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.
E-mail: 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

† Electronic supplementary information (ESI) available: Detailed analysis in the
types of compounds, frequencies of compounds presenting molecular features,
feature selection during the property predictions, algorithms of model development and predicting, external validation on the external set, application domain
and impact of isomeric features on predictive accuracy.

‡These authors contributed equally to this work.

Abstract
As an essential environmental property, octanol-water partition coefficient (KOW) quantifies the lipophilicity of a compound and it could be further employed to predict the toxicity. Thus, it is an indispensable factor should be considered in screening and development of green solvents with respect to unconventional and novel compounds. Herein, a deep-learning-assisted predictive model has been developed to accurately and reliably calculate log KOW values for organic compounds. An embedding algorithm was specifically established for generating signatures automatically for molecular structures to express structural information and connectivity. Afterwards, the Tree-structured long short-term memory (Tree-LSTM) network was used in conjunction with signature descriptor for automatic feature selection, and it was then coupled with the back-propagation neural network to develop a deep neural network (DNN), which is used for modeling quantity structure-property relationship (QSPR) to predict log KOW. Comparing with an authoritative estimation method, the proposed DNN-based QSPR model exhibited the better predictive accuracy and greater discriminative power in terms of the structural isomers and stereoisomers. As such, the proposed deep learning approach can act as a promising and intelligent tool for developing environmental property prediction methods for guiding development or screening of green solvents.

Introduction

As one of the cornerstones for the sustainable development,[1] environmental benefit drives chemical process technology and environmental science toward environmentally friendly technology.[2] The environmental impact is an indispensable factor that should be considered in the molecule design, chemical synthesis and solvent selection.[3-6] As an essential environmental property, the lipophilicity refers to the affinity of a compound for lipids and provides valuable information about the absorption, distribution and metabolism of compounds.[7-9]

Usually, the lipophilicity of a compound is measured as its partition coefficient between lipid and aqueous phases. Octanol has been widely accepted as a token representing cell membranes, tissue and lipids[7] and the partition coefficient between octanol and aqueous phases is frequently adopted as a measure for the lipophilicity of organic chemicals[9-11] as well as a physicochemical criterion for solvent selection.[12,13] This partition coefficient could be further employed to predict various indicators of toxicity (*e.g.* 50% effective concentration (EC50) and 50% lethal concentration (LC50)).[14,15]

The octanol-water partition coefficient (K_{OW}) describes the distribution of a substance between the octanol and aqueous phases in a two-phase octanol-water system at equilibrium.[16] It can be ideally measured by experiments, but the existing database is not enough for many compounds of interest. Additionally, the experimental determinations are not always feasible for those compounds with low water solubility. In this context, various methods that rely on property estimation were developed and continue to be proposed to solve the problems of generality and accuracy.

Many researchers reviewed the existing models on predicting log K_{OW} and they highlighted the frontiers and prospects of developing prediction approaches in this respect.[17-19] Additionally, some investigators elaborated the state-of-the-art and assessed the performances of the representative log K_{OW} predictive models.[20,21] A large number of studies have focused on developing empirical relationship models in which K_{OW} is described as a function of molecular physicochemical properties.[8,22,23] These empirical methods can be efficient in the computation but heavily rely on correlated properties which are not always available. In contrast, the quantity structure-property relationship (QSPR) methods, such as group contribution (GC) methods and topological methods, are more readily implemented since only structural information needs to be provided.[24]

Prior to the GC methods, atom and fragment contribution methods laid a solid foundation for the prediction of properties.[10,25] A molecule can be divided into atoms and fragments without any ambiguity and as such these methods achieved success in property estimation. However, molecules are much more than a collection of atoms.[21] In this regard, as extensions of atom and fragment contribution methods, modified GC methods have been put forward with the purpose of predicting properties for organic chemicals.[26-29] In the GC methods, various groups (*e.g.* substructures containing atoms and bonds) can be defined, and the target property value of a compound is given by summarizing the contributions of groups. A typical example is the three-level GC method proposed by Marrero and Gani[9,28] and it was applied to estimate K_{OW}. This method showed a better predictive accuracy regarding a large quantity of organic compounds. On the other hand, topology is an unambiguous feature and topological properties can be directly derived from molecular structures.[21] Therefore, different topological characteristics have been extensively adopted as descriptors to develop QSPR models and correlate properties.[30-33]

Although the GC and topological methods have revealed satisfactory performance in property estimation, a few shortcomings have limited their extensive applications, such as the limited discriminative power in isomers and the inadequate consideration in the holistic molecular structures.[34] To overcome these shortcomings, signature molecular descriptors were introduced which could capture a whole picture with connectivity information of each atom for a molecule.[35,36] This signature descriptor was developed specifically for molecular structures and it could be a potential tool for QSPR modelling without the need for calculation and selection of other numerical descriptors.[34] Moreover, it was further detailed and applied in QSPR researches.[37,38]

Meanwhile, a major expansion has appeared in the field of QSPR due to the advent of artificial intelligence (AI). Artificial neural networks have been extensively employed to determine the correlations between molecular structures and properties.[30,39-43] In this respect, based on the long short-term memory (LSTM),[44] the traditional LSTM network is usually structured as a linear chain and this exhibited the superior representational power and effectiveness. However, some

types of data (such as text) are better represented as the tree structures. In this context, an advanced Tree-structured LSTM (Tree-LSTM) network was put forward as a variant of the LSTM network to capture the syntactic properties of natural languages.[45] With regard to the complex and various molecular structures, the Tree-LSTM network is supposed as an attractive option in representing the relationships of the atoms or groups.

Recently, a deep learning approach for predicting the properties of chemical compounds was proposed, in which the Tree-LSTM network was successfully implemented with the purpose of expressing and processing chemical structures.[46] Additionally, taking the critical properties as examples, it proved that the proposed deep learning approach is suitable for a more diverse range of molecular structures and enables users to achieve more accurate predictions.

Based on the state-of-the-art, there are still three issues to be solved in K_{OW} estimation for organic compounds, and they are:

(i) Human intervention was involved in the feature selection of the molecular structures during the model development, which caused the omission of the important molecular information;

(ii) Too many topological features or physicochemical descriptors have been adopted, increasing the complexity of models and decreasing the computational efficiency;

(iii) The ability of differentiating structural isomers and stereoisomers is limited in the reported QSPR models, which constrains the application scope of the predictive model.

In order to overcome the three challenges, and motivated by the successful deep learning approach in property estimation, a QSPR model was developed in this research to accurately predict K_{OW} values for organic compounds and provide the valuable environmental information for guiding the selection and development of the important chemicals including green solvents. In this model, the automatic feature selection for molecules was achieved by coupling the canonical molecular signature and deep neural network (DNN).

