Journal of Hazardous Materials 383 (2020) 121141

Contents lists available at ScienceDirect

Journal of Hazardous Materials

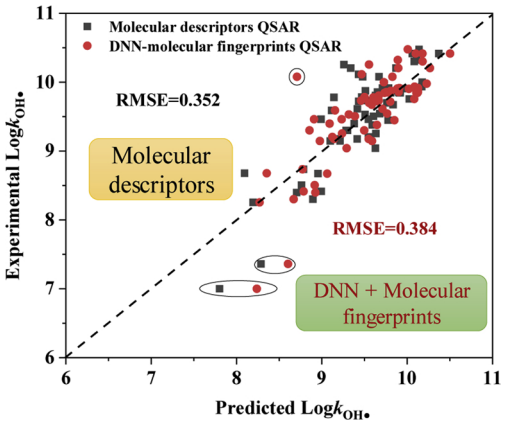
journal homepage: www.elsevier.com/locate/jhazmat

A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants

Shifa Zhong[a,1], Jiajie Hu[b,1], Xudong Fan[a], Xiong Yu[a,b], Huichun Zhang[a,*]

[a] Department of Civil Engineering, Case Western Reserve University, 2104 Adelbert Road, Cleveland, OH 44106-7201, USA
[b] Department of Electrical Engineering and Computer Science, Case Western Reserve University, 2104 Adelbert Road, Cleveland, OH 44106-7201, USA

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Xiaohong Guan

Keywords:
Deep neural network
Hydroxyl radical
Molecular fingerprints
QSAR
Advanced oxidation processes
Water treatment

ABSTRACT

This work combined a Deep Neural Network (DNN) with molecular fingerprints (MF) to develop models to predict the $OH^{\cdot}$ radical rate constants of 593 organic contaminants. Molecular descriptors, most often used in establishing quantitative structural-activity relationships (QSARs), were not used here because of their complicated generation processes that rely on advanced physicochemical and computational knowledge. Instead, we only fed the most basic information of the contaminant structures, i.e., MF encoding the types of atoms and how they are connected, to DNN and DNN then developed predictive models automatically. Here, a dataset containing 457 contaminants and their $OH^{\cdot}$ rate constants was first used to develop predictive models by DNN-MF. The hence developed models showed comparable accuracy to the traditional QSARs. The root mean square error (RMSE) values of the test sets were 0.358-0.384. The length of 2048 bits for the MF and 3 hidden layers (each with 1024 neurons) were found to be the optimal parameters for DNN. The model containing additional 89 micorpollutants in the training set was then successfully applied to predict the $OH^{\cdot}$ rate constants of 17 organophosphorus flame retardants and 29 additional micropollutants, with comparable accuracy to the reported molecular descriptors-based QSARs.

* Corresponding author.
E-mail address: hjz13@case.edu (H. Zhang).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.jhazmat.2019.121141
Received 31 July 2019; Received in revised form 29 August 2019; Accepted 2 September 2019
Available online 03 September 2019
0304-3894/

1. Introduction

Quantitative structure−activity relationships (QSARs) have been widely used for decades to correlate the reactivity with the chemical and/or structural features of organic compounds (Borhani et al., 2016; Free and Wilson, 1964; Su et al., 2018). Based on a given QSAR, the reactivity of new compounds can be predicted by applying their relevant features. This prediction is particularly significant in applications that need labor-intensive and expensive experiments, such as drug design (Hughes and Swamidass, 2017; Kubinyi, 1997; Olier et al., 2018). In the environmental field, QSARs have been often established to predict the reaction rate constants of organic contaminants with common reactants such as H_2O_2 (Lee and von Gunten, 2012; Su et al., 2018), O_3 (Lee and von Gunten, 2012; Sudhakaran and Amy, 2013), Fe(VI) (Ye et al., 2017), $OH^{\cdot}$ (Borhani et al., 2016; Cheng et al., 2017; Sudhakaran and Amy, 2013), $SO_4^{\cdot -}$ (Luo et al., 2018; Xiao et al., 2015), hydrated electrons (Li et al., 2018b), chlorine dioxide (Lee and von Gunten, 2012), MnO_2 (Salter-Blanc et al., 2016), and natural reductants such as Fe(II)-based and sulfite-based species (Canonica and Tratnyek, 2003; Colón et al., 2006; Salter-Blanc et al., 2015). However, the establishment of such QSARs is highly dependent on the calculation of a small group of pre-selected molecular descriptors, each derived to represent a portion of the molecular properties, such as Hammett constants, reduction potential, topological polar surface area, molar volume, dipole moment, and HOMO and LUMO energies. Different chemicals have different values of the molecular descriptors, which have often been obtained by calculations that require sophisticated physicochemical knowledge and the ability to use advanced software (Borhani et al., 2016; Cheng et al., 2017; Su et al., 2018; Ye et al., 2017). To establish working QSARs, one needs to intentionally choose the most relevant molecular descriptors from a large number of them – now over 5000 (Kamath and Pai, 2017), which again needs advanced knowledge of the descriptors and the types of reactions involved. Even with that knowledge, there is no guarantee that the selected descriptors are the most appropriate ones that can capture the entire picture of the reactivity, and the physical meanings of many theoretical descriptors are difficult to interpret (Borhani et al., 2016).

Deep neural network (DNN) is receiving increasing attention in recent years and has achieved great success in several areas (Coley et al., 2017; Moosavi et al., 2019; Ryan et al., 2018; Wei et al., 2016; Ye et al., 2018; Zhou et al., 2017), such as image recognition (Krizhevsky et al., 2012). One of the most exciting characteristics of DNN is that it can "learn" the features that are most relevant to the specific targets (e.g., reaction rate constants) by itself, and no sophisticated knowledge of the subject (e.g., reaction) is required. Hence, studies that used deep learning (e.g., artificial neural network) to develop QSARs have been recently published (Borhani et al., 2016; Ma et al., 2015). However, these studies still rely on complicated and sometimes arbitrary molecular descriptors as the inputs (Borhani et al., 2016; Fatemi, 2006; Ma et al., 2015). It is therefore of great interest to find out if QSARs can be developed without using any descriptor. Specifically, can we only provide the computer with the basic information about the compound structures, such as the types of atoms and how they are connected in the structure, and the corresponding reaction rate constants? If the computer can "learn" the relationship between the structures and the reactivity, then there is no need for any secondary molecular descriptors. The achievement of the above goal will greatly simplify and expand the development and applications of QSARs.

