Chemical Engineering Journal 408 (2021) 127998

Contents lists available at ScienceDirect

Chemical Engineering Journal

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

Molecular image-convolutional neural network (CNN) assisted QSAR models for predicting contaminant reactivity toward OH radicals: Transfer learning, data augmentation and model interpretation

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

[a] Department of Civil and Environmental 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

ARTICLE INFO

Keywords:
Convolutional neural network (CNN)
Hydroxyl radical
Model interpretation
Machine learning
Molecular images
QSARs

ABSTRACT

In this study, we used molecular images as a representation for organic compounds and combined them with a convolutional neural network (CNN) to develop quantitative structure-activity relationships (QSARs) for predicting compound rate constants toward OH radicals. We applied transfer learning and data augmentation to train molecular image-CNN models and the Gradient-weighted Class Activation Mapping (Grad-CAM) method to interpret them. Results showed that data augmentation and transfer learning can effectively enhance the robustness and predictive performance of the models, with the root-mean-square-error (RMSE) values on the test dataset ($RMSE_{test}$) decreasing from (0.395–0.45) to (0.284–0.339) after applying data augmentation, and the RMSE on the training dataset ($RMSE_{train}$) decreasing from (0.452–0.592) to (0.123–0.151) after applying transfer learning. The obtained molecular image-CNN models showed comparative predictive performance ($RMSE_{test}$ 0.284–0.339) with the molecular fingerprint-based models ($RMSE_{test}$ 0.30–0.35). Grad-CAM interpretation showed that the molecular image-CNN models correctly chose the molecular features in the images and identified key functional groups that influenced the reactivity. The applicability domain analysis showed that the molecular image-CNN models have a broader applicability domain than molecular fingerprints-based models and the reactivity of any new compounds with a maximum similarity of over 0.85 to the compounds in the training dataset can be reliably predicted. This study demonstrated that molecular image-CNN is a new tool to develop QSARs for environmental applications and can be used to build trustful models that make meaningful predictions.

1. Introduction

Quantitative structure-activity relationships (QSARs) play important roles in the environmental field [1–3], based on which one can readily predict the activity of new compounds so that labor-intensive and expensive experiments can be largely avoided. In water treatment, for example, QSARs are often developed for different oxidants or reductants, such as H_2O_2 [2,4], O_3 [4,5], Fe(VI) [6], HO• [1,5,7], $SO_4^{\bullet-}$ [8,9], and hydrated electrons [10]. An important step to develop QSAR models is to express different organic compounds numerically. Three most often used representations are molecular descriptors, molecular fingerprints and the group contribution method. These representations extract information of compounds to numbers with certain physicochemical meanings. For example, molecular descriptors can quantify different physicochemical properties of organic compounds, such as Hammett constants, reduction potential, and topological polar surface area. Molecular fingerprints encode a compound structure into a binary vector (i.e., only containing 0 s and 1 s), in which only 1 s represent certain atom groups in the structure and the positions of 1 s in the vector represent the identity of the atom groups. A more specific example of molecular fingerprints is listed in our recent papers [11,12]. The group contribution method decomposes a compound structure into several sub-groups, with each group contributing to a portion of the reactivity [13–15]. All of these three representations are sophisticated representations of organic compounds, and can be further combined with various statistical and/or machine learning algorithms to develop

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

https://doi.org/10.1016/j.cej.2020.127998
Received 10 September 2020; Received in revised form 30 November 2020; Accepted 1 December 2020
Available online 7 December 2020

QSAR models. Compared with traditional statistical methods, such as multiple linear regression, machine learning algorithms are particularly capable of handling complex non-linear relationships based on big data. Non-linear relationships between chemical structures and reactivity may exist when more and more organic contaminants are involved in the dataset. Hence, QSAR models developed by machine learning methods often show better predictive performance than those by multiple linear regression [1], especially for large datasets.