Methodology

A DNN model, which couples the Tree-LSTM network and back-propagation neural network (BPNN), was developed in this work based on the deep learning approach. It was built to specialize in the determination of the correlation between molecular structures and log K_{OW} values of organic compounds. The process of developing a reliable QSPR model with the DNN model is comprised of the following five basic steps, as illustrated in the Fig. 1.

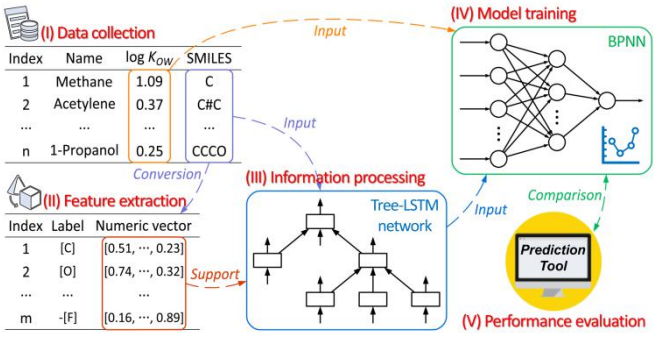

Fig. 1 The schematic diagram of the process for developing a QSPR model with the deep learning approach.

Step 1: Data collection.

The experimentally measured log K_{OW} values and simplified molecular-input line-entry system (SMILES) strings of compounds were collected since they are necessary for developing a QSPR model. Herein, the SMILES strings sufficed for representing the basic molecular structural information.

Step 2: Feature extraction.

The SMILES strings of compounds were utilised to generate a list of numeric vectors based on a proposed atom embedding algorithm which was implemented with the atomic signatures. The

vectors are able to describe molecular structures and represent their features.

Step 3: Information processing.

The SMILES strings were converted to canonical molecular signatures with the theory of the canonizing molecular graph.[47] On this basis, these signatures were mapped on the Tree-LSTM networks with the aim of creating vectors as inputs for the BPNN.

Step 4: Model training.

After receiving the inputs from the Tree-LSTM networks, the BPNN supported the correlation process and it was repeatedly run to learn a satisfactory QSPR model. In the training process, parameters were updated to optimize the parameters of the DNN model and finally the QSPR model with better performance was preserved for log K_{OW} estimation.

Step 5: Performance evaluation.

Based on the developed QSPR model, the generalization ability was assessed by the predictive performance of an external dataset. And the external competitiveness of the QSPR model was evaluated by comparing to an authoritative predictive model.

All the above steps for obtaining the QSPR model to predict the K_{OW} were achieved with a series of programs which were written in the Python language and successfully tested on Windows platforms.

Data acquisition and processing

The dimensionless K_{OW} values span over ten orders of magnitude and therefore the decimal logarithm of K_{OW} (log K_{OW}) was frequently adopted in property estimation. A large number of experimentally measured log K_{OW} values of chemical compounds were collected,[48] and all the experimental values were originated from references to guarantee the reasonability of the predictive model. To investigate the QSPR model for organic compounds, a number of irrelevant compounds were eliminated. The excluded irrelevant compounds involve the inorganic compounds (*e.g.* carbon dioxide, sulfur hexafluoride and hydrazine), metal-organic compounds (*i.e.*, the organic compounds containing metal atoms such as sodium, chromium or/and stannum) and mixtures consisting two or more compounds. Hence, the remaining 10754 pure organic compounds were assembled for the model development.

As a large dataset was collected, the data cleaning is essential to be carried out by detecting and removing outliers which contain gross errors. Accordingly, the Pauta criterion,[49] also referred to as the three sigma rule, was applied for the cleaning process. It describes that 99.73% of all values of a normally distributed parameter fall within three times the standard deviation (σ) of the average (μ). Any error beyond this interval is not a random error but a gross error. Accordingly, data points which include gross error are regarded as outliers and should be excluded from the sample data. The data cleaning process with Pauta criterion is graphically illustrated in the Fig. 2.

As a result, 86 out of 10754 organic compounds (about 0.8 percent of the dataset) were detected as outliers based on their experimental values and they were removed from the dataset. The remaining 10668 organic compounds were preserved as the final dataset for developing a QSPR model to predict log K_{OW}. The dataset of compounds spans a wide class of molecular structures including aliphatic and aromatic hydrocarbons, alcohols and phenols, heterocyclic compounds, amines, acids, ketones, esters, aldehydes, ethers and so on. A detailed analysis of the types of compounds is presented in Table S1 of the Electronic Supplementary Information (ESI).

In addition to the experimental values, the information of molecular structures is also indispensable in developing a QSPR model. SMILES[50] is a chemical language representing structural information in the text form and it is widely applied in the chemo-informatics software because it can be employed to build molecular two-dimensional or three-dimensional structures. Moreover, one can manually provide the SMILES string of any compound after simply learned the encoding rules.

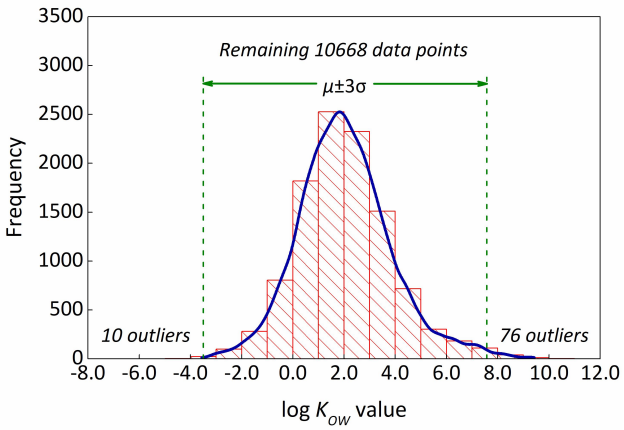

Fig. 2 The distribution of experimental data of 10754 organic compounds.

The open chemistry database, PubChem,[51] contains the largest collection of publicly available chemical information and provides two types of SMILES strings (*i.e.*, canonical SMILES and isomeric SMILES) for tens of millions of compounds. The canonical SMILES strings are available for all the compounds, whereas the isomeric SMILES strings which contain isomeric information are only provided for isomers. With respect to the dataset applied in this work, the SMILES string of each compound was derived from the PubChem according to its chemical abstracts service registry number (CAS). During the SMILES acquisition, the isomeric SMILES string was collected if available. Otherwise, the canonical SMILES string was adopted. Eventually, the experimental values and SMILES strings of 10668 organic compounds were adopted as the inputs for developing the QSPR model.

