A novel unambiguous strategy of molecular feature extraction in machine learning assisted predictive models for environmental properties

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

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

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

[c] *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*

[d] *Green Chemistry Centre of Excellence, University of York, York YO105DD, UK*

Introduction

Environmental properties of compounds play a crucial role in many fields such as sustainable chemistry,[1-3] process design,[4,5] environmental remediation and evaluation of chemicals' environmental behaviours.[6-8] Environmental benefits drive the development of green solvents, chemical synthesis and molecular design toward eco-friendly technology,[9-11] because environmental properties provide valuable information on the absorption, distribution and metabolism of compounds and direct the treatment

of organic pollutants which may pose serious threats to humans and wildlife.[12] However, reliably measuring the environmental properties for compounds is a costly task and sometimes tedious, especially for those compounds with very low vapour pressure, low aqueous solubility or high risk. Therefore, different approaches have been proposed in the open literature to predict properties for various types of chemical compounds.

Empirical relationship method is one of the popular approaches for property estimation, in which different physicochemical properties (*e.g.*, critical temperature, vapour pressure, and aqueous solubility) serve as input parameters to calculate target properties of compounds.[7,13,14] For instance, Gharagheizi *et al*.[14] developed a fairly accurate empirical model to predict Henry's law constant values of organic compounds relying on several basic properties (*e.g.*, normal boiling point temperature and critical pressure). This model can be easily applied for rapid estimation and it exhibits an absolute average deviation of about 10% with respect to 1816 organic compounds. However, empirical relationship approaches heavily depend on the availability and accuracy of the required input properties. Thus, it is not practical to use if one of the inputs is unavailable (or cannot be estimated).

Another popular type of the predictive tools has focused on the application of quantitative structure-property relationship (QSPR) models, in which the physicochemical properties are supposed to be related to molecular structures. A number of studies have made great contributions in this regard.[15-21] In addition, several QSPR models were put forward based on group contribution (GC) methods.[12,22-24] In such models, molecules of interest are divided into various groups (*e.g.*, atoms and

substructures containing atoms and chemical bonds), and each group is assigned a specific contribution value. Afterwards, the target property of a compound can be given by summarizing the contributions of groups. The GC methods therefore are regarded as multiple linear mathematical models. Whereas, the same groups in different GC methods have distinct contribution values and the definitions of groups are not entirely the same. Thus, different GC methods work in a similar way though exhibit different results. A classic GC method is the three-level GC estimation approach proposed by Marrero and Gani,[25] in which a total of 370 kinds of groups were defined for recognizing molecular structures. Attributed to its superior performance, the three-level GC method has been extensively applied for estimating various physicochemical properties such as critical properties, standard enthalpy of vaporization, and the octanol-water partition coefficient.[26-28]

GC methods are characterized by simple models, quick and fairly reliable estimations. Based on a comprehensive literature review, three typical shortcomings of the traditional GC methods have often emerged in applications:

(a) Difficulties in understanding the definitions and structures of complex groups;

(b) Computational error and time consumption due to complexities in recognizing groups and calculating the property;

(c) Scattered predicted values for certain groups of compounds resulting from different feasible strategies appearing in the structure recognition.

With the rapid development of artificial intelligence and computational power, many QSPR models have been investigated with the aid of artificial neural networks (ANNs)

and have gained popularity for estimating physicochemical properties.[29-36] Because of their ability to model and reproduce nonlinear processes, ANN-GC hybrid models have presented accurate predictive tools mainly with the aim of alleviation of the problems in the traditional GC methods.[37-39] As such, the assistance of computer-aided technologies enables the ANN-GC models to readily correct these shortcomings (a) and (b). Meanwhile, the shortcoming (c) can also be removed by predefining the priority rules of available strategies or developing new GC methods. However, the priority rules or the defined groups need to be updated when new chemicals and chemical structures are introduced.

In this research, an unambiguous strategy is proposed to rapidly recognize molecular structures, extract molecular features, and transfer features into identifiers according to encoding rules. The feature extraction algorithm for accomplishing these works is developed and is introduced in detail. Moreover, using the proposed strategy, a QSPR model is developed to predict property values for organic compounds in water, based on their experimental data and molecular structures. For this, adopting machine learning algorithms, a simple four-layer ANN is constructed to generate a predictive model which is expected to exhibit the following features:

(a) Using fewer molecular features to achieve more accurate predictions compared to the available models in the literature;

(b) Avoiding various feasible strategies appearing in structure recognition to prevent the scattered predicted values for certain groups of compounds;

(c) Enhancing the generality of the model with respect to the types of the organic

compounds.

Methodology

Herein, a strategy is proposed to rapidly recognize molecular structures and to extract molecular features without ambiguity followed by a neural network specially built for producing a predictive model to estimate physicochemical properties of the compounds of interest. Henry's law constant (HLC) for compounds in water is employed as a case study in this research. It is the air-water partition coefficient which describes the equilibrium distribution of a chemical between air and water, and it can be expressed as the ratio of partial pressure above water to the amount of dissolved gas in water.40,41 The HLC is an indicator of the chemical's volatility. It is important in describing the distribution and transport of chemicals between aquatic ecosystems and the atmosphere, which determines the fate of chemicals in environment. Compounds displaying higher HLC values, especially the lower molecular weight compounds, are more likely to volatilize from aqueous solutions, they must be handled carefully to improve air quality and avoid short- and long-term adverse health effects. Fig. 1 illustrates the procedure of model development as follows:

Step 1: Data collection.

The experimental data42 of organic compounds is essential for the development of a QSPR model. In addition, simplified molecular-input line-entry system (SMILES) string is also treated as a key parameter to the presented model, which expresses fundamental information of molecular structures.

Step 2: Feature extraction.

To ensure that the information of molecular structures can be processed by the neural network, molecular features are extracted with the proposed strategy and later converted to numeric vectors which are generated in a unique manner relying on built-in encoding rules. In this way, the molecular information can be introduced to the neural network and be correlated to the value of the target property.

Step 3: Neural network design.

On the basis of the experimental data and molecular feature vectors, a fully connected neural network is constructed to develop the predictive model. The structural parameters required in the design of neural network are optimized using cross-validation and grid search in order to provide stability and reliability in model training.