However, a simpler and more intuitive method to represent compounds is by 2D molecular images and every chemical has its unique molecular image. Different chemicals can be differentiated from each other in their images based on the type of atoms, their relative positions in the images and the connections of atoms in the molecules. For example, chlorine can be used to differentiate phenol and 2-chlorophenol while the position of chlorine can be used to differentiate 2-cholrophenol and 3-chlorophenol. For compounds with enantiomer structures, molecular images can still be used by using solid or dotted wedge-shapes to indicate the bond positions in the 3D space. Although traditional machine learning algorithms cannot handle image data efficiently, with the development of deep learning, especially the convolutional neural network (CNN), image data can be directly handled to, for example, develop QSAR models. The first example of such an application is "Chemception" [16], which feeds 2D images of molecules to a deep CNN. Chemception slightly outperforms molecular fingerprint-based QSAR models in predicting biochemical activity and solvation but slightly underperforms in predicting toxicity. In Chemception, the 2D images of molecules are not 2D drawing of chemicals but "grid" images, in which a 2D drawing is mapped onto a 80 × 80 grid, where each atom is assigned a number based on its atomic mass unit, bonds are assigned number 2, and the other parts of the grid are empty (i.e., vacuum) and defaulted to number 0. Following that work, Fernandez et al. directly used 2D drawings of chemicals to develop molecular image-CNN models to predict toxicity of compounds [17]. They found that the new models showed comparable predictive accuracy with molecular descriptor-based models. Shi et al. also used the same approach to building predictive models for absorption, distribution, metabolism, elimination, and toxicity of drug compounds and demonstrated a comparable performance to available machine learning models based on manual structural description and feature selection [18]. However, previous studies have not fully used the power of CNN, such as transfer learning and data augmentation, likely because the data volume in biomedicine is often large such that transfer learning and data augmentation are not needed. In the environmental field, unfortunately, data scarcity is a common issue; hence, transfer learning and data augmentation (details below) are expected to mitigate its impact [19,20]. After model development, CNN models in the previous studies are also not interpreted, although techniques of interpreting CNN models are available.

Transfer learning refers to applying a model being pre-trained for one task to another with some modification. There are many CNN architectures that have been well-trained on large datasets to effectively extract features from numerous images, such as edges, colors and shapes of objects. Adapting these architectures to a much smaller, different dataset will allow extraction of more relevant features from the dataset to achieve more accurate predictions. This is especially applicable here given that molecular images are often much less complex than many other images such as those in the ImageNet dataset [21]. Data augmentation can be readily achieved by, for example, rotating or flipping images, during which the objects in the images change their positions to generate new data points but their labels remain unchanged, which is not possible in the molecular descriptor-, molecular fingerprint- and group contribution-based methods. Both transfer learning and data augmentation can help address the issue of data scarcity, a common problem for experimental research such as measuring contaminant reactivity but detrimental to building robust ML models that require a large amount of data. For machine learning interpretation, our previous study has listed some interpretation methods [12]. For CNN interpretation, there are generally three methods: Gradients, DeconvNets and Guided Backpropagation, to generate saliency maps, i.e., heat maps. Adebayo et al. showed that only Gradients are effective among these three methods [22]. Hence, we employed a gradients-based method to interpret our CNN model in this study. The Gradient-weighted Class Activation Mapping (Grad-CAM) method was recently developed, which can highlight the regions of the molecular images that are linked to model predictions by the CNN. Grad-CAM works by using gradients of a given target to highlight the most related regions in images to the predicted target, and the detailed working mechanism has been fully described in the literature (a brief introduction to Grad-CAM is in Text S2) [23]. This interpretation step is crucial for validating and trusting our "black box" deep CNN models.