Tree structures in information processing

The signature molecular descriptor was introduced specifically for describing molecular structures, and all the connectivity information for every atom in a molecule was retained. Additionally, it can be theoretically applied to represent any organic compound which means that it is able to cover various molecular structures without limitation.

Herein, taking 1-propanol (CAS: 71-23-8; SMILES: CCCO) as an example. When a root atom was specified in the molecule, a tree spanning all atoms and bonds of the molecule was constructed (refer to Fig. 3(a)), and the signatures were generated relying on the theory of canonizing molecular graph.[47]

Up to a point, the syntactic property of natural languages is analogous to the connectivity information for atoms in a molecule. The former one is able to be captured by the Tree-LSTM network while the later one can be expressed with a signature. In addition, the tree structure of Tree-LSTM network (refer to Fig. 3(b)) is similar to the signature tree displayed in the Fig. 3(a). Therefore, it was assumed that the molecular structural information can be processed and transmitted by coupling the signatures and Tree-LSTM network, and this was proven to be practical.[46]

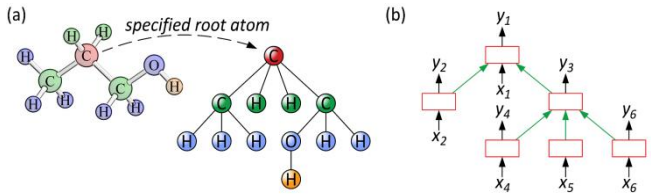

Fig. 3 The tree structures in expressing information of (a) the signature tree for the 1-propanol molecule and (b) the Tree-LSTM network.

Signature molecular descriptor and encoding rules

The structural information of molecules was extracted from the SMILES strings and expressed by atomic and molecular signatures with text form in this work. The atomic signatures can represent

the substructure of a molecule while the molecular signatures describe the whole molecular structure. To specify atomic features, atoms were converted to strings relying on encoding rules which refer the regulations defined in SMARTS[52] (a straightforward extension of SMILES for describing molecular substructures). In order to be well applied in this task, some new definitions were made as a complement of the encoding rules. Herein, RDKit[53] was adopted as an auxiliary tool for implementing the encoding rules by identifying the element symbols of atoms, the types of chemical bonds, the types of chirality centres and so forth.

Atomic signature of height 1, also called 1-signature, contains only the root atom and its chemical bonds along with connected atoms (refer to Fig. 4).[36] The 1-signature of each atom in molecules were generated with encoding rules, and subsequently a series of substrings representing molecular features were extracted with adopting the atom embedding program.[46] During the embedding process, each substring was assigned a numeric vector for distinction and adopted as the label for this vector. In spite of that these vectors were only used to represent molecular features. The structural information of molecules and atom connectivity will be totally preserved with the aid of the combination of signatures and the Tree-LSTM networks. For illustrative purpose, all the symbols involving in the labels of molecular features are listed and explained in Table 1.

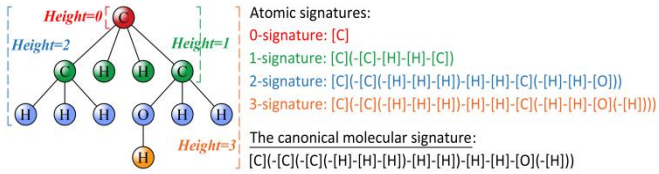

Fig. 4 The signature descriptors generated from the 1-propanol molecule.

The molecular signature was defined as the linear combination of atomic signatures covering all the atoms and bonds.[36] However, the molecular signatures involve redundant and duplicated information. Accordingly, canonical molecular signature, the lexicographically largest atomic signature, which suffices to represent the molecular graph, was introduced to simplify the molecular signature.[47] Herein, to be used in conjunction with the Tree-LSTM network, the canonical molecular signature of each compound was generated in a unique manner for describing the molecular structure. For instance, the canonical molecular signature for 1-propanol (CAS: 71-23-8; SMILES: CCCO) is represented as [C](-[C](-[C](-[H]-[H]-[H])-[H]-[H])-[H]-[H]-[O](-[H])) relying on the canonizing algorithm[46] and proposed encoding rules.

Table 1 The explanations and examples for symbols involved in the labels of molecular features.

Symbol	Explanation	Example
[A]	atom in aliphatic compound	[C] - carbon atom in an aliphatic compound
[a]	atom in aromatic compound	[c] - carbon atom in an aromatic compound
|r	atom in a ring	[C|r] - carbon atom in a ring
+ (inside [ ])	atom with a positive charge	[N+] - nitrogen atom with a positive charge
- (inside [ ])	atom with a negative charge	[N-] - nitrogen atom with a negative charge
- (outside [ ])	single bond	-[C] - carbon atom with a single bond
=	double bond	=[C] - carbon atom with a double bond
#	triple bond	#[C] - carbon atom with a triple bond
:	aromatic bond	:[c] - carbon atom with an aromatic bond
/=\	atoms in same side	/=\[C] - carbon atom in same side of connected atom
/=/	atoms in opposite side	/=/[C] - carbon atom in opposite side of connected atom
*	atom is a r-chirality center	[C*] - carbon atom is a r-chirality center
**	atom is a s-chirality center	[C**] - carbon atom is a s-chirality center

Table 2 The labels of molecular features classified by the chemical elements for representing molecular structures.

Chemical elements	Labels of molecular features
Carbon (C)	[C]; -[C]; =[C]; #[C]; [C|r]; -[C|r]; =[C|r]; [C*]; -[C*]; [C|r*]; -[C|r*]; [C**]; -[C**]; [C|r**]; -[C|r**]; [c|r]; -[c|r]; =[c|r]; :[c|r]; /=\[C]; /=/[C]; /=\[C|r]; /=/[C|r]; /=/[c|r]
Oxygen (O)	[O]; -[O]; =[O]; [O|r]; -[O|r]; [o|r]; :[o|r]; [O-]; -[O-]
Nitrogen (N)	[N]; -[N]; =[N]; #[N]; [N|r]; -[N|r]; =[N|r]; [n|r]; -[n|r]; :[n|r]; [N+]; -[N+]; =[N+]; #[N+]; [N-]; =[N-]; [N+|r]; -[N+|r]; =[N+|r]; [n+|r]; -[n+|r]; =[n+|r]; :[n+|r]; /=\[N]; /=/[N]; /=\[N+]
Phosphorus (P)	[P]; -[P]; =[P]; [P|r]; -[P|r]; =[P|r]; [P+]; -[P+]; [P+|r]; -[P+|r]
Sulphur (S)	[S]; -[S]; =[S]; [S|r]; -[S|r]; =[S|r]; [s|r]; :[s|r]
Others except as above	[H]; -[H]; [F]; -[F]; [Cl]; -[Cl]; [Br]; -[Br]; [I]; -[I]

Structural features and parameters of DNN

In the DNN model, the Tree-LSTM network was utilised in conjunction with the BPNN to develop a QSPR model for predicting log K_{OW}. The Tree-LSTM network was employed to describe molecular tree structures with canonical molecular signatures while the BPNN was used to correlate structures and properties. Back-propagation (BP) algorithm is a supervised learning procedure in the machine learning process and it was commonly used to train the DNN.[54-56] In this work, the BPNN was built with three layers including one input layer, one hidden layer and one output layer. The topological structure of the fully connected three-layer neural network is graphically presented in the Fig. 5. The input layer receives the vectors produced by Tree-LSTM network and the output layer gives the predicted log K_{OW} values. As single layers of linear neurons, the hidden layer take in a set of weighted inputs from the input layer and produce an output for the output layer.

