Electronic Supplementary Information (ESI)

Predictive deep learning models for environmental properties: the direct calculation of octanol-water partition coefficients from molecular graphs

Zihao Wang,‡[a] Yang Su,‡[a] Weifeng Shen,*[a] Saimeng Jin,*[a] James H. Clark,[b] Jingzheng Ren[c] and Xiangping Zhang[d]

[a] *School of Chemistry and Chemical Engineering, Chongqing University, Chongqing 400044, People's Republic of China. Email: shenweifeng@cqu.edu.cn, sj708@cqu.edu.cn*

[b] *Green Chemistry Centre of Excellence, University of York, York YO105D, UK*

[c] *Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, People's Republic of China*

[d] *Beijing Key Laboratory of Ionic Liquids Clean Process, CAS Key Laboratory of Green Process and Engineering, Institute of Process Engineering, Chinese Academy of Sciences, Beijing, 100190, People's Republic of China*

‡ These authors contributed equally to this work.

*Corresponding author:

Email: shenweifeng@cqu.edu.cn, sj708@cqu.edu.cn

Table of contents

Detailed analysis in the types of compounds.....S1
Frequencies of compounds presenting molecular features S2
Feature selection during the property predictions.....S3
Algorithms of model development and predicting.....S4
External validation on the external set.....S6
Application Domain.....S7
Impact of isomeric features on predictive accuracy S9
References S12

Detailed analysis in the types of compounds

The collected dataset of compounds spans a wide class of organic molecular structures including aliphatic and aromatic hydrocarbons, alcohols and phenols, heterocyclic compounds, amines, acids, ketones, esters, aldehydes, ethers and so on. In order to demonstrate the chemical diversity of the dataset, the corresponding counts of different types were detailed in Table S1, and their distributions in the training, test and external sets were also provided.

Table S1 The detailed analysis for the types of compounds in the entire dataset and three disjoint subsets.

	Training set	Test set	External set	Entire dataset
Aliphatic and aromatic hydrocarbons	731	80	114	925
Alcohols and phenols	523	67	67	657
Heterocyclic compounds	2074	275	243	2592
Amines	1596	188	192	1976
Acids	988	104	110	1202
Ketones	782	100	116	998
Esters	521	64	61	646
Aldehydes	49	7	11	67
Ethers	134	21	13	168
Others	1136	161	140	1437
Total	**8534**	**1067**	**1067**	**10668**

Since the subsets were divided with a random selection routine, the proportions of different types of compounds in each subset approximate the corresponding proportions for the compounds of the subset in the entire dataset.

Frequencies of compounds presenting molecular features

The detailed count for the number of compounds in the training set presenting each molecular feature is shown in Table S2. Such information is important for the future use of the predictive model to estimate the level of confidence on individual predictions considering the number of training compounds associated to each molecular feature.

Table S2 The frequencies of compounds presenting each molecular feature in model training.

Molecular feature	Frequency	Molecular feature	Frequency	Molecular feature	Frequency
[C]	5628	-[C]	6303	=[C]	381
#[C]	46	[C|r]	190	-[C|r]	2211
=[C|r]	526	[C*]	75	-[C*]	93
[C|r*]	5	-[C|r*]	638	[C**]	279
-[C**]	130	[C|r**]	22	-[C|r**]	618
[c|r]	541	-[c|r]	6201	=[c|r]	9
:[c|r]	6950	/=\[C]	102	/=/[C]	176
/=\[C|r]	15	/=/[C|r]	20	[O]	5
-[O]	4466	=[O]	5753	-[O|r]	677
[o|r]	3	:[o|r]	300	-[O-]	819
-[N]	3920	=[N]	153	#[N]	306
-[N|r]	1232	=[N|r]	217	-[n|r]	597
:[n|r]	2238	-[N+]	753	=[N+]	17
#[N+]	3	=[N-]	8	=[N+|r]	1
-[n+|r]	2	:[n+|r]	60	/=\[N]	42
/=/[N]	50	/=\[N+]	2	-[P]	186
-[P|r]	13	-[P+]	12	-[P+|r]	1
-[S]	866	=[S]	221	-[S|r]	291
[s|r]	1	:[s|r]	265	[H]	23
-[H]	8510	-[F]	662	[Cl]	1417
-[Cl]	705	[Br]	345	-[Br]	64
-[I]	112				

Feature selection during the property predictions

The molecular features chosen in the QSPR model rely on the molecular structure of the compound (refer to Fig. S1). Firstly, the canonical molecular signature was generated for a compound, and then the Tree-LSTM network for this compound was built according to the signature tree for mapping the molecular structure. Afterwards, the numeric vectors representing molecular features were fed to the nodes of the Tree-LSTM network. Finally, a vector generated in the Tree-LSTM network was introduced to the BPNN for training the predictive model.