Step 4: Model training.

Having received the feature vectors describing molecular structures, the neural network establishes a complex mathematical model and then produces the estimated property values. The training process runs repeatedly aiming to obtain a better predictive model which could provide more accurate predictions for HLC values of organic compounds in water.

All the above steps are achieved with a series of programs written in Python. The program has been run successfully on a desktop computer with Intel Core i3-8100 processor under Windows 10 operating system.

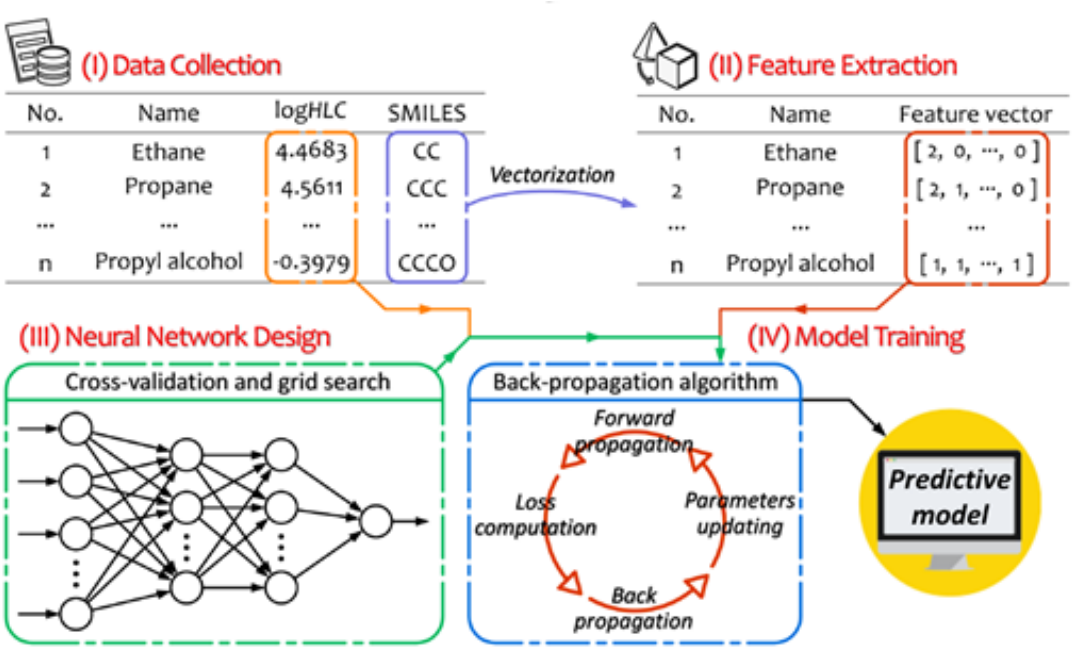

Fig. 1 The procedure for developing a predictive model to predict the logHLC values of organic compounds.

Data collection

To ensure the reliability of the predictive model, the experimental HLC values at 298.15 K are gathered from one of the most reliable and comprehensive databases.[42] The HLC is commonly reported in units of atm·m^3/mol (mole fraction basis) but here it is represented as its decimal logarithmic form (log*HLC*) because it spans over many orders of magnitude with regard to the collected massive samples. In this research, a number of irrelevant compounds (*e.g.*, inorganic compounds and ionic compounds) and the compounds provided with estimated HLC values have been discarded. Therefore, the model is applicable only to organic compounds and its reliability is significantly improved. As a consequence, the HLC values of 2566 diverse organic compounds in water are kept and assembled as the dataset for developing the predictive model. The compounds span 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. The distribution of the treated log*HLC* values is displayed in Fig. 2.

The other input for the development of the QSPR model is the information of molecular structures. The SMILES is a specification in the form of a line notation for describing the structure of chemical species, and it can be used to build two- or three-dimensional structure of a molecule.[43,44] As a chemical language, the SMILES string is sufficient to provide structural information for molecules required in model development. Thus, SMILES strings have been widely employed in the literature for developing QSPR-based models and cheminformatics software. Additionally, having learned the simple encoding rules of the SMILES strings, one can readily and correctly give the SMILES string of a compound from its molecular structure.

PubChem[45] is a massive open repository which provides over 200 million kinds of compounds with chemical information such as molecular formula, SMILES string, and so forth. It should be noted that there are two types of SMILES strings, canonical SMILES and isomeric SMILES. The former one is available for all the existing compounds, whereas the latter one is only provided for isomers since the isomeric SMILES strings contain isomeric information of molecules.

The SMILES strings for these investigated compounds have been collected from the PubChem database. In order to preserve the isomeric information, the isomeric SMILES string has been adopted if it is available for a given compound; otherwise the canonical SMILES string is employed. Therefore, the experimental data and SMILES

strings of the investigated 2566 organic compounds have been prepared for the correlation of molecular structures and properties.