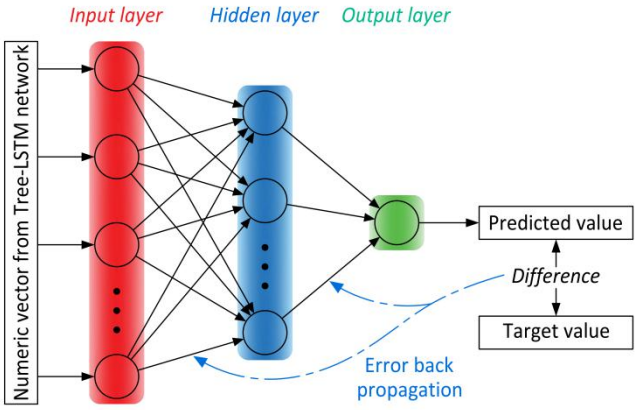

Fig. 5 The structure of the fully-connected BPNN model for log K_{OW} prediction.

As an open-source deep learning library for Python, PyTorch[57] mainly supported the development of the DNN model in this research. Huber loss is a common loss function which is characterized by rapid convergence and inclusiveness to outliers because it combines the advantages of two basic loss functions, *i.e.*, the mean square error and the mean absolute error. Therefore, the Huber loss[58] was adopted as the loss function in this research to evaluate the model performance during the training process. Additionally, Adam algorithm[59] was employed for optimizing the predictive model by minimizing the loss function due to the attractive benefits that it is computationally efficient and suitable for tasks with a large dataset.

A machine learning model is parameterized and it refers to numerous variables that can be classified into two types, model parameters and model hyper-parameters. The model parameters, such as weights and biases, are learned from the given dataset and updated with the BP algorithm

by calculating the gradient of the loss function during the model training. In contrast, with the purpose of controlling the learning process efficiently, model hyper-parameters were specified before the training activates.

In order to achieve the better performance as well as make the DNN model specialize in the prediction of log K_{OW}, hyper-parameters were specified and detailed as follows:

(i) The hidden layer of the BPNN has 32 neurons;

(ii) The batch size of the set for training, the number of training examples utilised in one iteration, is set as 250;

(iii) The Learning rate is set as 1.00E-03 to control the rate of convergence;

(iv) The weight decay rate is set as 1.00E-06 to alleviate the problem of the over-fitting.

Results and Discussion

List of molecular features

To ensure that molecular structures can be introduced to the Tree-LSTM network, all the atoms and bonds of a molecule need to be expressed in the form of numerical vectors. Initially, the 1-signatures of every atom in the 10668 chemical compounds were obtained relying on the aforementioned encoding rules and signature descriptors. Subsequently, 1-signatures were taken as samples to produce substring vectors with the atom embedding program. Finally, as exhibited in Table 2, 87 types of molecular features were extracted and assembled as a list for representing molecular structures in the Tree-LSTM networks. The detailed count for the number of compounds in the training set presenting each molecular feature is shown in the ESI (Table S2). Moreover, the way these features were chosen in the predictive model is detailed in the ESI (Page S3).

Training process

A network is said to generalize when it appropriately predicts the properties of compounds which are not in the set for training.[60] To measure the generalization ability of a QSPR model, an external dataset which is not used for training the QSPR model should be employed to evaluate the model performance. Consequently, the entire dataset containing 10668 organic compounds was divided into three disjoint subsets (*i.e.*, a training set, a test set and an external set) by a random selection routine according to their corresponding proportions (80%, 10% and 10%). The training set (8534 compounds), test set (1067 compounds) and external set (1067 compounds) were used to build and optimize the QSPR model, determine the timing for stopping training and measure the external predictive performance of the final model, respectively.

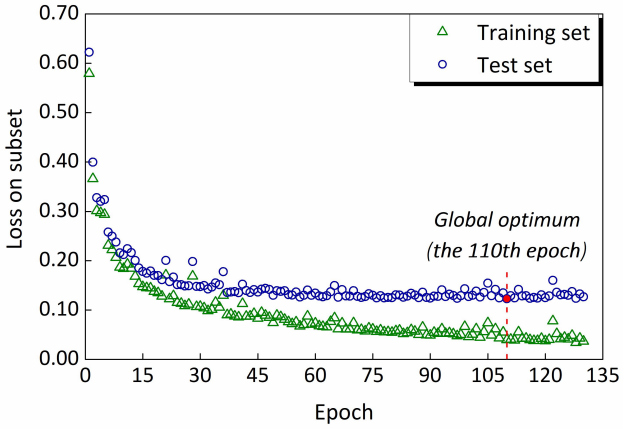

Fig. 6 The tendency in losses on training and test sets during training.

In the DNN model, the canonical molecular signatures were fed into the Tree-LSTM networks to describe molecular tree structures, and subsequently the BPNN was used to correlate molecular structures and log K_{OW} values. In the training process, the Huber loss function was minimized with the Adam algorithm to receive a better QSPR model which can provide more accurate predictions. The learning process proceeded epoch by epoch, and the losses on training set and

test set were calculated at the end of each epoch to evaluate the model performance and determine the timing for stopping training. Once the training process is activated, it does not terminate until there is no decrease in the loss on the test set within 20 consecutive epochs. Furthermore, early stopping was used to prevent the problem of the over-fitting and improve the performance of model on data outside of the training set. The QSPR model was preserved by storing the topological structure of the DNN on each epoch during the training process, and the final QSPR model can be rebuilt with its corresponding structural parameters of the DNN.

The training was terminated at the 130th epoch, and the tendency in losses on training and test sets is displayed in the Fig. 6. Since the 110th epoch, the loss on the training set significantly decreased while the loss on the test set kept at the same level in 20 consecutive epochs which means that the model was over-fitting during these epochs. Accordingly, the QSPR model obtained in the 110th epoch was considered to be the global optimum model and it was saved as the final model for log K_{OW} prediction to prevent an over-fitting model. Additionally, the algorithms of model development and prediction were provided in the ESI (Pages S4 and S5).