Molecular fingerprints (MF) encode structural and/or functional features of molecules as binary vectors (Glen et al., 2006), and have been commonly used in tasks such as virtual screening (Myint et al., 2012), similarity searching (Klopmand, 1992), and clustering (McGregor and Pallai, 1997). They have also been applied to developing QSAR models, such as predicting ligand biological activity (Myint et al., 2012) and toxicity (Mansouri et al., 2016; Wu and Wang, 2018). Each compound owns a unique vector (i.e., fingerprint) but the length of the vector is adjustable, for example, (0..1..0..1..0..0..1..1..) represents the MF of toluene (Figure S1 in the supplementary information (SI)). The values and positions in the vector store the structure information about the types of atoms and how they are connected to their neighboring atoms. Different compounds are described by different vectors, but the same structure feature in different compounds shares the same value and position within the vectors.

$OH^{\cdot}$ radical, a major reactive species in advanced oxidation processes that have been widely used in water treatment, can oxidize organic compounds through different structure-dependent pathways: (1) addition to olefin or aromatic systems; (2) abstraction of hydrogen from carbon atoms; (3) electron transfer reactions; and (4) reaction with sulfur-, nitrogen-, or phosphorus-moieties (Buxton et al., 1988; Lee and von Gunten, 2012; Minakata et al., 2009). Hence, the reactivity is highly structure-dependent such that similar reactivity can be found in compounds with the same structural features (Minakata et al., 2009). The encoding approach of MF is therefore suitable for deep learning to "learn" the relationship between structures and reactivity. However, to the best of our knowledge, MF has not been used to develop QSARs for the reactivity of $OH^{\cdot}$ radical toward organic contaminants.

The objective of this study was to demonstrate that DNN combined with MF can work well to develop QSARs without using any molecular descriptor. Here, the available dataset of 457 organic contaminants and their $OH^{\cdot}$-radical rate constants ($k_{OH\cdot}$) (Borhani et al., 2016) was first used to build QSAR, which was then validated by two other datasets of 17 organophosphorus fire retardants (Li et al., 2018a) and 118 micropollutants (Ortiz et al., 2017). The MF of these contaminants were obtained and used as the inputs for DNN. The results showed that the obtained DNN-MF models had comparable prediction accuracies to the traditional QSARs.

2. Materials and methods

2.1. Datasets

A dataset containing 457 organic contaminants from 27 diverse chemical classes and their $OH^{\cdot}$ radical rate constants was compiled by Borhani et al. (Borhani et al., 2016). This dataset was downloaded from the supplemental data and used in this study without any modification, except for adding the MF for all the compounds. Following Borhani et al., all the experimental rate constant data were extracted from the literature (Minakata et al., 2009; Monod et al., 2005; Wols and Hofman-Caris, 2012). If several rate constants were reported for the same contaminant, an average value was used. The rate constants $k_{OH^{\cdot}}$ ($M^{-1s^{-1}}$) were all obtained under standard conditions (i.e., 25 °C and 1 mol/L) and transformed into log units. The dataset was randomly split into a training set (80%) and a test set (20%). To ensure that both datasets contained all the chemical classes, we randomly selected 20% of the contaminants from each chemical class (27 in total) into the test set, and the remaining 80% was used in the training set. This was to ensure that DNN could "learn" the features of all the chemical classes. Following this approach, the dataset was split four times to form four groups (referred to as group 1 to 4), each group containing one training set and one test set. The reason for organizing the dataset into 4 groups was to investigate whether the data selection process affected the accuracy of the obtained model. In addition, another group (referred to as group 5) contained the same training set (90% of the contaminants) and test set (10% of the contaminants) as those used by Borhani et al. (Borhani et al., 2016), which was used to directly compare the performance of DNN-MF with the published results.

Another dataset that contained 18 organophosphorus flame retardants was recently reported and all the reaction constants were measured experimentally (Li et al., 2018a). 17 of them were chosen to test the prediction accuracy of the developed model because they had never been exposed to the model. The third dataset included 118 micropollutants and was used by Ortiz et al. to develop a descriptors-

based QSAR (Ortiz et al., 2017). Following their approach, we split the dataset to one training set (89 micropollutants) and one test set (29 micropollutants). This training set was then combined with the training set of group 1 to form a larger training set of 454 chemicals that contained more diverse structure features. DNN was next trained based on this larger training set to develop a QSAR that was finally applied to predict the $OH^{\cdot}$ radical rate constants of the test set of group 1, the 17 organophosphorus flame retardants, and the 29 micropollutants.

All the datasets, including the "SMILES" strings for all contaminants, are provided in the supplementary information (SI, Datasets excel file). "SMILES" stands for "simplified molecular-input line-entry system" and describes chemical structures using short ASCII strings in the form of line notation.

2.2. Generation of molecular fingerprints (MF)

To obtain the MF of all 593 organic compounds, we first obtained their "SMILES" strings by the ChemDraw program. Then, these "SMILES" strings were converted into MF by the RDKit program (https://www.rdkit.org/) using the command "*AllChem.GetMorganFingerprintAsBitVect()*". The generated MF were binary vectors (Figure S1) of the same length. The length was however adjustable, and the longer the length is, the more structural features are stored so that it was less likely for the features of different compounds to overlap.