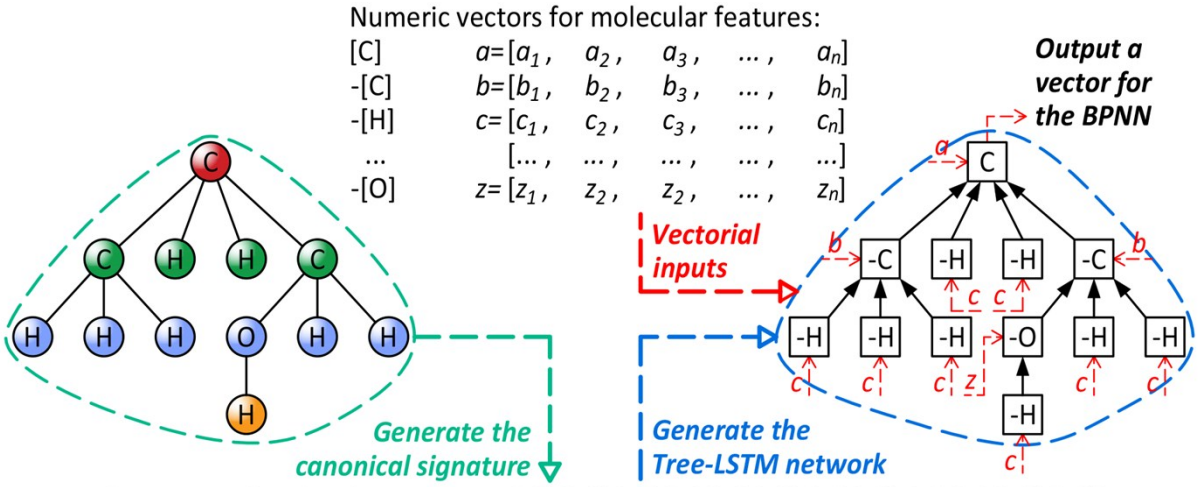

Fig. S1 The way of selecting molecular features for presenting the compound during predictions.