Here, we for the first time extended the molecular image-CNN method to the environmental field by developing QSARs to predict the rate constants of organic compounds toward HO• radicals ($\log k_{HO\bullet}$) in the aqueous phase. We applied transfer learning and data augmentation to train our molecular image-CNN models and interpret it by the Grad-CAM method. The effects of transfer learning and data augmentation on the robustness and predictive performance of models were investigated. A dataset containing 1089 compounds and their rate constants toward OH radicals in water, which was previously used to successfully develop molecular fingerprint-based models [12], was used here. We also compared the molecular image-CNN models with those molecular fingerprint-based models. Finally, we used Grad-CAM to interpret which regions of the images were chosen by the CNN to make prediction. To intuitively show how Grad-CAM works, we developed classification models to successfully recognize 5 randomly selected functional groups in the 1089 compounds, namely –OH, –CN, halogen, =O and aromatic carbon. Grad-CAM was then applied to check if this classification model chose the correct molecular features in the molecular images for these functional groups. After validating the reliability of Grad-CAM, we interpreted our final regression model to evaluate its trustability.

2. Materials and methods

2.1. Dataset and 2D molecular images

A dataset containing 1,089 organic compounds and their rate constants toward HO• radicals was compiled from our previous study and an detail description of the dataset can be found there [12]. Although we used the same dataset as the previous one, we employed a very different approach to develop QSAR models. In this way, we can make a fair comparison between the newly developed molecular image-CNN method and the previously published molecular fingerprint-machine learning method. Through this comparison, we highlighted a few advantages of the new approach, including transfer learning, data augmentation and structure visualization. This dataset contains $\log k_{\mathrm{HO}\bullet}$ values that are close to the diffusion-control limit [24,25]. This may cause problems for other methods such as the group contribution method but not for deep learning methods because the latter can still learn the structural characteristics of those compounds whose reactivity will approach the diffusion-control limit. For example, if the rate constants of one class of compounds are always close to the diffusion-control limit, deep learning models can still correctly predict the rate constants of similar compounds—also close to the diffusion-control limit. Solvation and steric hindrance can be expected to affect the reactions, but because these effects have already been reflected in the measured $\log k_{\mathrm{HO}\bullet}$, deep learning models can still learn the relationship between $\log k_{\mathrm{HO}\bullet}$ and these effects.

We used the same training, validation and test datasets (8:1:1) as in the previous study [12] so that we can make a fair comparison between molecular fingerprint-based and molecular image-CNN models. The dataset was also rearranged to form five groups of datasets, each containing one training, one validation and one test dataset. This was to check if the model performance was dependent on the data splitting or not. The training dataset was used to train a CNN model, the validation

