1998**, *37*, 2917−2928.
(37) Ruether, F.; Sadowski, G. Modeling the solubility of pharmaceuticals in pure solvents and solvent mixtures for drug process design. *J. Pharm. Sci.* **2009**, *98*, 4205−4215.

(38) Palmer, D. S.; Mitchell, J. B. O. Is experimental data quality the limiting factor in predicting the aqueous solubility of druglike molecules? *Mol. Pharmaceutics* **2014**, *11*, 2962−2972.
(39) Salahinejad, M.; Le, T. C.; Winkler, D. A. Aqueous solubility prediction: do crystal lattice interactions help? *Mol. Pharmaceutics* **2013**, *10*, 2757−2766.
(40) Emami, S.; Jouyban, A.; Valizadeh, H.; Shayanfar, A. Are crystallinity parameters critical for drug solubility prediction? *J. Solution Chem.* **2015**, *44*, 2297−2315.
(41) Thompson, J. D.; Cramer, C. J.; Truhlar, D. G. Predicting aqueous solubilities from aqueous free energies of solvation and experimental or calculated vapor pressures of pure substances. *J. Chem. Phys.* **2003**, *119*, 1661−1670.

https://doi.org/10.1021/acs.iecr.1c00998
Ind. Eng. Chem. Res. 2021, 60, 9259−9268