Algorithms of model development and predicting

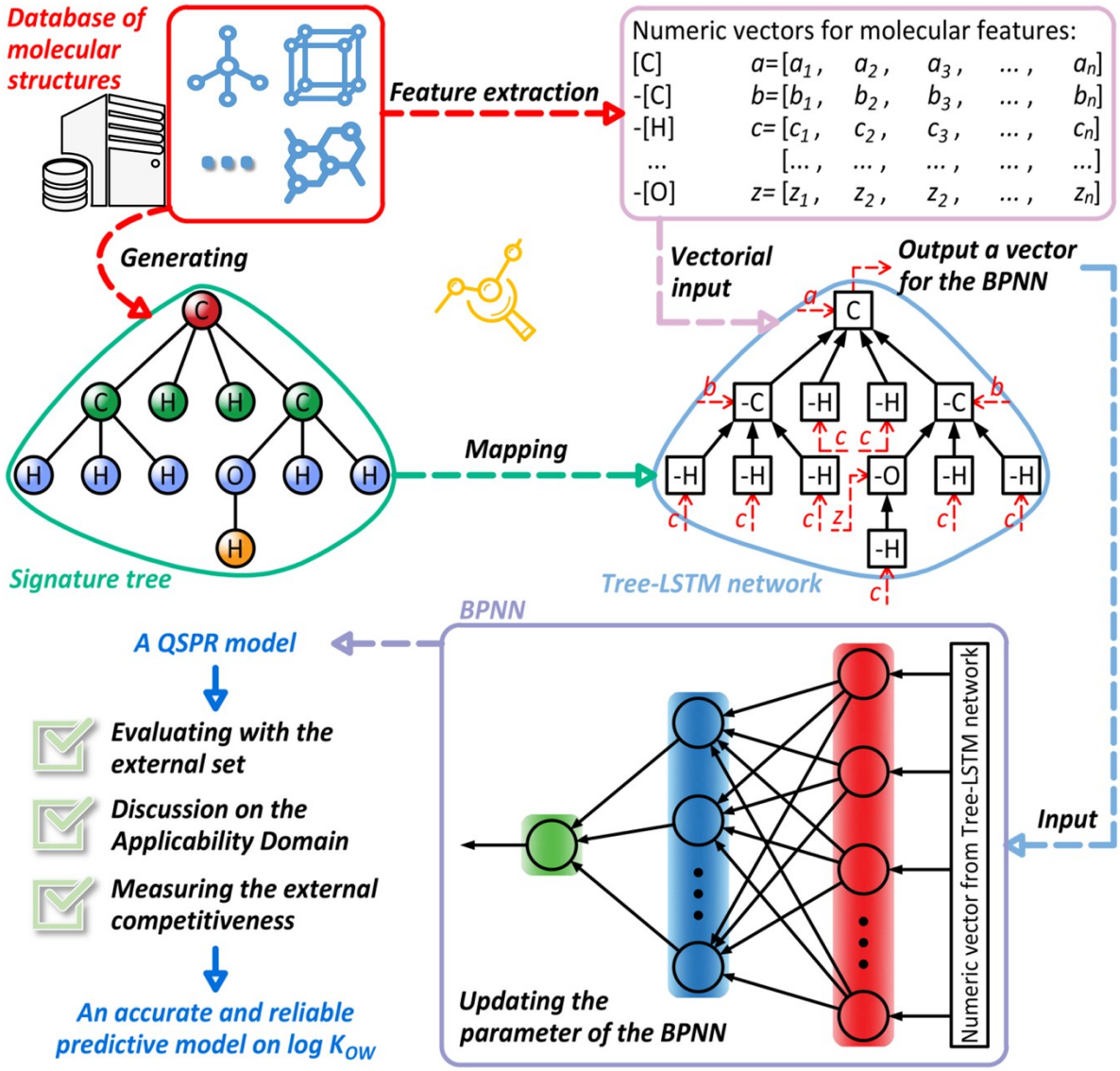

Fig. S2 The algorithm of model development with the Tree-LSTM network and BPNN.

The algorithm of model development with the Tree-LSTM network and BPNN is illustrated in Fig. S2. For supporting the development of the predictive model, molecular features were firstly extracted from the molecules of the collected dataset. Afterwards, the signature trees of compounds were generated for further mapping to the Tree-LSTM networks. Therefore, the vectors of molecular features can be inputted into the Tree-LSTM networks, and a vector was generated as an input for the BPNN. Within the BPNN, the properties were correlated to the molecular structures, and the QSPR model was obtained after massive training and testing. Afterwards, the QSPR model was evaluated with an external set, discussed on its

Applicability Domain and compared with the reported model to investigate its performance. As such, an accurate and reliable QSPR model was generated for predicting the log K_{OW} of organic compounds.

The algorithm of the proposed model for predicting with the Tree-LSTM network and BPNN is illustrated in Fig. S3. During predicting with the developed QSPR model, the molecular structure of a new compound is used to generate the signature tree which can be mapped to the Tree-LSTM network. Afterwards, using the vectors which were generated during the model development, the Tree-LSTM network outputs a vector integrated the features of the molecular structure for the BPNN. Relying on the parameters and hyper-parameters of the BPNN determined during model development, the BPNN makes a numeric prediction and outputs a predicted value for the log KOW of the compound.