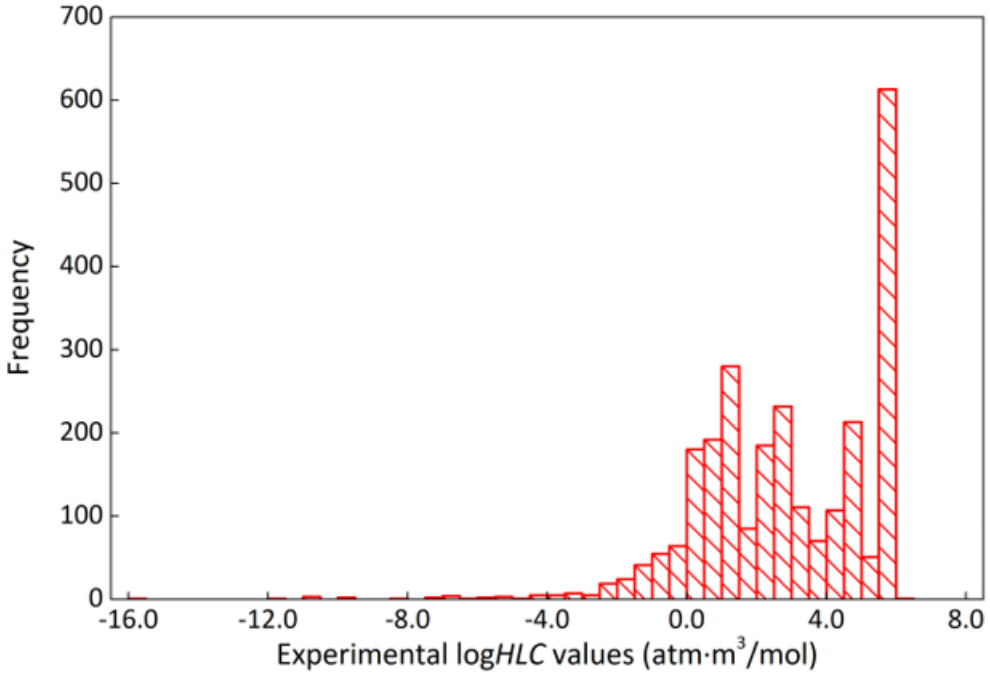

Fig. 1 The distribution of the collected experimental log*HLC* values for 2566 organic compounds.

Feature extraction

To be provided to the neural network, all types of data need to be translated into the numeric form contained in vectors. Accordingly, the molecular information of each compound needs to be converted and included in a numeric vector. For this purpose, an unambiguous strategy is proposed and programed to rapidly recognize molecular structures and extract molecular features. In the proposed strategy, each molecular feature represents a molecular substructure that only contains single non-hydrogen atom accompanied with its connected hydrogen atoms and chemical bonds. Therefore, only one strategy is feasible in subdividing a molecule into several substructures, and it avoids scattered predicted values. These features are created with built-in encoding

rules in which various traditional chemical information (such as type of the non-hydrogen atom, number of hydrogen atoms, and formal charge[46]) of substructures is taken into consideration. In addition, the types of chemical bonds between the substructure and its connected substructures in the molecule are considered in the encoding rules for creating molecular features, and meanwhile, stereoisomers are also identified by the encoding rules and the stereo-centres are recorded in the molecular features. In this way, molecular features are extracted with the encoding rules and similar substructures can be distinguished to the greatest extent. On this basis, molecular structures have been converted to numeric vectors according to the frequency of each feature. Therefore, similar to GC-based methods, the proposed strategy is characterized by good interpretability as molecule are the combination of fragments.

In order to preserve molecular information and specify distinct molecular features, the RDKit cheminformatics tool has been adopted for implementing the encoding rules to present the features with identifiers. The definitions of the characters incorporated in identifiers are provided in Table S1 of Electronic Supplementary Information (ESI). The procedure for the feature extraction and vectorization of molecular structures is comprised of the following three steps as depicted in Fig. 3.

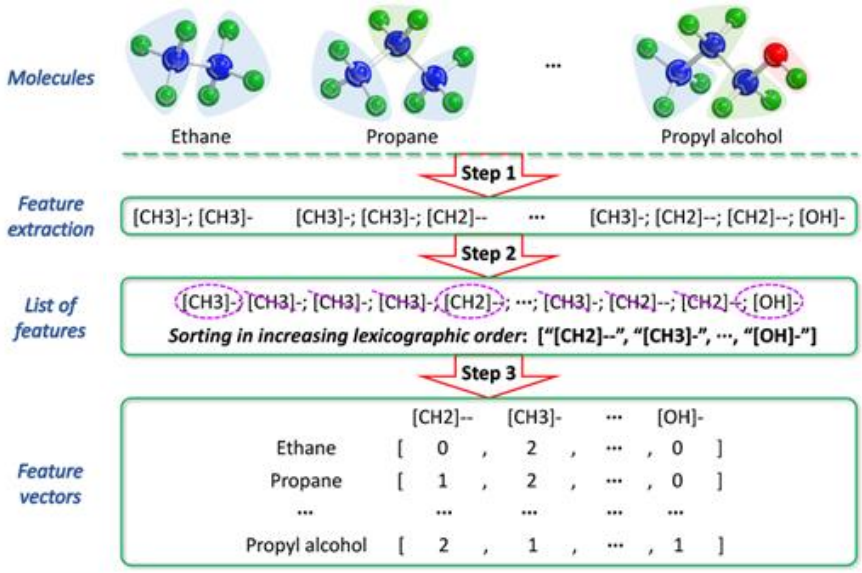

Fig. 2 The procedure of the feature extraction and vectorization for molecular structures.

Step 1: The molecular features are extracted from the molecular structures of organic compounds of interest which have been already expressed with identifiers using the pre-defined encoding rules. The process covers all the atoms and chemical bonds in a molecule to acquire the information of molecular structures without omissions.

Step 2: The molecular features represented with identifiers are assembled into a list and the duplicates are removed to ensure that each feature only appears once in the list. Then, all the remaining molecular features in the list are sorted in increasing lexicographic order (according to the Python function of "sorted") to fix the location of every feature in the list.

Step 3: For any individual compound, the feature extraction is performed again following step 1. Afterwards, the frequency of each feature in the molecule is assigned to the numeric vector according to its corresponding location in the feature list. Therefore, the final vectors include the required molecular information for all of the

compounds presented in the database.

In this way, molecular features are extracted and molecular vectors are generated. Attributed to the chemical information incorporated in molecular features, the proposed strategy is able to differentiate isomers, and whereas, part of structural isomers cannot be distinguished. Therefore, plane of best fit (PBF),[47] a rapid and amenable method for describing the 3D character of molecules is employed to retain the molecular information omitted in the proposed feature extraction strategy. The proposed strategy is improved with the introduction of PBF, and both structural and geometric isomers are well identified.

Neural network design

The input parameters to the developed predictive model are transferred from the first layer of neural network (the input layer) to its last layer (the output layer) through specific mathematical relations (neurons) and, accordingly, results in the predicted HLC values.

Layer is the basis to determine the architecture of a neural network. In this research, the neural network has been built with four layers including one input layer, two hidden layers and one output layer. The number of neurons in the input layer matches the number of numeric values in the input vector so that all extracted molecular features are completely loaded. In addition, the output layer only contains one neuron for producing predicted values for the target property. The network is fully connected which means that each neuron in a layer is connected to all neurons in the previous layer (see Fig. S1 of ESI).