To develop more accurate QSARs, we first investigated the effect of the vector length using the following values: 126, 512, 1024, 2048, 4096, and 8192 bits. Here, the root mean square error (RMSE) and coefficient of determination (R^2) were used to evaluate the performance of the developed models. RMSE is the standard deviation of the residuals (prediction errors) (eq. 1). When we applied the R^2 to the test set, the R^2 values were equal to the external explained variance (Q^2). The lower the RMSE and the higher the R^2 value are, the better the model is. The absolute relative error (ARE) (eq. 2) and the average absolute relative error (AARE) (eq. 3) for each chemical class were also calculated to obtain the numbers of chemicals in each chemical class that had prediction errors of < 2%, 2–4%, 4–6% or > 6%.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (logk_{OH\cdot}^{exp} - logk_{OH\cdot}^{pred})^2}{n}} \quad (1)$$

$$\mathrm{ARE} = \frac{|logk_{OH\cdot}^{exp} - logk_{OH\cdot}^{pred}|}{logk_{OH\cdot}^{exp}} \quad (2)$$

$$\mathrm{AARE} = \frac{\sum_{i=1}^{n} ARE_i}{n} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (logk_{OH\cdot}^{exp} - logk_{OH\cdot}^{pred})^2}{\sum_{i=1}^{n} (logk_{OH\cdot}^{exp} - mean(logk_{OH\cdot}^{exp}))^2} \quad (4)$$

where $mean(logk_{OH\cdot}^{exp})$ is the mean experimental $\log k_{OH\cdot}$.

2.3. Structure of DNN

DNN is a computer program designed to emulate human brains in terms of learning from the data in a manner similar to the human nervous system. A typical DNN is composed of a number of neurons from a few to millions, which are arranged in a series of layers (Figure S2). The input neurons in the input layer are designed to receive the external data, such as the MF used here, and the output neurons in the last layer are the final predictions made by the DNN, which will be used to compare with the true target data, such as $\log k_{OH\cdot}$. Between the input layer and the output layer are hidden layers, often more than one layer. The input data go into the DNN through the input layer, are then transformed in the hidden layers, and finally become the predictions in the output layer. The values in all neurons in the hidden and output layers are calculated by (sum of the values in the previous neurons × weight + bias), in which weights and biases can be updated based on the errors between the predictions and the target until the errors reach a minimum value. This process is the "learning" process of DNN. The number of layers and neurons is also called the "depth" and "width" of DNN, respectively. Larger numbers of layers and neurons mean deeper and wider DNN, which often have more powerful fitting ability and can achieve better accuracy on the prediction. However, too many layers and neurons often have the overfitting problem, that is, accurate prediction on the training set but worse prediction on the test set. The model development process is hence to develop an optimum architecture of the DNN with an appropriate fitting ability.

In this study, our DNN is composed of an input layer, several hidden layers, and an output layer (Figure S2). In each layer, there are numerous neurons accepting values from the neurons of the neighboring layer. In the input layer, the number of neurons was equal to the length of the MF. For instance, if the length of a MF was 512 bits, then there were 512 neurons in the input layer. The number of neurons in the output layer was 1 because there was only one reaction rate constant for each compound. The number of neurons in the hidden layers, called "hyperparameter", was set manually before the learning process began. Here, we focused on two most important hyperparameters: the number of hidden layer and the number of neurons, and investigated their effects on the performance of the DNN. The RMSE, ARE, and AARE values were also calculated to evaluate the effects of the hyperparameters. Detailed description of the theory behind DNN has been adequately described elsewhere (Fatemi, 2006; Lek and Guégan, 1999; Zupan and Gasteiger, 1993).

To avoid overfitting, the "dropout" method was applied to each hidden layer. Overfitting means that the model has a low RMSE on the training set but a high RMSE on the test set. It often results from the complexity of the models being too high. Models with high complexity may extract odd features that fit the training set well but are not applicable to the test set. The "dropout" value is the probability that the value of a neuron is not passed to the neurons in the next layer (abandoned). For example, a dropout value of 0.5 means that for each neuron, there is a 50% probability that its value is not used to calculate the values in the next layer, or in other words, 50% of randomly selected neurons will be abandoned. Employing the "dropout" method would lower the complexity of the established DNN models, thus controlling the potential of overfitting. The "dropout" value was also a hyperparameter. We set it at 0.2 for the neurons in the first hidden layer and 0.5 for the ones in all other hidden layers. The model training was stopped after 500 epochs (iterations).

3. Results and discussion

For DNN, the prediction accuracy is highly related to its structure, i.e., the numbers of layers and neurons. Moreover, the length of MF, i.e., the number of neurons in the input layer, should be adjusted manually. Hence, their effects on the prediction accuracy of DNN were first investigated.

3.1. Effects of the MF length and hyperparameters

MF store the structural features of compounds; therefore, the effect of their length on the performance of DNN was first investigated. Note that this parameter is not a hyperparameter of DNN. During the modeling process, the structure of the DNN was fixed, i.e., one input layer, three hidden layers each with 1024 neurons, and one output layer. The model was first established by training DNN with the training set of group 1 and then applied to the test set of group 1 to calculate the RMSE.