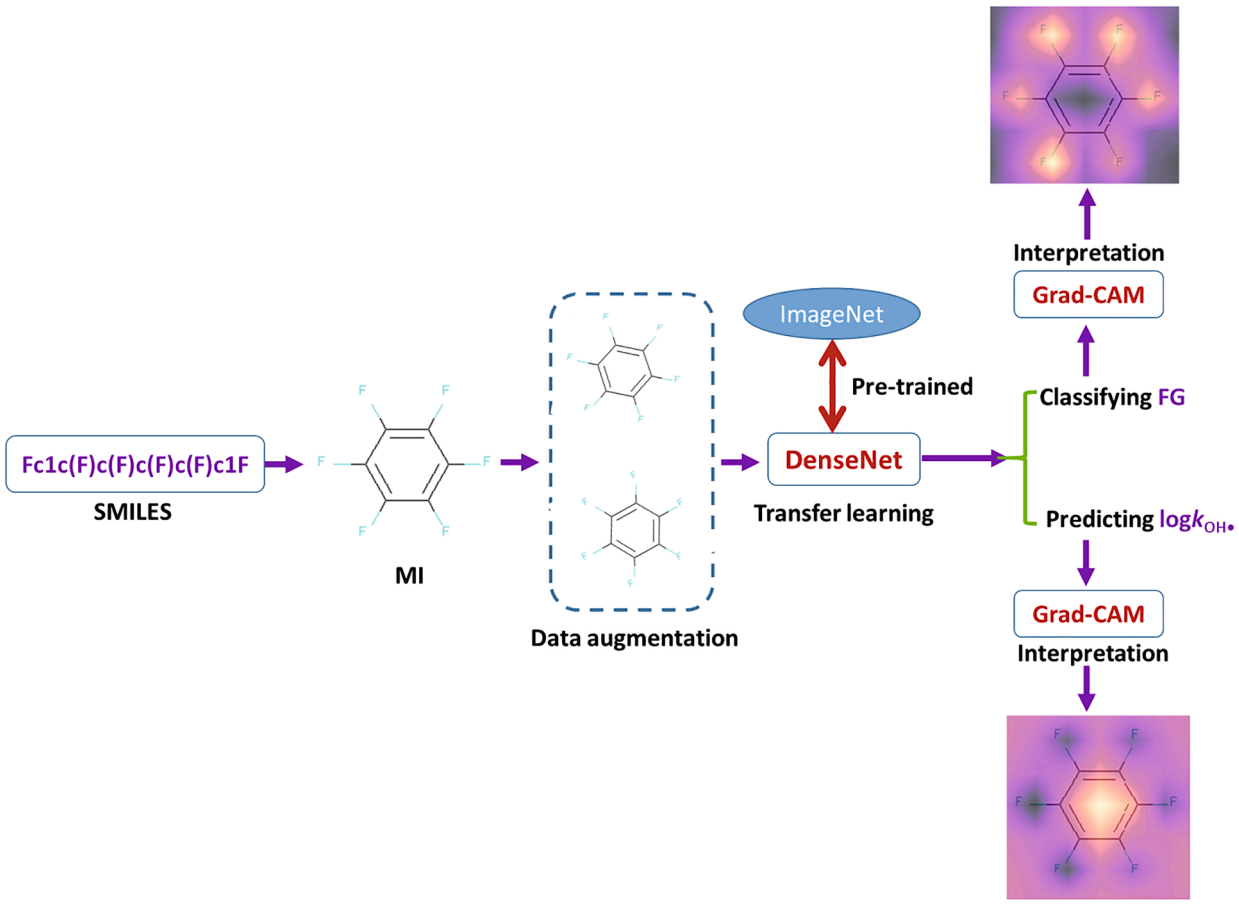

Fig. 1. Flow chart of the model development and interpretation process for classification and regression applications.

dataset was used to avoid overfitting the data, and the test dataset was used to evaluate the generalization ability of the CNN model. Note that the test dataset had never been exposed to the model during the training process.

The 'SMILES' strings for all compounds were obtained by the ChemDraw program and then converted to the corresponding 2D structural images with a uniform style by the RDKit package in Python®. "SMILES" is short for Simplified Molecular-Input Line-Entry System, which refers to a line notation for encoding molecular structures and specific instances [26]. We then trained one model for each of the five groups and obtained five models as model-C-x or model-R-x, where C represents the classification application, R represents the regression application, and × = 1–5. For the classification application, five functional groups, namely, –OH, –CN, halogen, =O and aromatic carbon, were randomly chosen as the targets for demonstration of the efficiency of the Grad-CAM method. The chosen functional groups do not have special reactivity or mechanisms toward OH radicals. Note that we did not have to limit to these five functional groups and any other functional groups can be chosen here. This classification example was only used to show how Grad-CAM works; therefore, it did not matter which functional groups were chosen. Definitely, choosing representative functional groups for OH radical reaction is also feasible and we expect to see a similar finding as what we have observed. For each functional group, we manually labeled the 1,089 compounds with a combination of 0s and 1s in which 1 represents its presence and 0 is for its absence. This is different from the regression modeling, where the labels, i.e., log-transformed rate constants $\log k_{HO•}$, were recorded when we collected the data points. When making predictions for $\log k_{HO•}$, we only need the 2D images of the chemicals as the input to the molecular image-CNN models.