Generalization ability

The generalization ability acts as an important indicator to evaluate the predictive performance of a model in machine learning. The problem of over-fitting in the DNN model is bound to cause the loss of generalization ability and low external prediction accuracy.[61] The traditional models for property prediction were developed with the entire dataset for training, and therefore the generalization ability was unable to be guaranteed around two decades ago. However, efforts have been made to avoid the publication of models without external validations and the generalization ability of predictive models has been highly improved in the last decade.[62,63]

As stated, early stopping was used in this work to avoid over-fitting and ensure that the final QSPR model has satisfactory generalization ability. In addition, the external set was employed to measure the generalization ability of the final QSPR model. The training and external sets were applied in the final QSPR model, and the predicted values were compared against the original experimental values. The predictive performance is visualized in the Fig. 7 with the scatter graphs of predicted values versus experimental values. It is observed that the distribution and predictive accuracy of the external set are similar to those of the training set. It demonstrated that the developed QSPR model has the satisfactory generalization ability in predicting log K_{OW} values for organic compounds. Furthermore, the external validation indices[64] have been calculated to measure the performance of the predictive model on the external set in the ESI (Page S6).

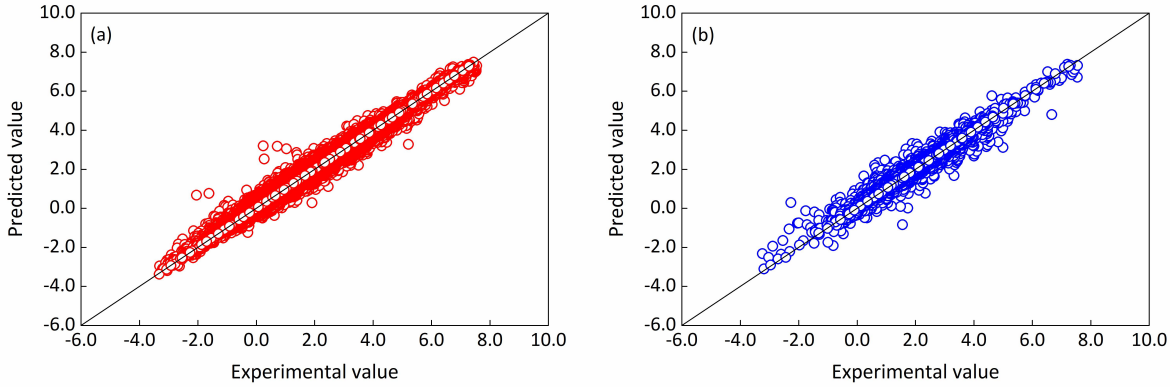

Fig. 7 The scatter plots of predicted - experimental value with DNN model for (a) training set and (c) external set.

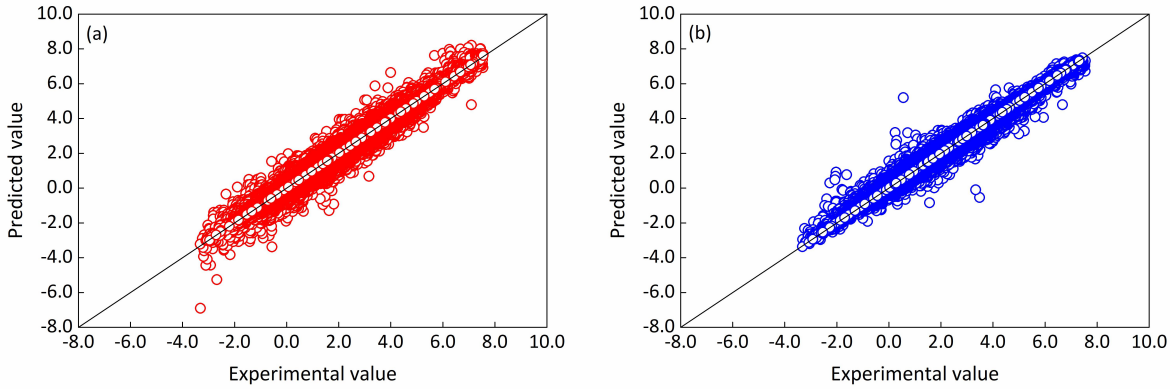

Fig. 8 The scatter plots of predicted - experimental value with (a) KOWWIN model and (b) ISO-DNN model.

Applicability 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 predictive model is visualized with the Williams plot which is displayed with the plot of standardized residuals versus hat values. The discussion on the AD of the predictive model is provided in the ESI (Pages S7 and S8). Moreover, the way for calculating the standardized residual, hat value and critical hat value can be found in the published works.[65,66]

External competitiveness

The satisfactory predictive capability of a new QSPR model needs to be proven by its external competitiveness. As an authoritative log K_{OW} prediction tool relying on an atom and fragment contribution method, the KOWWIN program was developed and maintained by the United State Environmental Protection Agency and the Syracuse Research Corporation.[65] In this research, a comparison between the KOWWIN and developed QSPR model (represented as ISO-DNN model) was conducted to measure their predictive capabilities. It should be noted that it is not always possible to compare the performance of different models unless they are evaluated using the same dataset. Therefore, a dataset of KOWWIN predicted values was collected.[47] For a fair comparison, the samples in this dataset are consistent with the whole dataset adopted in this work. The overall predictive capabilities of the ISO-DNN and KOWWIN models were further assessed and compared in the form of scatter graphs (refer to Fig. 8). Overall, the data points in the Fig. 8(b) were closer to the diagonal line (predicted value equals experimental value) when compared to the data points in Fig. 8(a). This suggests that the ISO-DNN model has better predictive accuracy in log K_{OW} estimation, although some data points still exhibited relatively large deviations.

The scatter plots can only appear to indicate a better predictive accuracy of the ISO-DNN model. It is more persuasive if the external competitiveness of the developed QSPR model can be demonstrated from perspective of statistics. The residual (experimental value minus predicted value) of each compound in the dataset was calculated using the KOWWIN and ISO-DNN models. The residual distributions in the Fig. 9 show that the residuals produced by the ISO-DNN model are more densely gathered around the zero value in contrast with those obtained using the KOWWIN model. This indicates that the residual distribution of the ISO-DNN model has a lower standard deviation. According to the analysis, it was further proven that the develop QSPR model enables a more accurate prediction for log K_{OW} from a statistical point of view.

Table 3 The statistical results of the KOWWIN and ISO-DNN models in log K_{OW} prediction.