As shown in Fig. 1A, with increasing length of MF, the RMSE value first decreased and then remained low when the length was over 2048 bits. This result indicates that longer MF yielded better models. This was because MF of longer lengths could store more structural features

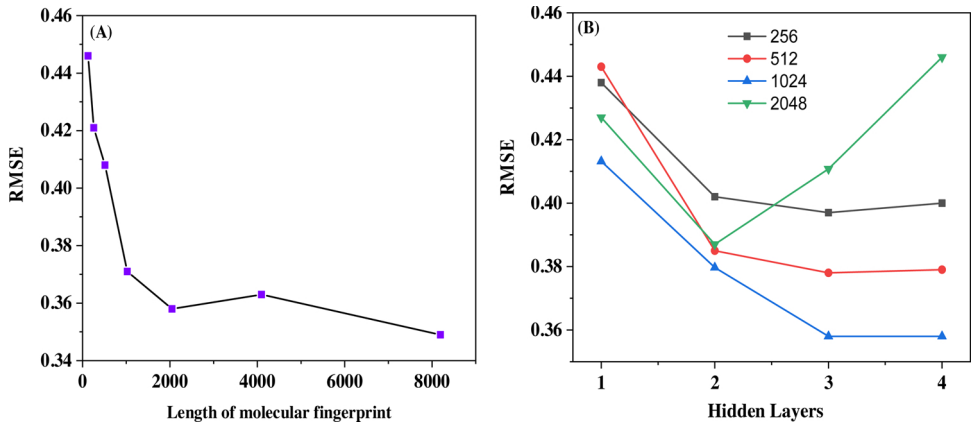


Fig. 1. The effects of (A) the length of molecular fingerprints and (B) the numbers of hidden layers and neurons on the RMSE values. Group 1 datasets were employed in the modeling.

so that it was less likely for the features of different compounds to overlap. Once more features were extracted by DNN, more accurate models were developed. Here, 2048 was chosen as the length of the MF for all the calculations below. This was because further increasing the length to more than 2048 bits only led to slight decrease in the RMSE, but made calculations more demanding.

Fig. 1B shows the effects of the hyperparameters on the performance of the models. Generally, the numbers of hidden layers and neurons were seen as the "depth" and "width" of DNN, respectively. Wider or deeper DNN can produce more complex models that yield lower RMSE on the training set because it can extract more features, but can more easily over fit to generate higher RMSE on the test set. Here, the number of hidden layers was changed from 1 to 4 and the number of neurons in all the hidden layers was varied from 256 to 2048. The length of all the MF was fixed at 2048 bits, as obtained above. As shown in Fig. 1B, the 2048 neurons in the hidden layers can be first excluded because of the largest RMSE values in most cases, and the RMSE also increased as the number of hidden layers increased from 2 to 4. Because the DNN with 2048 neurons was already the widest among all the DNNs (i.e., produced models with the highest complexity), further increasing the number of hidden layers would make the model even more complex, thus leading to higher RMSE on the test set, that is, overfitting. In this study, only the models with the smallest RMSE values were selected to be the optimum model to avoid the overfitting problem.

As the number of neurons increased from 256 to 1024, the RMSE almost always became smaller for the DNN with 1–4 hidden layers. This result indicated that a reasonably larger number of neurons in the hidden layers contributed to lower RMSE. For a given number of neurons in the hidden layers, increasing the number of hidden layers from 1 to 3 decreased the RMSE, but there was only a slight decrease in the RMSE when that number exceeded 3. Adding the fourth hidden layer negligibly contributed to further reduction in RMSE but significantly burdened the calculation process. Based on the above sensitivity study of the DNN structure, the optimal numbers of neurons and hidden layer were set as 1024 and 3, respectively.

3.2. Performance of the models developed by DNN-MF

Fig. 2A shows the scatter plot of the experimental versus the predicted $\log k_{OH\cdot}$ of the 96 compounds in the test set of group 1 based on the model trained using the training set of group 1 (RMSE = 0.358, R^2 = 0.747). For the datasets of groups 2, 3 and 4, the obtained models for the training sets showed similar RMSE values (0.387, 0.360, and 0.372, Table S1) and R^2 values (0.678, 0.651, and 0.664, Table S1) as the respective test set, indicating that the prediction accuracy of the "learned" models was robust, or independent of data splitting. Group 5 contained the same dataset as Borhani et al.'s, in which the training set was used to train the DNN model, and the obtained model was then applied to predict the test set of group 5 to get a RMSE and R^2 value of 0.384 and 0.669 (Table S1, group 5). This accuracy is comparable to the one reported by Borhani et al. (Table S1, RMSE = 0.352 and R^2 = 0.724) (Borhani et al., 2016). It should be noted that there were 9 duplicate contaminants in Borhani et al.'s dataset, but we did not delete them to make the comparison between their model and ours under the identical conditions. Also, the RMSE values only slightly increased after the 9 duplicate values were deleted (Table S1). As shown in Fig. 2B, the predicted versus the experimental $\log k_{OH\cdot}$ also showed similar patterns in both models. This result indicated that the model developed by DNN-MF had comparable prediction accuracy to the traditional one based on molecular descriptors; however, the new model avoided using any molecular descriptor.

When the model based on the training set of group 5 was used to predict the $\log k_{OH\cdot}$ values for the group 5 test set, the predictions deviated more for those with $\log k_{OH\cdot}$ values less than 9 than those greater than 9 (Fig. 2B), e.g., for acetic acid and pentachlorehane. This may be because the number of compounds with $\log k_{OH\cdot}$ less than 9 in the training set was much less than that with $\log k_{OH\cdot}$ greater than 9 (Fig. 2D). In other words, DNN "learned" less about the compounds with smaller $\log k_{OH\cdot}$. This phenomenon is understandable because DNN is a highly data-dependent method. The more data it has, the more accurate the prediction is. Nevertheless, DNN still showed promising results in developing predictive tools for reactivity estimation and its performance can be enhanced by providing a more diverse range of data for each structure group.

Table S2 lists the numbers of chemicals in each chemical class that have prediction errors of < 2%, 2–4%, 4–6% and > 6%. In most chemical classes, there are more chemicals with small ARE (< 2%) (377 vs. 229) and less chemicals with larger ARE (2–4%, 4–6% and > 6%) (48 vs. 119, 10 vs. 57 and 22 vs. 52, respectively) in our model than in the reported descriptors-based binary particle swarm optimization (BPSO) algorithm and multiple-linear regression (MLR) model (BPSO-MLR) (Borhani et al., 2016). The AARE values in our model were also mostly smaller than those in BPSO-MLR. These results indicated that our model was more accurate than BPSO-MLR in predicting the rate constants for most chemicals. Our model was only less accurate in predicting the rate constants for three classes, i.e., carboxyl, imidazole and triazines, as there were 3, 3 and 5 contaminants with ARE > 6% using our model while 1, 0, and 3 using BPSO-MLR, respectively.