2.2. *Transfer learning, DenseNet121 architecture, and data augmentation*

For transfer learning, we used a pre-trained DenseNet121 that had been well trained on the ImageNet dataset [21], which contains over 1,000,000 images in 1,000 categorical classes. This well-trained DenseNet121 can thus effectively extract feature information from images. DenseNet121, the state-of-the-art architecture of CNNs developed in 2016 [27], is a CNN architecture of DenseNet with 121 layers. The details about DenseNet have been well-documented (a brief introduction to DenseNet121 is in Text S1) [27]. When applying transfer learning to our task, we first froze the convolutional layers of DenseNet121 and only trained its last fully connected layer. The model performance, however, was not satisfactory because molecular images are still different from the images in the ImageNet. While allowing the last fully connected layer of DenseNet121 being well trained, we then unfroze the convolutional layers and trained them again so they were more adaptive to our molecular images. The effect of transfer learning was investigated by comparing the predictive performance of the models trained with and without applying transfer learning to the same dataset.

When training models for the molecular images, we also applied data augmentation to expand the volume of our dataset. This can help the models to well recognize molecules independent of their positions and angles in the images. For example, phenol is still phenol no matter how we flip or rotate its image. For every molecular image, we randomly flipped it horizontally or vertically or rotated it with a random degree of

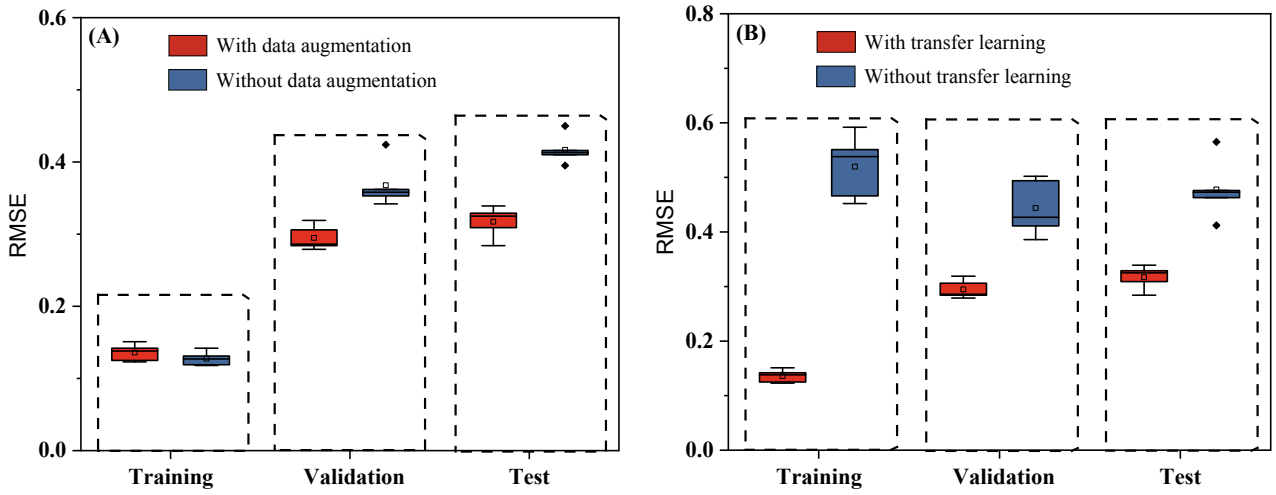

Fig. 2. The effect of data augmentation (A) and transfer learning (B) on the predictive performance of the CNN models.

less than 90° in every training epoch. Early stopping was used to control overfitting during the training process. The effect of data augmentation was investigated by comparing the predictive performance of models trained with and without applying data augmentation to the same dataset.

After training the models, we randomly chose one model from the 5 models for either the classification or regression application (i.e., Model-C-x or Model-R-x) and applied Grad-CAM to interpret which regions of the images were chosen as the most related feature(s) for the predictions. Fig. 1 summarizes the process of model development and interpretation, including molecular image generation, data augmentation, transfer learning, the CNN architecture used, and the model interpretation by Grad-CAM.