The four-layer neural network has been developed using Python as follows:

(i) PyTorch is an open-source machine learning library for Python which is rising in popularity, and it is used to build different structures of the neural network in a flexible way;[48]

(ii) Root mean square error measures the differences between predicted and experimental values, and it is adopted as the loss function to quantify the performance of the developed model;

(iii) Adam algorithm[49] is an optimization method to update the weights and biases of the neural network, and it is applied to optimize the predictive model because of its high computational efficiency;

(iv) Back-propagation algorithm, a supervised learning procedure commonly used to train neural networks, is employed to update the weights and biases of neurons by calculating the gradient of the loss function.

The parameters of the neural networks are generally divided into two categories: model parameters and hyper-parameters. Model parameters (*e.g.*, weights and biases) are automatically tuned or optimized by calculating the gradient of the loss function during training. On the other hand, model hyper-parameters are commonly set by the operators in advance before the neural network is functional. With the aim of efficiently controlling the training process and generating a robust model, the hyper-parameters are herein optimized using the approaches of cross-validation and grid search. Learning rate is set to 1.00×10^{-3} to control the rate of convergence for the neural network. Activation functions map inputs to outputs and enhance the ability of neural networks

in handling complex tasks. Two types of activation functions, “sigmoid” and “softplus” (corresponding equations are provided in Table S2 of ESI), are introduced to hidden layer 1 and 2, respectively. Moreover, the application of the ANN-based predictive model is illustrated in the ESI (Page S9).

Model training

During iterative training of the neural network, one epoch represents one forward pass (regression process from input layer to output layer) and one backward pass (back-propagation process from output layer to input layer) for all the data of a dataset. As per the batch size of training set, an epoch is divided into several iterations. The weights and biases of neurons are updated after every iteration completed so that the model can be optimized multiple times during one epoch.

The predictability of the neural network is generally verified with an external dataset which is not involved in the training of neural network. Herein, the collected dataset (including the HLC values measured in water for 2566 compounds) has been divided into two subsets: a modelling set and a test set, holding 80% and 20% of the whole dataset by using a random selection routine or k-means clustering method (*i.e.*, random sampling and cluster sampling). Data points might be distributed very non-uniformly in the input space, and therefore, adopting the k-means clustering in the data partitioning would lead to better training, validation and test sets than simply using randomization. The modelling and test sets are employed to, respectively, build the predictive model and evaluate the predictability of the developed model. The best set of hyper-parameters are determined by the five-fold cross-validation. In the five-fold

cross-validation, the dataset is equally partitioned into five subsets and the model training is carried out five times. During each training process, one of the five subsets is regarded as the validation set and the remaining four subsets are assigned to the training set. Therefore, each subset is used for training four times and for validation once. After training five times, the model performance is finally evaluated with the results from five independent validation sets.

During training the neural network, the error in the validation set is compared with that in the training set. Usually where both learning curves meet the tolerance is the point at which training should stop. The error is measured with the adopted loss function, and the tolerance is set to 1.00×10^{-3}.

Results and discussion

Feature vector

58 types of molecular features have been extracted from the molecular structures relying on the proposed unambiguous strategy, and they are summarized in Table S3 of ESI. These features are represented by identifiers involving various chemical information. For instance, the molecular feature "[CH0]-#" indicates an aliphatic carbon atom attached with zero hydrogen atoms, a single bond, and a triple bond.

Afterwards, the extracted molecular features are sorted in increasing lexicographic order as mentioned earlier. On this basis, the frequency of each feature appeared in a molecule is computed. Integers represented the frequencies of features are assigned in the corresponding locations of a numeric vector. In this way, the numeric vector containing 58 nonnegative integers is generated to describe the structural information

of the molecule. Three small molecules (ethane, propane, and 1-propanol) are taken as examples to illustrate the production of vectors as shown in the Fig. 3. Once the numeric vectors have been prepared, they act as input parameters for the neural network to correlate the relationship between structures and properties.

Training process

The numeric vectors characterizing molecular information are introduced as input parameters to the neural network. The number of neurons in the input layer is equivalent to the number of numeric values in the feature vector, and thus, all the molecular information can be completely loaded to the neural network. During training of the model presented in this research, the loss function of training set has been minimized by the optimizer to search for a fairly accurate predictive model to describe the relationship between the molecular structures and the target property. Once a batch passes through the neural network, the molecular information traverses all the neurons from the first to last layer, and the neural network produces predicted values for this batch. Subsequently, the deviations between experimental and predicted values are calculated, and then the weights and biases of the neurons are updated from the output layer to the input layer with the back-propagation algorithm.

In order to improve the robustness of the neural network, the numbers of neurons in two hidden layers are optimized using the five-fold cross-validation and grid search method. Four models are investigated considering two different input vectors (*i.e.*, feature vector and feature vector supplemented with PBF) and two different sampling methods (random sampling and cluster sampling). As highlighted in Fig. 4, the optimal

set of structural parameters for each model is determined by the lowest loss function value, and for four discussed schemes, the numbers of neurons in hidden layers are 7 and 10, 13 and 12, 10 and 7, and 15 and 6, respectively. It is worth mentioning that the number of cluster centres in cluster sampling is optimized by calculating Calinski-Harabasz index,[50] and the results show that the clustering is better when four cluster centres are given (see Fig. S7 of ESI). Moreover, the model performance is directly compared with the number of cluster centres as discussed in the ESI (Pages S10-S12). The learning curves of these models with optimal sets of structural parameters are provided in Fig. S2 of ESI, which compare the errors in the training and validation sets for predictive models trained with different input vectors and dataset dividing methods.

Model performance

The problem of over-fitting in the neural network eventually leads to the loss of the model's predictability. In the traditional methods for property prediction, the whole dataset is employed to train and test the predictive models. Thus, these traditional prediction models may have weak predictability in predicting properties for new compounds of interest.