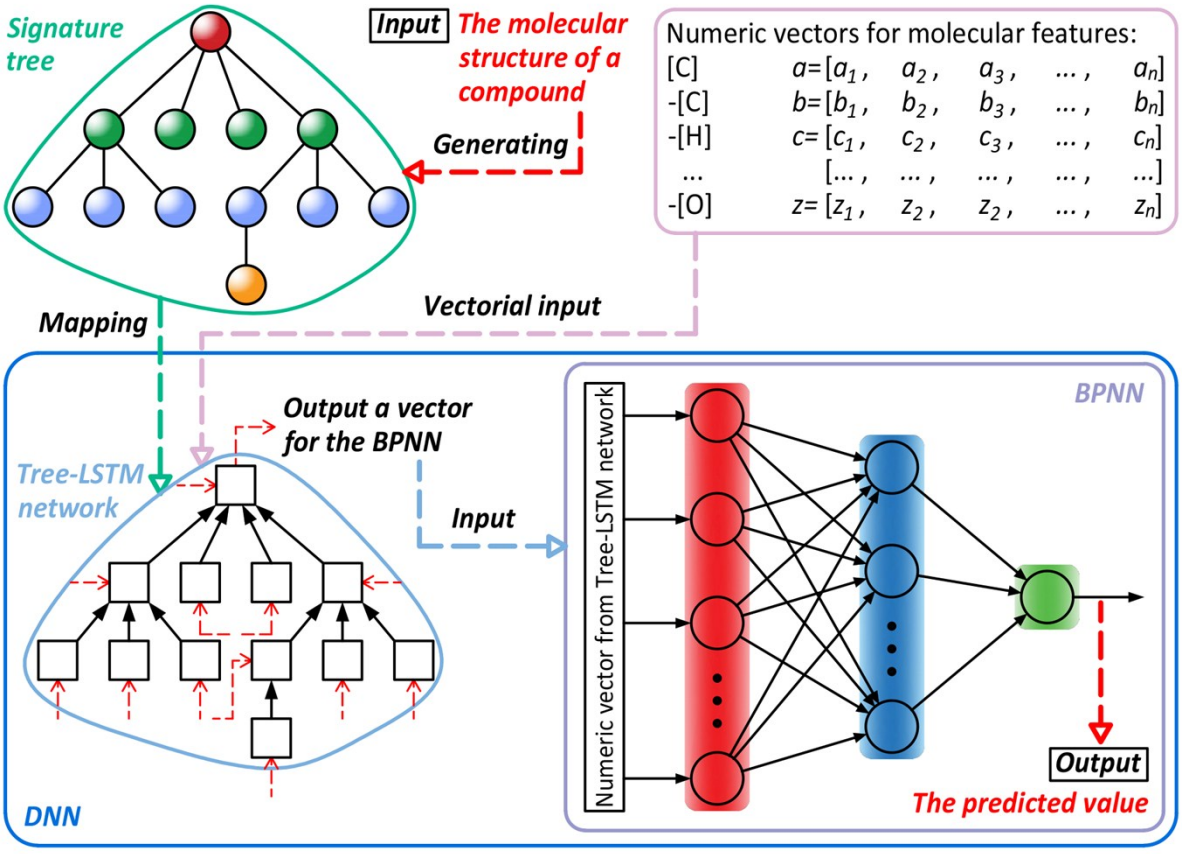

Fig. S3 The algorithm of the model for predicting with the Tree-LSTM network and BPNN.

External validation on the external set

Apart from the training and test sets, an external set was adopted as the additional test set to evaluate the predictivity of the final model in this research. According to the research of Chirico et al.,[1] the external validation indices (i.e., root mean square error (*RMSE*), mean absolute error (*MAE*) and determination coefficient (R^2)) have been calculated to measure the performance of the predictive model as summarized in Table S3.

Table S3 The statistics results for the ISO-DNN model on the training and external sets.

	N	*RMSE*	*MAE*	R^2
Training set	8534	0.2836	0.2101	0.9741
External set	1067	0.4656	0.3355	0.9285

[a] The number of data points;

[b] $RMSD = \sqrt{\sum_{n=1}^{N}(x_n^{exp} - x_n^{pre})^2 / N}$;

[c] $MAE = \frac{1}{N}\sum_{n=1}^{N}\left|x_n^{exp} - x_n^{pre}\right|$;

[d] $R^2 = 1 - [\sum_{n=1}^{N}(x_n^{exp} - x_n^{pre})^2 / \sum_{n=1}^{N}(x_n^{exp} - \mu)^2]$ (where $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n^{exp}$).

As it turns out, the RMSE and MAE of the training set is lower than those of the external set, and the R2 of the training set is closer to 1.0000. It indicated that the developed QSPR model can make more accurate predictions for the training set. In view of that the similar results of these external validation indices between the training and external sets, the developed predictive model is acceptable and it has satisfactory applicability.