2.3. Evaluation metrics

For the classification application, accuracy, F1 score and AUC (Area Under The Curve)-ROC (Receiver Operating Characteristics) were used to evaluate performance of the developed models. Accuracy is the percentage of compounds that are correctly classified in the total number of compounds, as shown in eq. (1). F1 score ranges from 0 to 1 and is calculated by the precision and recall, as shown in eq. (2), in which a F1 score of close to 1 suggests a good model performance. A detailed explanation of F1 score is in Text S4. An AUC-ROC curve quantifies the classification performance at various thresholds. With the maximum AUC-ROC value of 1, higher AUC-ROC values indicate better model performance.

$$\text{Accuracy} = \frac{TN + TP}{TN + TP + FP + FN} \tag{1}$$

$$\text{F1score} = 2\text{x}\frac{precision^{*}recall}{precision + recall} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

where TN and TP are the true negative and positive cases that are correctly identified while FP and FN are the false negative and positive cases that are not correctly identified.

For the regression application, the root mean square error (RMSE), mean absolute error (MAE) and R^2 were used to evaluate the performance of the developed models. RMSE is the standard deviation of the residuals (prediction errors) (eq. (5)), MAE measures the average absolute difference between the predicted and real values (eq. (6)), and R^2 is the coefficient of determination ranging from 0 to 1 (eq. (7)) [28]. Good models have low RMSE and MAE values and R^2 values closer to 1.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(logk_{HO.}^{exp} - logk_{HO.}^{pred}\right)^2}{n}} \tag{5}$$

$$\text{MAE} = \frac{\sum_{i}^{n}\left|logk_{HO.}^{exp} - logk_{HO.}^{pred}\right|}{n} \tag{6}$$

$$\text{R2} = 1 - \frac{\sum\left(logk_{HO.}^{exp} - \overline{logk_{HO.}^{exp}}\right)^2}{\sum\left(logk_{HO.}^{exp} - logk_{HO.}^{pred}\right)^2} \tag{7}$$

where $logk_{HO.}^{exp}$, $logk_{HO.}^{pred}$, and $\overline{logk_{HO.}^{exp}}$ are the experimental, predicted and average experimental $\log k_{HO\bullet}$.

Table 1
The predictive performance of molecular image-CNN QSARs on the test datasets from 5 groups.

Groups	Model	Training			Validation			Test		
		MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
1	Model-R-1	0.09	0.12	0.96	0.24	0.31	0.75	0.22	0.33	0.60
2	Model-R-2	0.10	0.14	0.95	0.22	0.29	0.73	0.24	0.32	0.73
3	Model-R-3	0.08	0.14	0.95	0.20	0.28	0.72	0.23	0.34	0.68
4	Model-R-4	0.09	0.13	0.96	0.23	0.32	0.70	0.22	0.31	0.73
5	Model-R-5	0.10	0.15	0.95	0.20	0.28	0.76	0.20	0.28	0.67