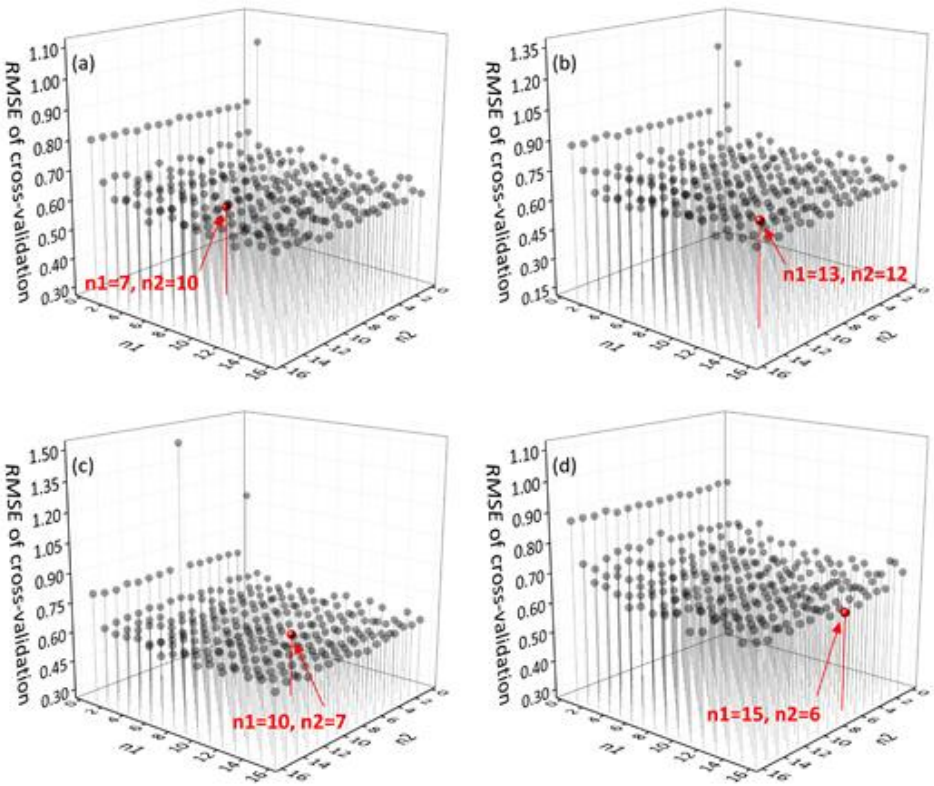

Fig. 3 The optimization and determination for the numbers of neurons in two hidden layers (n1 and n2) trained using (a) feature vector under random sampling (**Scheme 1**); (b) feature vector under cluster sampling (**Scheme 2**); (c) feature vector supplemented with PBF under random sampling (**Scheme 3**); (d) feature vector supplemented with PBF under cluster sampling (**Scheme 4**).

The four-layer fully connected neural network with given parameters is in fact a correlation between molecular structures and log*HLC* values. Based on the modelling and test sets, the predictability of the developed predictive model is measured by estimating log*HLC* values from the molecular structures. The numeric vectors of structure features of the independent compounds (not used in training and validation steps) are fed into the neural network and the estimated values are given based on the developed model.

The statistical analysis for the modelling and test sets is carried out with three indicators

based on the experiment value (x^{exp}), predicted value (x^{pre}) and number of data points (N). The first is the root mean squared error ($RMSE$) which measures the standard deviation of differences between estimated and experimental values. The second is mean absolute error (MAE) which indicates the magnitude of differences between estimated and experimental values. Another is coefficient of determination (R^2) which provides information about the quality of the model fit.

The statistical parameters for four predictive models in Table 1 reveal that these models have satisfactory predictability and can make accurate prediction on the new data. In comparison, the model trained using feature vector supplemented with PBF under cluster sampling (*i.e.*, **Scheme 4**) is significantly better than others, which indicates that introducing PBF descriptor as input and adopting k-means clustering in sampling lead to better predictive performance. The logHLC values calculated with **Scheme 4** versus the corresponding experimental data is illustrated in Fig. 5. Moreover, the weight and bias matrixes for this predictive model are provided in Table S4 of ESI.

Table 1 The statistical analysis for the subsets and whole dataset in log*HLC* prediction using different input vectors and sampling methods.

Predictive model	Dataset	N [a]	$RMSE$ [b]	MAE [c]	R^2 [d]
	Modelling set	2052	0.3197	0.1686	0.9824
Scheme 1	Test set	514	0.6469	0.2553	0.9453
	Whole dataset	***2566***	***0.4069***	***0.1860***	***0.9732***
	Modelling set	2052	0.2886	0.1579	0.9866
Scheme 2	Test set	514	0.5683	0.2558	0.9467
	Whole dataset	***2566***	***0.3623***	***0.1775***	***0.9787***
	Modelling set	2052	0.2875	0.1535	0.9858
Scheme 3	Test set	514	0.5619	0.2410	0.9587
	Whole dataset	***2566***	***0.3596***	***0.1710***	***0.9791***
	Modelling set	2052	0.2592	0.1399	0.9888
Scheme 4	Test set	514	0.4188	0.2121	0.9741
	Whole dataset	***2566***	***0.2981***	***0.1544***	***0.9856***

[a] Number of data points;