Predictive model	N [a]	$RMSE$ [b]	MAE [c]	R^2 [d]
KOWWIN	10668	0.4224	0.3045	0.9451
ISO-DNN	10668	0.3386	0.2376	0.9606

[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 $\mu = \frac{1}{N}\sum_{n=1}^{N} x_n^{exp}$).

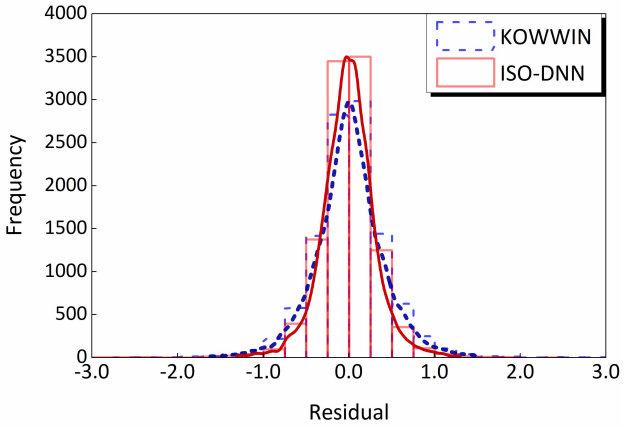

Fig. 9 The residual distributions of log K_{OW} estimation with KOWWIN and ISO-DNN models.

Table 4 The capabilities of the KOWWIN and ISO-DNN models in distinguishing the structural isomers.

Chemical name	CAS [a]	x^{exp} [b]	$x^{pre\text{-}K}$ [c]	$x^{pre\text{-}D}$ [d]	Δx_K [e]	Δx_D [f]
Pyridazine-4-carboxamide	88511-47-1	-0.96	-1.31	-0.75	0.35	0.21
Pyrimidine-5-carboxamide	40929-49-5	-0.92	-1.31	-0.63	0.39	0.29
Thiophene-2-carboxylic acid	527-72-0	1.57	1.69	1.61	0.12	0.04
Thiophene-3-carboxylic acid	88-13-1	1.50	1.69	1.49	0.19	0.01

[a] Chemical abstracts service registry number; [b] Experimental value; [c] KOWWIN predicted value; [d] DNN predicted value; [e] $\Delta x_K = |x^{exp} - x^{pre\text{-}K}|$; [f] $\Delta x_D = |x^{exp} - x^{pre\text{-}D}|$.

Table 5 The capabilities of the KOWWIN and ISO-DNN models in distinguishing the stereoisomers.

Chemical name	CAS [a]	x^{exp} [b]	$x^{pre\text{-}K}$ [c]	$x^{pre\text{-}D}$ [d]	Δx_K [e]	Δx_D [f]
(2R,6R)-2,6-dimethylcyclohexan-1-ol	39170-83-7	2.10	2.47	2.55	0.37	0.45
(2R,6S)-2,6-dimethylcyclohexan-1-ol	39170-84-8	2.37	2.47	2.38	0.10	0.01
(1R)-cis-(alphaS)-cypermethrin	65731-84-2	6.05	6.38	5.78	0.33	0.27
(1R)-trans-(alphaS)-cypermethrin	65732-07-2	6.06	6.38	6.05	0.32	0.01

[a] Chemical abstracts service registry number; [b] Experimental value; [c] KOWWIN predicted value; [d] DNN predicted value; [e] $\Delta x_K = |x^{exp} - x^{pre\text{-}K}|$; [f] $\Delta x_D = |x^{exp} - x^{pre\text{-}D}|$.

Furthermore, the model performance was also quantified using three evaluation indexes, the root mean squared deviation (*RMSD*), average absolute error (*AAE*) and determination coefficient (R^2). These evaluation indexes are related to the experimental values (x^{exp}), predicted values (x^{pre}) and the number of data points (N). From Table 3, it can be observed that both *RMSD* and *AAE* present lower values in the ISO-DNN model for the investigated compounds, and the R^2 for the ISO-DNN model is closer to 1. The results demonstrate that the developed QSPR model has a better agreement between the experimental and predicted log K_{OW} values for organic compounds.

Discriminative power in isomers

In chemistry, isomers are these molecules with identical formulae but distinct structures and they do not necessarily have similar properties. The isomers include structural isomers whose atoms and functional groups are joined together in different ways and stereoisomers that differ in three-dimensional orientations of their atoms and functional groups in space.

Although the KOWWIN program exhibits the strong ability in the log K_{OW} prediction, it is found that, under most circumstances, structural isomers can be differentiated but stereoisomers cannot. Because the KOWWIN program was developed relying on an atom and fragment contribution method, its discriminative power was limited in isomers. In contrast, the developed DNN-based QSPR model is able to differentiate the structural isomers and stereoisomers and account for stereochemistry due to the interaction between the canonical molecular signatures and Tree-LSTM networks in the processing and transmitting structural information for molecules.

Two pairs of structural isomers and two pairs of stereoisomers were extracted from the investigated dataset and they were taken as examples for illustrating the capabilities of two models in discriminating isomers. The experimental values, KOWWIN predicted values and ISO-DNN predicted values of these structural isomers and stereoisomers are summarized in Tables 4 and 5 respectively.

Regarding the isomers as shown in Tables 4 and 5, both compounds in each pair of isomers have different experimental values, while they were given the same KOWWIN predicted value since the predictive model does not have obvious advantages over differentiating isomers. In contrast, the ISO-DNN model has greater discriminative power and the predicted values were assigned depending on the distinct structures of structural isomers and stereoisomers. Meanwhile, the predictive accuracy was guaranteed. As it turns out, the strong discriminative power of the developed ISO-DNN model can be attributed to the interaction between the canonical molecular signatures and Tree-LSTM networks. Furthermore, considering that only part of compounds was featured with the isomeric information, the impact of isomeric features on the predictive accuracy of the model is discussed in the ESI (Pages S9 - S11).

Conclusions

In this work, a QSPR model under the deep learning approach was developed for accurately and reliably predicting the octanol-water partition coefficients for organic compounds and providing valuable environmental information for guiding selection and development of important chemicals including green solvents. Prior to the learning process, molecular features were extracted to describe structural information and connectivity. The canonical molecular signatures were produced relying on the theory of canonizing molecular graph and mapped on the Tree-LSTM networks to generate input parameters in preparation for obtaining a QSPR model. Afterwards, the learning process was performed by the built DNN model combining the

Tree-LSTM network and the BPNN, and the final QSPR model for the log K_{OW} estimation was determined after the massive training and test. The evaluations were finally carried out for exhibiting better predictive accuracy and external competitiveness of the DNN-based QSPR model in contrast with an authoritative log K_{OW} prediction tool. Moreover, the developed QSPR model revealed greater discriminative power in the structural isomers and stereoisomers.