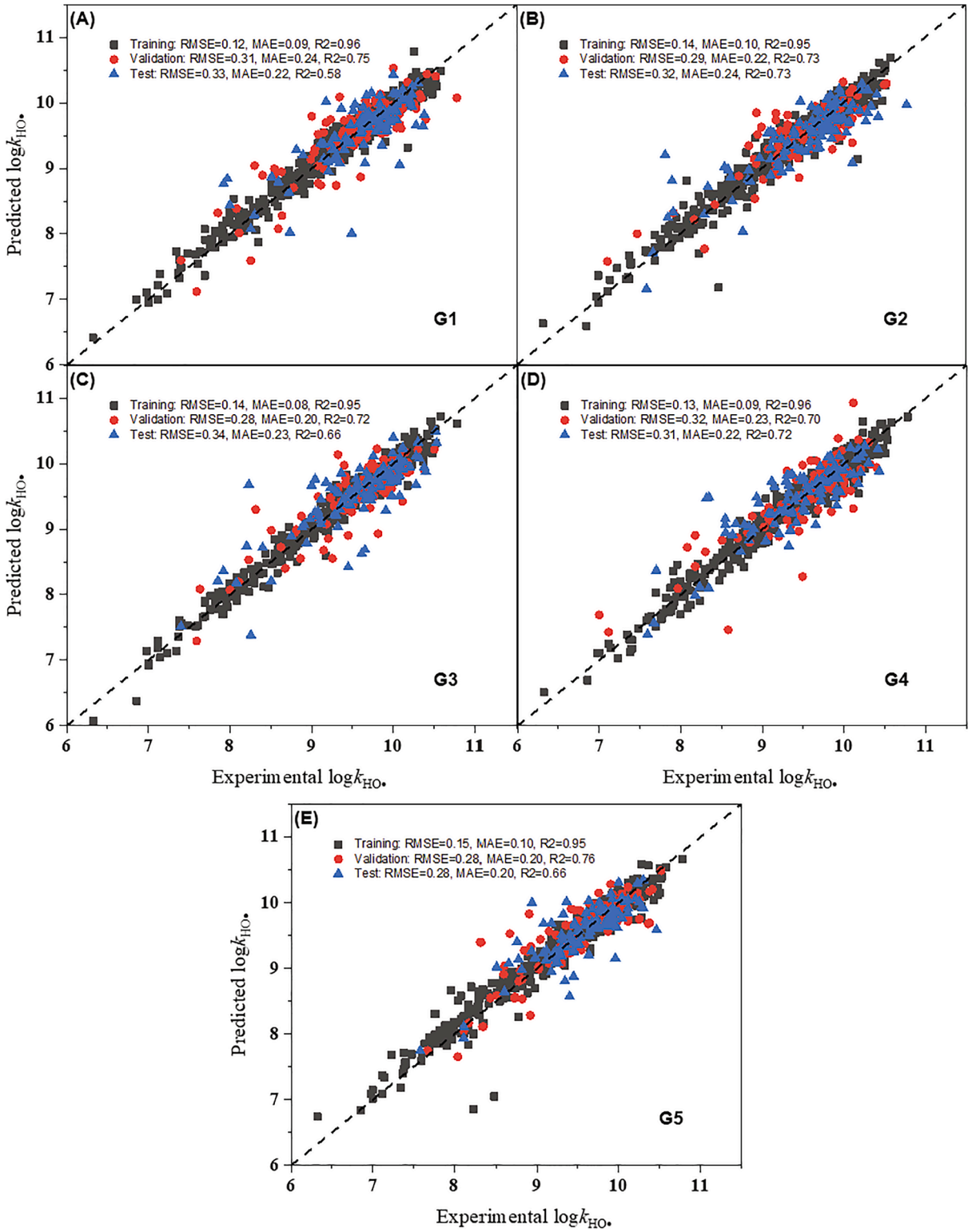

Fig. 3. The scatter plots of the experimental versus predicted log$k_{HO\bullet}$ by the molecular image-CNN models for the five groups: (A) G1, (B) G2, (C) G3, (D) G4, and (E) G5.