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

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

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

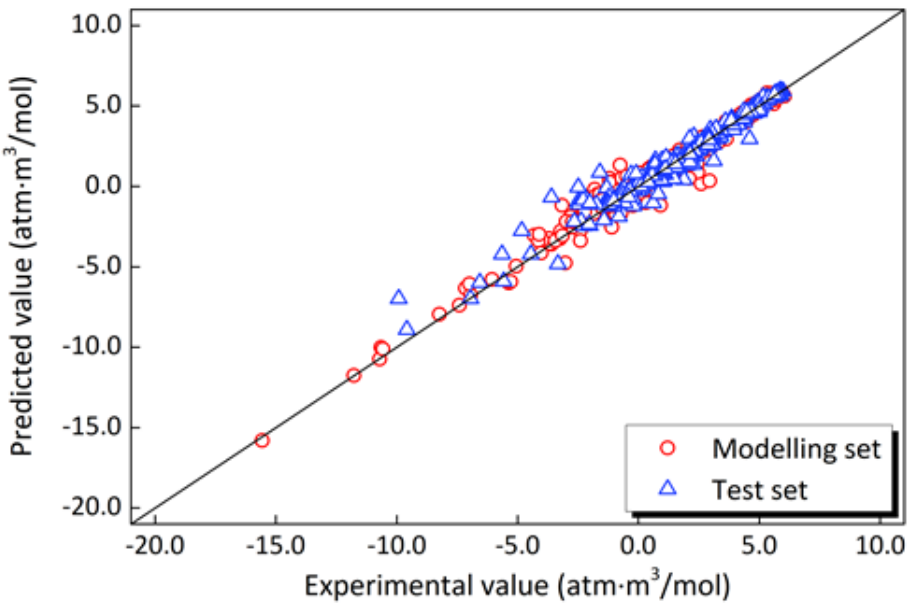

Fig. 4 The scatter plot of experimental and predicted log*HLC* values for the modelling and test sets.

Comparison with reported models

The ultimate objective of developing the predictive model is to accurately estimate log*HLC* values in water for organic compounds. Although a number of predictive models have been reported in the literature for this purpose, different models need be compared using the same experimental dataset.

Herein, the performance of the developed neural-network-assisted predictive model based on feature vector supplemented with PBF under cluster sampling (*i.e.*, **Scheme 4**) in this research (represented as NN model) is compared with a few of available models in the literature. An empirical relationship method[14] (represented as ER model) is picked in contrast with the NN model to measure the predictive power of the developed model. A comprehensive comparison shows that over 80% compounds (1475 out of 1816) used in the ER model are included in the development of the NN model, which proves that both models have employed similar datasets.

As far as the 1475 organic compounds are concerned, both models exhibit satisfactory predictive accuracy. As displayed in Fig. 6, it is clear that the NN model produced relatively small deviations. In other words, the NN model has a better agreement between the predicted and experimental values in terms of the overlapped 1475 organic compounds.

From the view of statistics, residual (experimental value minus estimated value) of each compound is calculated to compare the residual distribution plots of both ER and NN models (see Fig. S3 of ESI). With respect to the residual distribution, the residuals produced by the NN model are more densely gathered around the zero value which indicates that the proposed NN model perfectly estimated the log*HLC* values for more compounds than the conventional ER model.

On the other hand, several statistical indicators such as *RMSE*, *MAE*, and R^2 are analysed based on the same data subset as shown in Table 2. For the overlapped 1475 organic compounds, the *RMSE* and *MAE* of the NN model are significantly lower than those of the ER model which means that the NN model generated smaller errors in predicting log*HLC*. Meanwhile, the R^2 of the NN model is closer to 1.0000 which donates that the predicted values given by the NN model are better fitted with the experimental values. All these statistical results further confirmed the conclusion drew with the residual distribution, and it demonstrated the stronger predictive capability of the NN model.

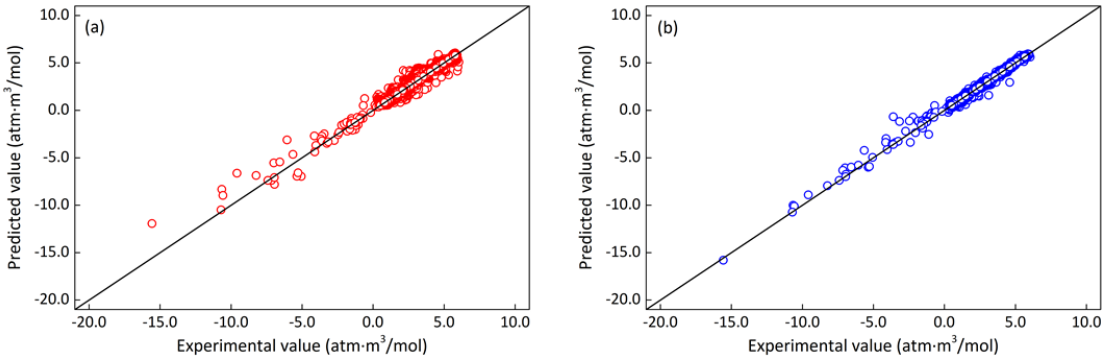

Fig. 5 The scatter plots of experimental and predicted log*HLC* values for (a) ER model and (b) NN model.

Except for the empirical relationship method, a hybrid method coupling the GC method and the neural network was proposed to develop a predictive model (represented as GN model) for estimating log*HLC* values of organic compounds.[37] In the GN model, 107 functional groups are extracted from 1940 compounds, and on this basis, a four-layer neural network was built to produce a nonlinear model for property estimation. In comparison with previous studies, it covered a lager dataset and showed a lower *RMSE* value.

Although different approaches for structure representation are used in the GN and NN models, the neural network is adopted as a tool to develop predictive models for estimating log*HLC*. Accordingly, it is necessary to evaluate the predictive performance of NN model in contrast with the GN model. A thorough comparison reveals that over 80% organic compounds (1567 out of 1940) employed in the GN model are used to develop the predictive model in this research. In terms of the overlapped 1567 compounds, the predictive capabilities of both GN and NN models are visualized in Fig. 7 with the scatter plots of estimated values versus experimental values.

Table 2 The comparison for statistical results of the ER and NN models in log*HLC* prediction.

Predictive model	*N*	*RMSE*	*MAE*	R^2
ER model (Gharagheizi *et al.*[14])	1475	0.4400	0.2898	0.9660
NN model (this research)	1475	0.2124	0.1069	0.9921

Table 3 The comparison for statistical results of the GN and NN models in log*HLC* prediction.