Application Domain

The predictions can be considered reliable for the compounds which fall in the Applicability Domain (AD) of the predictive model. The Williams plot is a recommended leverage approach for AD investigation which provides a graphical detection of both the response outliers and the structurally influential outliers in a predictive model. In this research, the AD of the developed model was visualized with the Williams plot which is displayed with the plot of standardized residuals versus hat values as exhibited in Fig. S4.

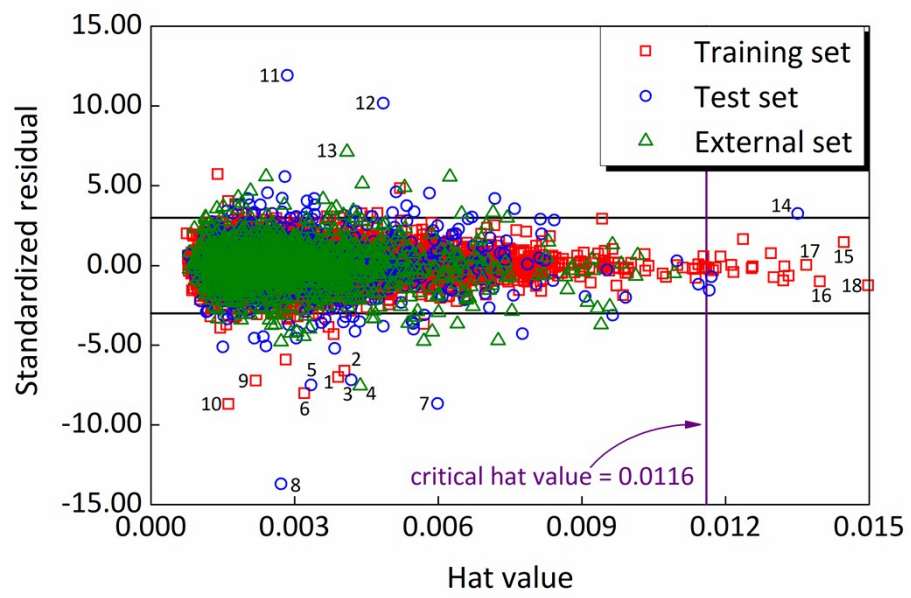

Fig. S4 The Applicability Domain of the developed QSPR model.

In the Williams plot, the compounds with standardized residuals greater than 3 standard deviation units are identified as response outliers. Part of the response outliers of the developed predictive model were marked with numbers and detailed as follows: cephaloridine (1), methyl 3-[5-acetyl-2-[2-[[3-ethyl-5-[(3-ethyl-4-methyl-5-oxopyrrol-2-yl) methylidene]-4-methylpyrrol-2-ylidene]methyl]-3-methyl-4-oxo-1H-cyclopenta[b]pyrrol-6-ylidene]-4-methyl-3,4-dihydropyrrol-3-yl]propanoate (2), sarmoxicillin (3), (4aR,6R,7R,7aS)-6- (6-amino-2-bromopurin-9-yl)-2-hydroxy-2-oxo-4a,6,7,7a-tetrahydro-4H-furo[3,2-d][1,3,2] dioxaphosphinin-7-ol (4), prolylphenylalanine (5), thienylglycine (6), 8-thiomethyl cyclic AMP (7), 1-bromo-4-[2-[2-(4-methoxyphenoxy)ethoxy]ethoxy]-2,5-dimethylbenzene (8), N,N-diethyl-3-methoxy-4-(2-hydroxy-5-sec-butylphenylazo)benzenesulfonamide (9), 1-butyl-5-

[[3-[(2-chlorophenyl)methyl]-4-oxo-2-sulfanylidene-1,3-thiazolidin-5-ylidene]methyl]-4-methyl-2-oxo-6-(4-phenylpiperazin-1-yl)pyridine-3-carbonitrile (10), (E)-4-(dimethylamino)-4-oxobut-2-enoic acid (11), (E)-2-cyano-3-amino-3-(isopropylamino)propenoic acid methyl ester (12) and [4-[2-(diaminomethylidene)hydrazinyl]phenyl]iminourea (13). Moreover, the hat value of a compound greater than the critical hat value indicates the compound is outside of the model's structural AD and it could lead to unreliable predictions. As it turns out, 22 compounds were detected as structurally influential outliers and part of them were marked with numbers and exhibited as follows: mellitic acid (14), perfluoromethylcyclohexylpiperidine (15), 2-nitrostrychnidin-10-one (16), strychnine (17) and perfluorocyclohexane (18). The way for calculating the standardized residual, hat value and critical hat value can be found in the published works.[2,3]

Impact of isomeric features on predictive accuracy

A comprehensive investigation found that 1663 out of the 10668 compounds contain isomeric features in developing the QSPR model and they were described with the isomeric features, while the remaining 9005 compounds described with the canonical SMILES strings contain no isomeric features.