To further validate the reliability of our model, the second dataset containing 17 organophosphorus flame retardants was applied to the model trained by using the training set of group 1. Compared to the experimental $\log k_{OH\cdot}$ values, the predicted $\log k_{OH\cdot}$ had RMSE = 0.306 and R^2 = 0.737 (Fig. 2A), which was similar to RMSE = 0.235 and R^2 = 0.862 by using a conventional descriptors-based QSAR (Li et al.,

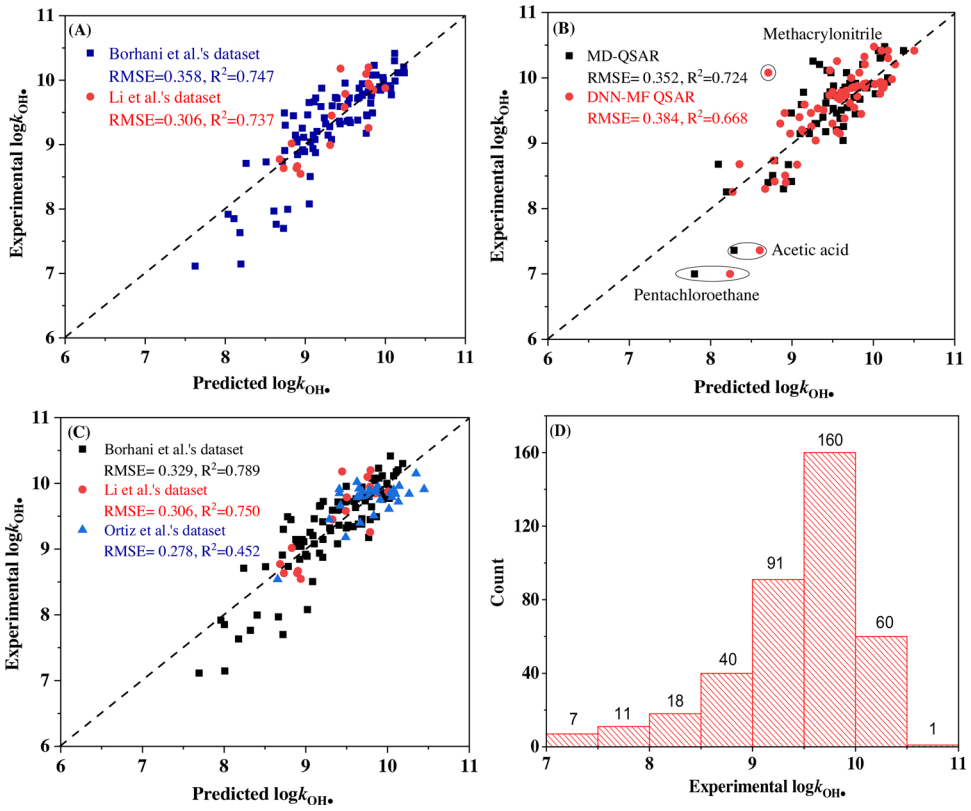


Fig. 2. The scatterplot of the predicted vs the experimental values of log$k_{OH\cdot}$ for (A) 96 compounds in the test set of group 1 and the 17 organophosphorus flame retardants in Li et al. and (B) the 69 compounds in the test set of group 5 based on the conventional QSARs versus the new DNN-MF model, both trained by the training set of group 5; (C) Comparison between the experimental and the predicted values of log$k_{OH\cdot}$ for the 96 compounds in the test set of group 1 (Borhani et al.), the 17 newly reported compounds in Li et al., and the 29 micropollutants in Ortiz et al. by the DNN-MF model trained by the larger training set of 454 chemicals. (D) The number of compounds in the training set of group 5 that has the reported log$k_{OH\cdot}$ values within each range, for example, 160 compounds with log$k_{OH\cdot}$ values between 9.5 and 10.

2018a). It should be noted that all of the chemicals in this dataset had never been exposed to our model. This satisfactory accuracy was nevertheless expected because all the functional groups in these 17 chemicals, including alkane, alcohol, benzene, ether, and halogenated groups, had already been "learned" by the DNN-MF model. For micropollutants that mostly have more complicated structures, this approach was still effective. As shown in Fig. 2C, the model obtained by combining the training sets of group 1 and the 89 micropollutants in Ortiz et al. had comparable prediction accuracies for the test set of group 1, the 17 flame retardants, and the 29 micropollutants (RMSE = 0.278-0.329) (Ortiz et al., 2017).

3.3. Comparison among different models

Table 1 lists the performance of six reported models that were developed based on different algorithms. In general, to obtain models with satisfactory prediction performance on an increasing number of chemicals (n from 55 to 526), more molecular descriptors (p from 4 to13) had to be involved in the reported models. However, our DNN-MF model showed comparable accuracy (R^2_{test} = 0.747, RMSE = 0.329) to the reported models even on the largest dataset (n = 593) and without using any molecular descriptor (p = 0). It should be noted that both R^2_{train} (0.972) and $RMSE_{train}$ (0.135) on the training set were better than the R^2_{test} (0.747) and $RMSE_{test}$ (0.329) on the test set. This phenomenon is common and reasonable because we trained our DNN on the training set and then applied it to the test set that it had never seen before (Myint et al., 2012). This should not be seen as overfitting, because we always selected the model with the smallest $RMSE_{test}$ when optimizing the DNN structure, as shown in Fig. 1B and explained in Section 3.1. This means that further increase or decrease in the complexity of DNN would lead to higher RMSE, which would yield overfitting or underfitting. Note that although our DNN-MF showed a slightly higher $RMSE_{test}$ and lower R^2_{test} than the MD-based models, its accuracy can be further increased by including a larger number of compounds in the modeling process because more meaningful features will be learned by the model. For the MD-based models, however, with increasing number of organic compounds, their accuracy might decrease, as shown in Table 1 where the $RMSE_{test}$ increased with an increasing number of parameters in the model ($RMSE_{test}$: 0.079-0.356). Given the fact that more and more contaminants may arise in the future, the DNN-MF based approach will show great applications in the environmental field.