Predictive model	*N*	*RMSE*	*MAE*	R^2
GN model (Gharagheizi *et al.*[37])	1567	0.3283	0.1356	0.9822
NN model (this research)	1567	0.2187	0.1123	0.9921

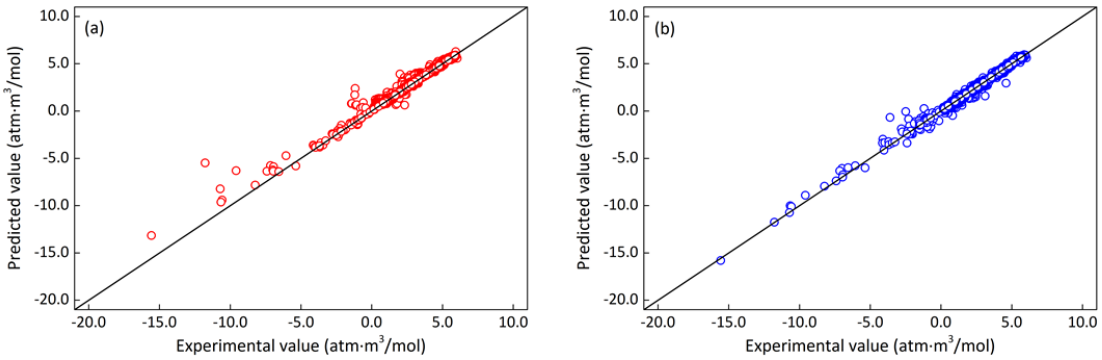

Fig. 7 The scatter plots of experimental and predicted logHLC values for (a) GN model and (b) NN model.

From the scatter plots, it is observed that both models exhibit satisfactory predictive accuracy although some data points represent relatively large deviations. In this regard, it is hard to conclude that which model is better in estimating log*HLC* for organic compounds. Thus, analysis is carried out in the statistical perspective to evaluate the

predictive performance of both models. From the residual distribution plots displayed in Fig. S4 of ESI, almost all the residuals produced by GN and NN models are within ±0.5 log units from the zero value. Accordingly, with respect to the overlapped 1567 compounds, both models made accurate estimation for log*HLC*. However, there are no obvious differences between the distributions of the residuals produced by the GN and NN models. Thus, their predictive performances are further quantified with several statistical indicators as shown in Table 3.

The consistency of these indicators for both models suggested that they have similar predictive performance in this task. Diving into this situation, NN model has a slightly lower *RMSE* and *MAE* together with a bit higher R^2. These subtle differences in statistical results prove that the NN method is slightly better than GN method in predictive accuracy for the (overlapped 1567) organic compounds.

From the above, the developed neural-network-assisted predictive model based on feature vector supplemented with PBF under cluster sampling (*i.e.*, the NN model) exhibits a distinct advantage over the empirical model (*i.e.*, the ER model) in the predictive accuracy and application scope. On the other hand, the NN model is slightly better than the GC-based neural network model (*i.e.*, the GN model) with the aid of the proposed feature extraction algorithm. Nevertheless, the GN model extracted 107 functional groups to develop the predictive model, whereas the NN model only adopted 58 molecular features. In other words, the NN model achieved a higher predictive accuracy with fewer molecular features. Moreover, the predictability of the developed model is further evaluated with compounds outside the adopted dataset as presented in

the ESI (Page S13).

Another point is worth mentioning that the proposed neural-network-assisted predictive model is developed relying on a large dataset of 2566 organic compounds with a R^2 of 0.9856. In the available predictive models, the largest dataset for model development contains 1954 pure compounds and the model exhibits an R^2 of 0.9828.[16] Therefore, the developed predictive model is considered to be the most comprehensive model for predicting log*HLC*. Accordingly, using the neural network and the proposed algorithm for extracting molecular features, the developed predictive model is able to provide accurate and reliable prediction for log*HLC* of organic compounds.

Conclusions

This research proposes an unambiguous feature extraction strategy to avoid different feasible strategies in the characterization of molecular structures. It therefore can overcome some shortcomings of GC-based methods, such as the scattered predicted values for certain groups of compounds. A four-layer neural network is then constructed to correlate the molecule structures with target property values for organic compounds. With the frequencies of molecular features as inputs, the neural network is trained with the acquired experimental data and evaluated with a test set which is not involved in the training process. During the training process, the numbers of neurons in the neural network are optimized to achieve a robust model using the five-fold cross-validation and grid search. As such, a hybrid predictive model is obtained with the combination of the proposed strategy and machine learning algorithm.

With respect to the logHLC values of pure organic compounds in water, the predictive

model is built based on the experimental values of 2566 organic compounds in water. Moreover, the introduction of the PBF descriptor and two dataset dividing methods are investigated in regard to the model performance. As it turns out, four predictive models are characterized by good predictability and predictive accuracy. The statistical analysis indicates that the predictive model developed with feature vector supplemented with PBF under cluster sampling shows significantly better predictive ability. It proves that the introduction of the PBF descriptor and adopting k-means clustering in sampling enhanced the model performance.

In contrast with the reported predictive models in the literature, the developed predictive model demonstrates higher predictive accuracy although fewer molecular features were used in its development. Moreover, it exhibits enhanced generality and covers more diverse organic compounds than reported models with respect to the employed comprehensive database. Therefore, the proposed strategy and model development methods can serve as a promising and effective approach to develop property predictive models, directing the reduction of pollutants in environment and the development of greener solvents. We can reasonably expect them to be further popularized to use for some other important environmental properties such as water solubility and the bioconcentration factor, which reveals their vital potential 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 (No. 21878028); the Fundamental Research Funds for the Central Universities (No. 2019CDQYHG021); the Chongqing Innovation Support Program for Returned Overseas Chinese Scholars (No. CX2018048); the Beijing Hundreds of Leading Talents Training Project of Science and Technology (No. Z171100001117154).

Notes and references

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

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

3 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.

4 N. G. Chemmangattuvalappil, C. C. Solvason, S. Bommareddy and M. R. Eden, *Comput. Chem. Eng.*, 2010, **34**, 582-591.

5 T. Zhou, K. McBride, S. Linke, Z. Song and K. Sundmache, *Curr. Opin. Chem. Eng.*, 2020, **27**, 35-44.

6 J. Sedlbauer, G. Bergin and V. Majer, *AIChE J.*, 2002, **48**, 2936-2959.

7 S. H. Hilal, S. N. Ayyampalayam and L. A. Carreira, *Environ. Sci. Technol.*, 2008, **42**, 9231-9236.

8 W. Shen, L. Dong, S. Wei, J. Li, H. Benyounes, X. You and V. Gerbaud, *AIChE J.*, 2015, **61**, 3898-3910.

9 D. Prat, A. Wells, J. Hayler, H. Sneddon, C. R. McElroy, S. Abou-Shehada and P. J. Dunne, *Green Chem.*, 2015, **18**, 288-296.