Differing from the traditional property prediction models, the developed deep-learning-assisted model avoids the human intervention in the feature selection of molecular structures. Meanwhile, by coupling the canonical molecular signatures and Tree-LSTM networks, the molecular features were automatically extracted from chemical structures which circumvents numerous topological features and physicochemical descriptors. Therefore, the deep learning approach was successfully implemented to develop a QSPR model for predicting the log K_{OW} for the organic compounds which provide valuable information in the absorption, distribution and metabolism of chemicals and could be further employed to predict various indicators of toxicity. It proved that the deep learning approach can serve as a promising and intelligent approach to develop property prediction models with high predictive accuracy and a wide application scope. Although this research focused on predicting the log K_{OW} for measuring the lipophilicity of organic chemicals, the proposed approach can be further popularized to some other environmentally important properties such as water solubility and bioconcentration factor, which exhibits its vital potentials in the development of green chemistry.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The authors acknowledge the financial support provided by the National Natural Science Foundation of China (Nos. 21606026, 21878028); the Fundamental Research Funds for the Central Universities (Nos. 2019CDQYHG021, 2019CDXYHG0013); the Beijing Hundreds of Leading Talents Training Project of Science and Technology (No. Z171100001117154).

Notes and references

1 J.H. Clark, *Green Chem.*, 2006, **8**, 17-21.

2 (*a*) D. Prat, A. Wells, J. Hayler, H. Sneddon, C. R. McElroy, S. Abou-Shehada and P. J. Dunne, *Green Chem.*, 2015, **18**, 288-296; (*b*) S. Jin, A. J. Hunt, J. H. Clark and C. R. McElroy, *Green Chem.*, 2016, **18**, 5839-5844; (*c*) S. Jin, Y. Tian, C. R. McElroy, D. Wang, J. H. Clark and A. J. Hunt, *Catal. Sci. Technol.*, 2017, **7**, 4859-4865; (*d*) S. Jin, F. Byrne, C. R. McElroy, J. Sherwood, J. H. Clark and A. J. Hunt, *Faraday Discuss.*, 2017, **202**, 157-173; (*e*) R. Luque, J. A. Menendez, A. Arenillas and J. Cot, *Energy Environ. Sci.*, 2012, **5**, 5481-5488; (*f*) C. S. K. Lin, L. A. Pfaltzgraff, L. Herrero-Davila, E. B. Mubofu, S. Abderrahim, J. H. Clark, A. K. Apostolis, K. Nikolaos, S. Katerina, D. Fiona, T. Samarthia, M. Zahouily, B. Robert and L. Rafael, *Energy Environ. Sci.*, 2013, **6**, 426-464.

3 J.H. Clark, *Green Chem.*, 1999, **1**, 1-8.

4 (*a*) W. Shen, L. Dong, S. Wei, J. Li, H. Benyounes, X. You and V. Gerbaud, *AIChE J.*, 2015, **61**, 3898-3910; (*b*) A. Jayswal, X. Li, A. Zanwar, H. H. Lou and Y. Huang, *Comput. Chem. Eng.*, 2011, **35**, 2786-2798.

5 Y. Hu, Y. Su, S. Jin, I. L. Chien and W. Shen, *Sep. Purif. Technol.*, 2019, **211**, 723-737.

6 A. Yang, H. Zou, I. Chien, D. Wang, S. Wei, J. Ren and W. Shen, *Ind. Eng. Chem. Res.*, 2019, **58**, 7265-7283
7 S. Neidle, *Cancer Drug Design and Discovery*, Elsevier, New York, 2011, pp. 131-154.
8 A. Rybinska, A. Sosnowska, M. Grzonkowska, M. Barycki and T. Puzyn, *J. Hazard. Mater.*, 2016, **303**, 137-144.
9 J. Marrero and R. Gani, *Ind. Eng. Chem. Res.*, 2002, **41**, 6623-6633.
10 T. Cheng, Y. Zhao, X. Li, F. Lin, Y. Xu, X. Zhang, Y. Li and R. Wang, *J. Chem. Inf. Model.*, 2007, **47**, 2140-2148.
11 M. Turchi, Q. Cai and G. Lian, *Chem. Eng. Sci.*, 2019, **197**, 150-158.
12 (*a*) F. P. Byrne, S. Jin, G. Paggiola, T. H. M. Petchey, J. H. Clark, T. J. Farmer, A. J. Hunt, C. R. McElroy and J. Sherwood, *Sustainable Chem. Processes.*, 2016, **4**, 7; (*b*) F. P. Byrne, B. Forier, G. Bossaert, C. Hoebers, T. J. Farmer and A. J. Hunt, *Green Chem.*, 2018, **20**, 4003-4011.
13 M. Tobiszewski, S. Tsakovski, V. Simeonov, J. Namieśnik and F. Pena-Pereira, *Green Chem.*, 2015, **17**, 4773-4785.
14 M. D. Ertürk and M. T. Saçan, *Ecotoxicol. Environ. Saf.*, 2013, **90**, 61-68.
15 S. Bakire, X. Yang, G. Ma, X. Wei, H. Yu, J. Chen and H. Lin, *Chemosphere*, 2018, **190**, 463-470.
16 S. G. Machatha and S. H. Yalkowsky, *Int. J. Pharm.*, 2005, **294**, 185-192.
17 M. Reinhard and A. Drefahl, Handbook for Estimating Physicochemical Properties of Organic Compounds, Wiley, New York, 1999.
18 C. Nieto-Draghi, G. Fayet, B. Creton, X. Rozanska, P. Rotureau, J. C. Hemptinne, P. Ungerer, B. Rousseau and C. Adamo, *Chem. Rev.*, 2015, **115**, 13093-13164.
19 J. C. Dearden, P. Rotureau and G. Fayet, *SAR QSAR Environ. Res.*, 2013, **24**, 279-318.
20 R. Mannhold, G. I. Poda, C. Ostermann and I. V. Tetko, *J. Pharm. Sci.*, 2009, **98**, 861-893.
21 R. Mannhold and H. van de Waterbeemd, *J. Comput.-Aided Mol. Des.*, 2001, **15**, 337-354.
22 (a) C. W. Cho, S. Stolte and Y. S. Yun, *Sci. Total Environ.*, 2018, **633**, 920-928. (b) B. Admire, B. Lian and S. H. Yalkowsky, *Chemosphere*, 2015, **119**, 1441-1446; (b) C. C. Bannan, G. Calabró, D. Y. Kyu and D. L. Mobley, *J. Chem. Theory Comput.*, 2016, **12**, 4015-4024.
23 (*a*) K. B. Hanson, D. J. Hoff, T. J. Lahren, D. R. Mount, A. J. Squillace and L. P. Burkhardb, *Chemosphere*, 2019, **218**, 616-623; (*b*) E. Wyrzykowska, A. Rybińska-Fryca, A. Sosnowska and T. Puzyn, *Green Chem.*, 2019, **21**, 1965-1973.
24 G. M. Kontogeorgis and R. Gani, *Computer Aided Property Estimation for Process and Product Design*, Elsevier, Amsterdam, 2004. pp. 3-26.
25 (*a*) W. M. Meylan and P. H. Howard, *J. Pharm. Sci.*, 1995, **84**, 83-92; (*b*) A. J. Leo, *Chem. Rev.*, 1993, **93**, 1281-1306.
26 K. G. Joback and R. C. Reid, *Chem. Eng. Commun.*, 1987, **57**, 233-243.
27 L. Constantinou and R. Gani, *AIChE J.*, 1994, **40**, 1697-1710.
28 J. Marrero and R. Gani, *Fluid Phase Equilib.*, 2001, **183**, 183-208.
29 S. Jhamb, X. Liang, R. Gani and A. S. Hukkerikar, *Chem. Eng. Sci.*, 2018, **175**, 148-161.
30 I. V. Tetko, V. Y. Tanchuk and A. E. P. Villa, *J. Chem. Inf. Model.*, 2001, **41**, 1407-1421.
31 V. K. Gombar and K. Enslein, *J. Chem. Inf. Model.*, 1996, **36**, 1127-1134.
32 J. J. Huuskonen, D. J. Livingstone and I. V. Tetko, *J. Chem. Inf. Model.*, 2000, **40**, 947-955.
33 J. I. García, H. García-Marín, J. A. Mayoral and P. Pérez, *Green Chem.*, 2013, **15**, 2283-2293.
34 N. D. Austin, N. V. Sahinidis and D. W. Trahan, *Chem. Eng. Res. Des.*, 2016, **116**, 2-26.
35 D. P. Visco, R. S. Pophale, M. D. Rintoul and J. L. Faulon, *J. Mol. Graphics Modell.*, 2002, **20**, 429-438.