We also compared our model with the reported group contribution based model that used error goals ($EG = \frac{\text{Predicted value}}{\text{Experimental value}}$) to evaluate the model performance (RMSE was not provided) (Minakata et al., 2009). Note that their dataset was already included in ours so we directly compared the performance of their model with that of ours. Their model could predict the rate constants of 334 out of 435 contaminants

Table 1
Comparison among different models for their performance in predicting aqueous $k_{OH\cdot}$ values.

Model	Algorithm	n[a]	p[b]	Training Set		Test Set	
				R^2_{train}	$RMSE_{train}$	R^2_{test}	$RMSE_{test}$
(Wang et al., 2009)	MLR	55	4	0.905	0.139	0.962	0.079
(Kui et al., 2009)	GA-MLR[c]	78	4	0.735	0.174	0.76	0.20
(Sudhakaran and Amy, 2013)	PCA-MLR[d]	83	2	0.918	–	–	–
(Jin et al., 2015)	MLR	118	7	0.823[f]	0.204	0.772	0.329
(Borhani et al., 2016)	BPSO-MLR[e]	457	8	0.716	0.347	0.724	0.356
(Luo et al., 2017)	MLR	526	13	0.805[f]	0.165	0.802	0.232
This study	DNN-MF	593	0	0.972	0.135	0.789	0.329[g]

[a] n = total number of chemicals in the dataset.
[b] p = number of molecular descriptors.
[c] GA: genetic algorithm.
[d] PCA: principal component analysis.
[e] BPSO: binary particle swarm optimization.
[f] The R^2_{adj} value was reported instead.
[g] This value was for Borhani et al.'s test set (Fig. 2C).

(77%) that had the error goal (EG) ($EG = \frac{\text{Predicted value}}{\text{Experimental value}}$) in the range of 0.5–2 (Minakata et al., 2009), whereas our model could predict 416 out of 457 contaminants (91%) within the same EG range. In short, these comparison results confirmed that the DNN-MF had satisfactory performance in predicting chemical reactivity.

3.4. The feasibility of the DNN-molecular fingerprints (MF) approach: model utilization and model development

The DNN-MF approach combines simple MF with powerful DNN to develop QSAR models. To extend this approach to a wide range of applications, we'd like to address two important questions: one is how to apply the established QSARs to new compounds, i.e., model utilization, and the other is how to develop new models for different tasks, i.e., model development. A related question is whether the knowledge of DNN is necessary in the above context.

For model utilization, the background of DNN is not necessary. Once a model has been established, the parameters in DNN (weights and biases in Figure S2) have been fixed. Users just need to input the MF of new compounds into the model, and DNN will calculate the prediction results using these fixed parameters. MF can be more easily obtained than molecular descriptors, simply by transforming "SMILES" strings of new compounds using the RDKit program. The "SMILES" strings of new compounds can be obtained from their chemical names, CAS numbers or chemical structures (in this study, we used chemical structures).

To make the established DNN-ML models broadly available, it will be useful to develop APPs or web applications for automatic calculation. Users can simply input the chemical names, CAS numbers or the chemical structures of new compounds and click the prediction button. The APP or web applications will automatically retrieve the MF, feed them into the trained DNN, and return the prediction results. In comparison, for molecular descriptors-based QSARs, molecular descriptors are typically not available for new compounds and one needs both advanced physicochemical knowledge to select appropriate descriptors and the ability to use different software such as Gaussian to calculate the descriptors.

For model development, the knowledge of DNN is necessary if one wants to develop new MF-based QSARs. Fortunately, learning how to use DNN is much easier than learning all necessary chemical, computational and software knowledge for molecular descriptors. There are many packages that can be directly used in Python. In MATLAB, there are also a number of built-in DNNs. One just needs to adjust the hyperparameters of DNN for a specific application. With the rapid development of computer science and data science, many more researchers will have the ability to run DNN in the near future. For molecular descriptors-based QSARs, however, the learning curve is much steeper. This is probably part of the reason that although numerous QSARs have been developed and available for a few decades, they are mostly used by a small group of researchers.

4. Conclusions

This work combined MF with DNN to develop predictive models for the oxidation rate constants of 593 organic compounds by $OH^{\cdot}$ radical. The optimum length of MF was 2048 bits, and the optimized architecture of DNN was 1 input layer with 2048 neurons, 3 hidden layers each with 1024 neurons, and 1 output layer with 1 neuron. The model developed by the DNN-MF approach showed low RMSE values, e.g., 0.329 on the largest dataset (n = 593) and without using any molecular descriptor. This work for the first time showed that the combination of MF and DNN could provide simple, robust models that had comparable prediction accuracies to the traditional QSARs that relied on complex molecular descriptors. Simply providing the DNN with the most basic chemical information (i.e., MF) was sufficient for the DNN to "learn" the relationship between the reactivity and the chemical structures. It also should be noted that the generation of MF was straightforward, the MF can be much more easily understood than molecular descriptors, and the development and use of DNN-MF models was independent of any advanced knowledge of chemical reactions, chemical properties, and descriptor calculations. Developing QSARs using this new approach is thus both useful and desirable for environmental researchers who do not have rich chemical and computational knowledge.