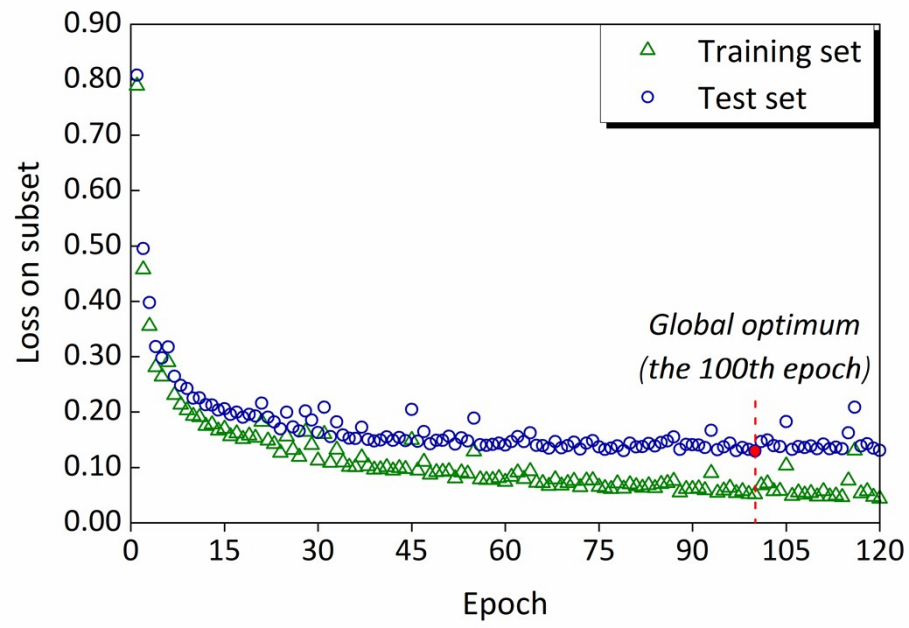

Fig. S5 The tendency in losses on the training and test sets during training for the CAN-DNN model.

In order to evaluate the impact of isomeric features on the predictive accuracy of the model, we replace all the isomeric SMILES strings with the canonical SMILES strings, and on this basis a new QSPR model was developed with the same training, test and external sets used for developing the ISO-DNN model. The QSPR model obtained in the 100th epoch was considered to be the global optimum model (refer to Fig. S5) and it was saved as the final model (represented as CAN-DNN model) for log KOW prediction. And the predictive performance of the training, test and external sets is visualized in Fig. S6.