36 J. L. Faulon, D. P. Visco and R. S. Pophale, *J. Chem. Inf. Model.*, 2003, **43**, 707-720.
37 N. G. Chemmangattuvalappil and M. R. Eden, *Ind. Eng. Chem. Res.*, 2013, **52**, 7090-7103.
38 N. G. Chemmangattuvalappil, C. C. Solvason, S. Bommareddy and M. R. Eden, *Comput. Chem. Eng.*, 2010, **34**, 2062-2071.
39 U. Safder, K. J. Nam, D. Kim, M. Shahlaei and C. K. Yoo, *Ecotoxicol. Environ. Saf.*, 2018, **162**, 17-28.
40 A. Eslamimanesh, F. Gharagheizi, A. H. Mohammadi and D. Richona, *Chem. Eng. Sci.*, 2011, **66**, 3039-3044.
41 A. Lusci, G. Pollastri and P. Baldi, *J. Chem. Inf. Model.*, 2013, **53**, 1563-1575.
42 F. Gharagheizi, A. Eslamimanesh, F. Farjood, A. H. Mohammadi and D. Richon, *Ind. Eng. Chem. Res.*, 2011, **50**, 11382-11395.
43 J. Zheng, Y. Zhu, M. Zhu, G. Sun and R. Sun, *Green Chem.*, 2018, **20**, 3287-3301.
44 S. Hochreiter and J. Schmidhuber, *Neural Computation*, 1997, **9**, 1735-1780.
45 K. S. Tai, R. Socher and C. D. Manning, *arXiv preprint*, 2015, arXiv:1503.00075.
46 Y. Su, Z. Wang, S. Jin, W. Shen, J. Ren and M. R. Eden, *AIChE J.*, 2019, doi: 10.1002/aic.16678.
47 J. L. Faulon, M. J. Collins and R. D. Carr, *J. Chem. Inf. Model.*, 2004, **44**, 427-436.
48 KOWWIN Data, http://esc.syrres.com/interkow/KowwinData.htm, (accessed November 28, 2018).
49 L. Li, Z. Wen and Z. Wang, *Theory, Methodology, Tools and Applications for Modeling and Simulation of Complex Systems*, Springer, Singapore, 2016, pp. 497-503.
50 D. Weininger, *J. Chem. Inf. Model.*, 1988, **28**, 31-36.
51 S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E. E. Bolton, *Nucleic Acids Res.*, 2019, **47**, D1102-1109.
52 Daylight Chemical Information Systems, Inc., http://www.daylight.com/, (accessed December 5, 2018).
53 RDKit: Open-Source Cheminformatics Software, http://www.rdkit.org/, (accessed December 5, 2018).
54 D. E. Rumelhart, G. E. Hinton and R. J. Williams, *Cognitive Modeling*, 1988, **5**, 1.
55 M. A. Nielsen, *Neural Networks and Deep Learning*, Determination press, California, 2015.
56 Y. LeCun, Y. Bengio and G. Hinton, *Nature*, 2015, **521**, 436-444.
57 PyTorch, https://pytorch.org/, (accessed December 5, 2018).
58 P. J. Huber, *Ann. Math. Stat.*, 1964, **35**, 73-101.
59 D. P. Kingma and J. Ba, *arXiv preprint*, 2014, arXiv:1412.6980.
60 N. J. Nilsson, *Artificial Intelligence: A New Synthesis*, Morgan Kaufmann, California, 1998.
61 F. Melnikov, J. Kostal, A. Voutchkova-Kostal, J. B. Zimmerman and P. T. Anastas, *Green Chem.*, 2016, **18**, 4432-4445.
62 N. Chirico and P. Gramatica, *J. Chem. Inf. Model.*, 2012, **52**, 2044-2058.
63 A. Tropsha, *Mol. Inform.*, 2010, **29**, 476-488.
64 P. Gramatica, *Mol. Inform.*, 2014, **33**, 311-314.
65 A. Rybinska, A. Sosnowska, M. Grzonkowska, M. Barycki and T. Puzyn, *J. Hazard. Mater.*, 2016, **303**, 137-144.
66 P. P. Roy, S. Kovarich and P. Gramatica, *J. Comput. Chem.*, 2011, **32**, 2386-2396.
67 EPI Suite™-Estimation Program Interface, https://www.epa.gov/tsca-screening-tools/epi-suitetm-estimation-program-interface, (accessed February 3, 2019).