Compared with traditional QSARs, our DNN models did not indicate which chemical features were most related to the reactivity and why the model chose those features. The features extracted by DNN might be statistical results, but it was also possible that meaningful mechanistic features had been "learned", which can facilitate mechanistic understanding. On the one hand, one would argue that mechanistic knowledge may not be necessary for prediction purposes, especially given the large number of chemicals that exist in the environment and the likely increasing number of them to be developed and released. On the other hand, our on-going work aims to unveil the "black-box" model and identify which features can be learned by DNN for the prediction purposes, which will assist in obtaining physical insights into the QSARs. Nevertheless, the DNN-MF approach is promising in the development of simple, powerful QSARs. In addition to predicting $k_{OH\cdot}$ of organic compounds, this new approach will likely show exciting applications in and can be easily expanded to many other areas of the environmental field, such as prediction of various biotic and abiotic degradation rate constants, adsorption, transport, plant uptake, and toxicity.

Acknowledgements

This material is based upon work supported by the U.S. National Science Foundation under Grants CBET-1804708.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.jhazmat.2019.121141.

References

Borhani, T., Saniedanesh, M., Bagheri, M., Lim, J., 2016. QSPR prediction of the hydroxyl radical rate constant of water contaminants. Water Res. 98, 344–353.
Buxton, G.V., Greenstock, C.L., Helman, W.P., Ross, A.B., 1988. Critical Review of rate constants for reactions of hydrated electrons, hydrogen atoms and hydroxyl radicals (·OH/·O− in Aqueous Solution. J. Phys. Chem. Ref. Data 17 (2), 513–886.
Canonica, S., Tratnyek, P.G., 2003. Quantitative structure-activity relationships for oxidation reactions of organic chemicals in water. Environ. Toxicol. Chem. 22, 1743–1754.
Cheng, Z., Yang, B., Chen, Q., Shen, Z., Yuan, T., 2017. Quantitative relationships between molecular parameters and reaction rate of organic chemicals in Fenton process in temperature range of 15.8 °C - 60 °C. Chem. Eng. J. 350, 534–540.
Coley, C.W., Barzilay, R., Jaakkola, T.S., Green, W.H., Jensen, K.F., 2017. Prediction of organic reaction outcomes using machine learning. ACS Cent. Sci. 3, 434–443.
Colón, D., Weber, E.J., Anderson, J.L., 2006. QSAR Study of the Reduction of Nitroaromatics by Fe(II) Species. Environ. Sci. Technol. 40 (16), 4976–4982.
Fatemi, M.H., 2006. Prediction of ozone tropospheric degradation rate constant of organic compounds by using artificial neural networks. Anal. Chim. Acta 556 (2), 355–363.
Free, S.M., Wilson, J.W., 1964. A mathematical contribution to structure-activity studies. J. Med. Chem. 7 (4), 395–399.
Glen, R.C., Bender, A., Arnby, C.H., Carlsson, L., Idrugs, B.-S., 2006. Circular fingerprints: flexible molecular descriptors with applications from physical chemistry to ADME. IDrugs 9, 199–204.
Hughes, T.B., Swamidass, J.S., 2017. Deep learning to predict the formation of quinone species in drug metabolism. Chem. Res. Toxicol. 30, 642–656.
Jin, X., Peldszus, S., Huck, P.M., 2015. Predicting the reaction rate constants of micropollutants with hydroxyl radicals in water using QSPR modeling. Chemosphere 138, 1–9.
Kamath, V., Pai, A., 2017. Application of Molecular Descriptors in Modern Computational Drug Design-An Overview.Application of Molecular Descriptors in Modern Computational Drug Design-An Overview.
Klopmand, G., 1992. Concepts and applications of molecular similarity, by Mark A. Johnson and Gerald M. Maggiora, eds., John Wiley & Sons, New York, 1990, 393 pp. Price: $65.00. J. Comput. Chem. 13 (4), 539–540.
Krizhevsky, A., Sutskever, I., in neural, H.-G.E, 2012. Imagenet classification with deep convolutional neural networks. Commun. ACM 60, 84–90.