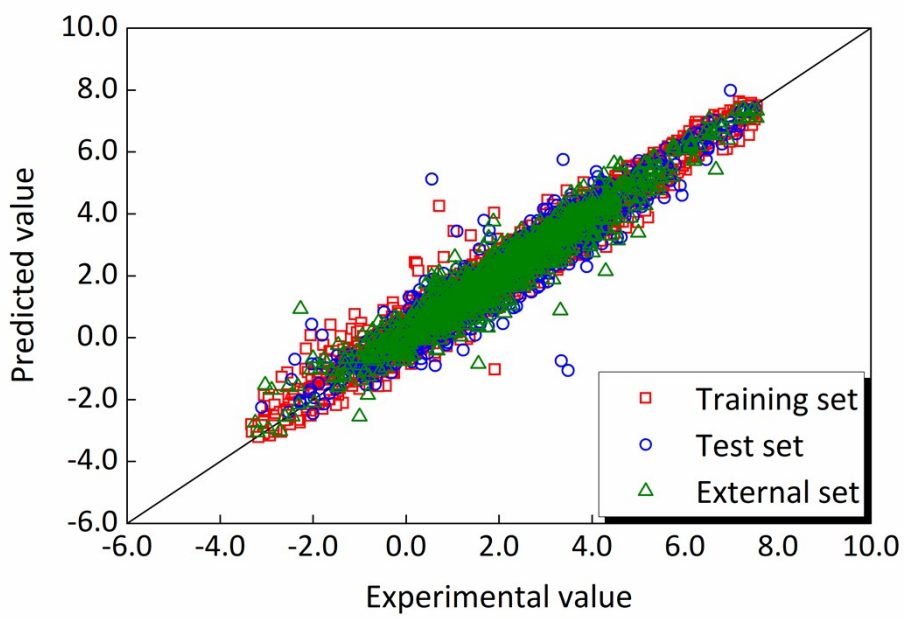

Fig. S6 The scatter plots of predicted - experimental value with the CAN-DNN model for (a) training set, (b) test set and (c) external set.

Furthermore, the predictive accuracy of the two model was measured with several statistical indices as summarized in Table S4. The results show that the predictive accuracy of the CAN-DNN model is lower than that of the ISO-DNN model. It proves that the isomeric features involving in model development are conducive to improve the accuracy of the predictive model.

Table S4 The statistics results for the ISO-DNN and CAN-DNN models in log K_{OW} prediction.

	N	*RMSE*	*MAE*	R^2
ISO-DNN	10668	0.3386	0.2376	0.9606
CAN-DNN	10668	0.3707	0.2634	0.9552

[a] The number of data points;

[b] $RMSD = \sqrt{\sum_{n=1}^{N}(x_n^{exp} - x_n^{pre})^2 / N}$;

[c] $MAE = \frac{1}{N}\sum_{n=1}^{N}\left|x_n^{exp} - x_n^{pre}\right|$;

[d] $R^2 = 1 - [\sum_{n=1}^{N}(x_n^{exp} - x_n^{pre})^2 / \sum_{n=1}^{N}(x_n^{exp} - \mu)^2]$ (where $\mu = \frac{1}{N}\sum_{n=1}^{N}x_n^{exp}$).

Additionally, the impact of isomeric features on the predictive accuracy of the model was investigated by comparing the predictive accuracy of the KOWWIN model and the developed ISO-DNN model performing on the 1663 isomeric compounds and the remaining 9005

compounds. As displayed in Table S5, the predictive accuracy on the 1663 isomeric compounds is much lower than that on the other 9005 compounds with the KOWWIN model. However, using the ISO-DNN model, the predictive accuracy on the 1663 isomeric compounds is close to that on the other 9005 compounds. Meanwhile, comparing with the KOWWIN model, the improvement in the predictive accuracy for the 1663 isomer compounds is markedly higher than the other 9005 compounds using the ISO-DNN model. It also demonstrated that the predictive accuracy on log KOW is significantly improved with the participation of the isomeric features. Therefore, the isomeric features are considered beneficial to improve the predictive accuracy of the model.

Table S5 The statistics results for the KOWWIN and ISO-DNN models in log K_{OW} prediction for isomeric compounds and other compounds.

	Isomeric compounds		Other compounds	
	KOWWIN	ISO-DNN	KOWWIN	ISO-DNN
N [a]	1663	1663	9005	9005
$RMSE$ [b]	0.5563	0.3957	0.3928	0.3269
MAE [c]	0.4060	0.2546	0.2857	0.2345
R^2 [d]	0.9352	0.9619	0.9480	0.9603

[a] The number of data points;

[b] $RMSD = \sqrt{\sum_{n=1}^{N}(x_n^{exp} - x_n^{pre})^2 / N}$;

[c] $MAE = \frac{1}{N}\sum_{n=1}^{N}\left|x_n^{exp} - x_n^{pre}\right|$;

[d] $R^2 = 1 - [\sum_{n=1}^{N}(x_n^{exp} - x_n^{pre})^2 / \sum_{n=1}^{N}(x_n^{exp} - \mu)^2]$ (where $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n^{exp}$).

References

1 N. Chirico and P. Gramatica, *J. Chem. Inf. Model.*, 2012, **52,** 2044-2058.

2 A. Rybinska, A. Sosnowska, M. Grzonkowska, M. Barycki and T. Puzyn, *J. Hazard. Mater.*, 2016, **303**, 137-144.

3 P. P. Roy, S. Kovarich and P. Gramatica, *J. Comput. Chem.*, 2011, **32**, 2386-2396.