10 M. Tobiszewski, S. Tsakovski, V. Simeonov, J. Namieśnik and F. Pena-Pereira, *Green Chem.*, 2015, **17**, 4773-4785.

11 S. Jin, A. J. Hunt, J. H. Clark and C. R. McElroy, *Green Chem.*, 2016, **18**, 5839–5844.

12 N. K. Razdan, D. M. Koshy and J. M. Prausnitz, *Environ. Sci. Technol.*, 2017, **51**, 12466-12472.

13 F. Gharagheizi, A. Eslamimanesh, A. H. Mohammadi and D. Richon, *Ind. Eng. Chem. Res.*, 2011, **50**, 5877-5880.

14 F. Gharagheizi, A. Eslamimanesh, A. H. Mohammadi and D. Richon, *J. Chem. Thermodyn.*, 2012, **47**, 295-299.

15 T. Puzyn, P. Rostkowski, A. Świeczkowski, A. Jędrusiak and J. Falandysz, *Chemosphere*, 2006, **62**, 1817-1828.

16 F. Gharagheizi, P. Ilani-Kashkouli, S. A. Mirkhani, N. Farahani and A. H. Mohammadi, *Ind. Eng. Chem. Res.*, 2012, **51**, 4764-4767.

17 Y. Su, Z. Wang, S. Jin, W. Shen, J. Ren and M. R. Eden, *AIChE J.*, 2019, **65**, e16678.

18 Z. Wang, Y. Su, W. Shen, S. Jin, J. H. Clark, J. Ren and X. Zhang, *Green Chem.*, 2019, **21**, 4555-4565.

19 S. Datta, V. A. Dev and M. R. Eden, *Comput. Chem. Eng.*, 2019, **127**, 150-157.

20 M. Barycki, A. Sosnowska and T. Puzyn, *Green Chem.*, 2018, **20**, 3359-3370.

21 J. I. García, H. García-Marín, J. A. Mayoral and P. Pérez. *Green Chem.*, 2013, **15**, 2283-2293.

22 J. Sedlbauer, G. Bergin and V. Majer, *AIChE J.*, 2002, **48**, 2936-2959.

23 S. T. Lin and S. I. Sandler, *Chem. Eng. Sci.*, 2002, **57**, 2727-2733.

24 Y. Huang, H. Dong, X. Zhang, C. Li and S. Zhang, *AIChE J.*, 2013, **59**, 1348-1359.

25 J. Marrero and R. Gani, *Fluid Phase Equilib.*, 2001, **183**, 183-208.

26 J. Marrero and R. Gani., *Ind. Eng. Chem. Res.*, 2002, **41**, 6623-6633.

27 S. Jhamb, X. Liang, R. Gani and A. S. Hukkerikar, *Chem. Eng. Sci.*, 2018, **175**, 148-161.

28 T. Zhou, S. Jhamb, X. Liang, K. Sundmacher and R. Gani, *Chem. Eng. Sci.*, 2018, **183**, 95-105.

29 X. Yao, M. Liu, X. Zhang, Z. Hu and B. Fan, *Anal. Chim. Acta.*, 2002, **462**, 101-117.

30 A. Eslamimanesh, F. Gharagheizi, A. H. Mohammadi and D. Richon, *Chem. Eng. Sci.*, 2011, **66**, 3039-3044.

31 F. Gharagheizi, A. Eslamimanesh, A. H. Mohammadi and D. Richon, *Ind. Eng. Chem. Res.*, 2010, **50**, 221-226.

32 Y. Pan, J. Jiang and Z. Wang, *J. Hazard. Mater.*, 2007, **147**, 424-430.

33 M. Safamirzaei, H. Modarress and M. Mohsen-Nia, *Fluid Phase Equilib.*, 2008, **266**, 187-194.

34 N. J. English and D. G. Carroll, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 1150-1161.

35 M. Safamirzaei and H. Modarress, *Fluid Phase Equilib.*, 2012, **332**, 165-172.

36 D. R. O'Loughlin and N. J. English, *Chemosphere*, 2015, **127**, 1-9.

37 F. Gharagheizi, R. Abbasi and B. Tirandazi, *Ind. Eng. Chem. Res.*, 2010, **49**, 10149-10152.

38 F. Gharagheizi, A. Eslamimanesh, A. H. Mohammadi and D. Richon, *Ind. Eng. Chem. Res.*, 2011, **50**, 5815-5823.

39 F. Gharagheizi and R. Abbasi, *Ind. Eng. Chem. Res.*, 2010, **49**, 12685-12695.

40 D. Yaffe, Y. Cohen, G. Espinosa, A. Arenas and F. Giralt, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 85-112.

41 H. P. Chao, J. F. Lee and C. T. Chiou, *Water Res.*, 2017, **120**, 238-244.

42 C. L. Yaws, Yaws' Critical Property Data for Chemical Engineers and Chemists, https://app.knovel.com/hotlink/toc/id:kpYCPDCECD/yaws-critical-property/yaws-critical-property, (accessed 3rd April 2019).

43 Simplified molecular-input line-entry system, https://en.wikipedia.org/wiki/Simplified_molecular-input_line-entry_system, (accessed 9th May 2019).

44 D. Weininger, *J. Chem. Inf. Comput. Sci.*, 1988, **28**, 31-36.

45 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.

46 Formal Charge. https://en.wikipedia.org/wiki/Formal_charge, (accessed 6th March 2020).

47 N. C. Firth, N. Brown and J. Blagg, *J. Chem. Inf. Model.*, 2012, **52**, 2516-2525.

48 N. Ketkar, *Deep Learning with Python*, Apress, Berkeley, 2017.

49 D. P. Kingma and J. Ba, *arXiv preprint*, 2014, 1412.6980.

50 T. Caliński and J. Harabasz, *Commun. Stat.-Theory Methods*, 1974, **3**, 1-27.