Kubinyi, H., 1997. QSAR and 3D QSAR in drug design Part 2: applications and problems. Drug Discov. Today 2 (12), 538–546.
Kušić, H., Rasulev, B., Leszczynska, D., Leszczynski, J., Koprivanac, N., 2009. Prediction of rate constants for radical degradation of aromatic pollutants in water matrix: a QSAR study. Chemosphere 75 (8), 1128–1134.
Lee, Y., von Gunten, U., 2012. Quantitative structure–activity relationships (QSARs) for the transformation of organic micropollutants during oxidative water treatment. Water Res. 46 (19), 6177–6195.
Lek, S., Guégan, J.F., 1999. Artificial neural networks as a tool in ecological modelling, an introduction. Ecol. Modell. 120 (2), 65–73.
Li, C., Wei, G., Chen, J., Zhao, Y., Zhang, Y.-N., Su, L., Qin, W., 2018a. Aqueous OH radical reaction rate constants for organophosphorus flame retardants and plasticizers: experimental and modeling studies. Environ. Sci. Technol. 52, 2790–2799.
Li, C., Zheng, S., Li, T., Chen, J., Zhou, J., Su, L., Zhang, Y.-N., Crittenden, J.C., Wang, D., Zhu, S., Zhao, Y., 2018b. Quantitative structure-activity relationship models for predicting reaction rate constants of organic contaminants with hydrated electrons and their mechanistic pathways. Water Res. 151, 468–477.
Luo, S., Wei, Z., Spinney, R., Villamena, F.A., Dionysiou, D.D., Chen, D., Tang, C.-J., Chai, L., Xiao, R., 2018. Quantitative structure–activity relationships for reactivities of sulfate and hydroxyl radicals with aromatic contaminants through single–electron transfer pathway. J. Hazard. Mater. 344, 1165–1173.
Luo, X., Yang, X., Qiao, X., Wang, Y., Chen, J., Wei, X., Peijnenburg, W.J.G.M., 2017. Development of a QSAR model for predicting aqueous reaction rate constants of organic chemicals with hydroxyl radicals. Environ. Sci. Process. Impacts 19 (3), 350–356.
Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E., Svetnik, V., 2015. Deep neural nets as a method for quantitative structure–Activity relationships. J. Chem. Inf. Model. 55 (2), 263–274.
Mansouri, K., Abdelaziz, A., Rybacka, A., Roncaglioni, A., Tropsha, A., Varnek, A., Zakharov, A., Worth, A., Richard, A.M., Grulke, C.M., Trisciuzzi, D., Fourches, D., Horvath, D., Benfenati, E., Muratov, E., Wedebye, E., Grisoni, F., Mangiatordi, G.F., Incisivo, G.M., Hong, H., Ng, H.W., Tetko, I.V., Balabin, I., Kancherla, J., Shen, J., Burton, J., Nicklaus, M., Cassotti, M., Nikolov, N.G., Nicolotti, O., Andersson, P.L., Zang, Q., Politi, R., Beger, R.D., Todeschini, R., Huang, R., Farag, S., Rosenberg, S.A., Slavov, S., Hu, X., Judson, R.S., 2016. CERAPP: collaborative estrogen receptor activity prediction project. Environ. Health Perspect. 124 (7), 1023–1033.
McGregor, M.J., Pallai, P.V., 1997. Clustering of large databases of compounds: using the MDL “Keys” as structural descriptors. J. Chem. Inf. Comput. Sci. 37 (3), 443–448.
Minakata, D., Li, K., Westerhoff, P., Crittenden, J., 2009. Development of a group contribution method to predict aqueous phase hydroxyl radical (HO•) reaction rate constants. Environ. Sci. Technol. 43 (16), 6220–6227.
Monod, A., Poulain, L., Grubert, S., Voisin, D., Wortham, H., 2005. Kinetics of OH-initiated oxidation of oxygenated organic compounds in the aqueous phase: new rate constants, structure–activity relationships and atmospheric implications. Atmos. Environ. 39 (40), 7667–7688.
Moosavi, S., Chidambaram, A., Talirz, L., Haranczyk, M., Stylianou, K.C., Smit, B., 2019. Capturing chemical intuition in synthesis of metal-organic frameworks. Nat. Commun. 10 (1), 539.
Myint, K.-Z., Wang, L., Tong, Q., Xie, X.-Q., 2012. Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. Mol. Pharm. 9 (10), 2912–2923.
Olier, I., Sadawi, N., Bickerton, G.R., Vanschoren, J., Grosan, C., Soldatova, L., King, R.D., 2018. Meta-QSAR: a large-scale application of meta-learning to drug design and discovery. Mach. Learn. 107 (1), 285–311.
Ortiz, E.V., Bennardi, D.O., Bacelo, D.E., Fioressi, S.E., Duchowicz, P.R., 2017. The conformation-independent QSPR approach for predicting the oxidation rate constant of water micropollutants. Environ. Sci. Pollut. Res. - Int. 24 (35), 27366–27375.
Ryan, K., Lengyel, J., Shatruk, M., 2018. Crystal structure prediction via deep learning. J. Am. Chem. Soc. 140, 10158–10168.
Salter-Blanc, A.J., Bylaska, E.J., Johnston, H.J., Tratnyek, P.G., 2015. Predicting Reduction Rates of Energetic Nitroaromatic Compounds Using Calculated One-Electron Reduction Potentials. Environ. Sci. Technol. 49 (6), 3778–3786.
Salter-Blanc, A.J., Bylaska, E.J., Lyon, M.A., Ness, S.C., Tratnyek, P.G., 2016. Structure–Activity relationships for rates of aromatic amine oxidation by manganese dioxide. Environ. Sci. Technol. 50 (10), 5094–5102.
Su, H., Yu, C., Zhou, Y., Gong, L., Li, Q., Alvarez, P., Long, M., 2018. Quantitative structure–activity relationship for the oxidation of aromatic organic contaminants in water by TAML/H2O2. Water Res. 140, 354–363.
Sudhakaran, S., Amy, G.L., 2013. QSAR models for oxidation of organic micropollutants in water based on ozone and hydroxyl radical rate constants and their chemical classification. Water Res. 47 (3), 1111–1122.
Wang, Y., Chen, J., Li, X., Zhang, S., Qiao, X., 2009. Estimation of aqueous-phase reaction rate constants of hydroxyl radical with phenols, Alkanes and alcohols. QSAR Comb. Sci. 28 (11–12), 1309–1316.
Wei, J.N., Duvenaud, D., Aspuru-Guzik, A., 2016. Neural networks for the prediction of organic chemistry reactions. ACS Cent. Sci. 2, 725–732.
Wols, B.A., Hofman-Caris, C.H.M., 2012. Review of photochemical reaction constants of organic micropollutants required for UV advanced oxidation processes in water. Water Res. 46 (9), 2815–2827.
Wu, Y., Wang, G., 2018. Machine learning based toxicity prediction: from chemical structural description to transcriptome analysis. Int. J. Mol. Sci. 19 (8).
Xiao, R., Ye, T., Wei, Z., Luo, S., Yang, Z., Spinney, R., 2015. Quantitative structure-–activity relationship (QSAR) for the oxidation of trace organic contaminants by sulfate radical. Environ. Sci. Technol. 49 (22), 13394–13402.
Ye, T., Wei, Z., Spinney, R., Dionysiou, D.D., Luo, S., Chai, L., Yang, Z., Xiao, R., 2017. Quantitative structure–activity relationship for the apparent rate constants of aromatic contaminants oxidized by ferrate (VI). Chem. Eng. J. 317, 258–266.
Ye, W., Chen, C., Wang, Z., Chu, I.-H., Ong, S., 2018. Deep neural networks for accurate predictions of crystal stability. Nat. Commun. 9 (1), 3800.
Zhou, Z., Li, X., Zare, R.N., 2017. Optimizing chemical reactions with deep reinforcement learning. ACS Cent. Sci. 3, 1337–1344.
Zupan, J., Gasteiger, J., 1993. Neural Networks for Chemists: An Introduction. John Wiley Sons, Inc.