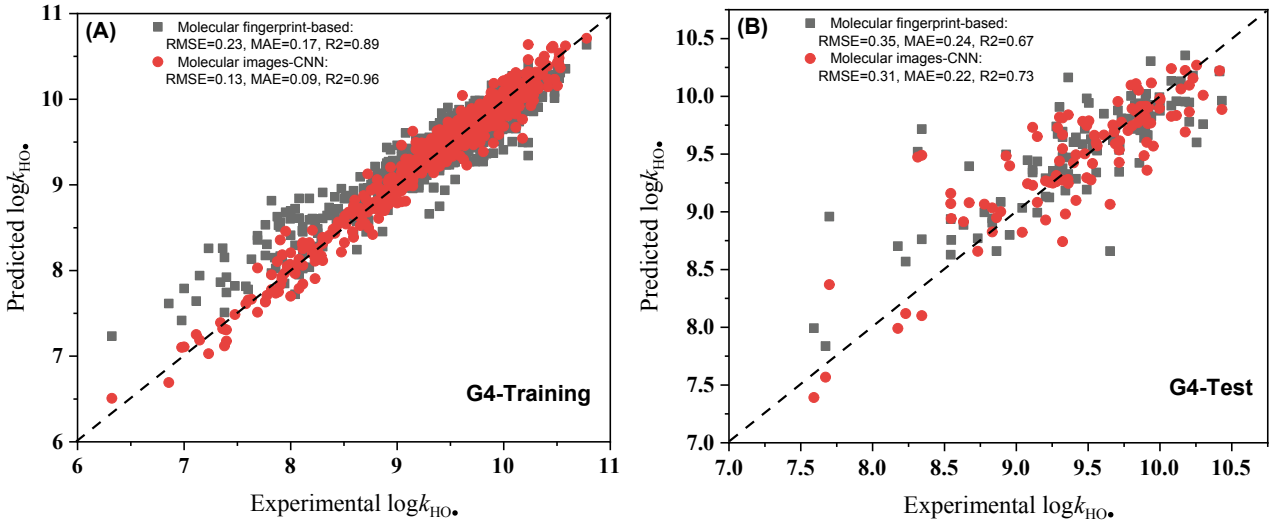

Fig. 4. The comparison of scatter plots of the experimental versus predicted log$k_{HO•}$ between the molecular fingerprint-based and molecular image-CNN models for group 4.

2.4. Applicability domain (AD)

AD is used to evaluate if a reliable prediction can be made for a new compound. It can be determined by comparing the similarity between a new compound and the compounds in the training dataset. The more similar the new compound is to the compounds in the training dataset, the more reliable the prediction is. Here, we compared the image of each compound in the test dataset with that of every compound in the training dataset one by one, and obtained sets of similarity values. Similar to our previous study [12], two similarity metrics were used: maximum similarity and mean similarity, which were calculated by taking the maximum and mean values from one set of similarity values, respectively. If the similarity of a compound in the test dataset was lower than the threshold value, this compound was determined to be outside the AD and then removed from the test dataset. $RMSE_{test}$ was then recalculated and the optimum threshold value was the one that led to the minimum recalculated $RMSE_{test}$ with the least number of compounds determined to be outside the AD. The similarity between two images is quantified by the Structural SIMilarity (SSIM) index [29]. The SSIM index can be viewed as a quantitative measure of one of the images as compared to that of another image that is regarded as of perfect quality. The range of SSIM is 0 to 1, in which the higher the SSIM index value, the more similar these two images are. A detailed description of this index can be found in the paper [29]. The similarity was obtained by the package of scikit-image with the command "*skimage.measure.compare_ssim(image1, image2, multichannel = True)*".

Table 2
The prediction accuracy, F1 scores and AUC-ROC values for functional group recognition on the test datasets from 5 groups.

Group	Model	Accuracy (%)	F1 score	AUC-ROC
1	Model-C-1	91.7	0.977	0.98
2	Model-C-2	95.4	0.987	0.99
3	Model-C-3	94.5	0.984	0.98
4	Model-C-4	96.3	0.992	0.99
5	Model-C-5	93.5	0.983	0.98

3. Results and discussion

3.1. Effect of transfer learning and data augmentation on the predictive performance of the models

Fig. 2 shows the comparison of predictive performance of models trained with/without applying data augmentation or transfer learning. Fig. 2A shows that without applying data augmentation the generalization ability of models on the test dataset became worse ($RMSE_{test}$: 0.284–0.339 vs. 0.395–0.45), although its predictive performance on the training dataset was similar to that with data augmentation applied ($RMSE_{train}$: 0.123–0.151 vs. 0.118–0.142). This is because without data augmentation the same atom groups at different positions in images cannot be well recognized by CNN [30]. By flipping or rotating images in data augmentation this issue can be largely mitigated. Fig. 2B shows that without transfer learning the obtained CNN models have poor performance even on the training dataset ($RMSE_{train}$: 0.452–0.592). This is because the CNN models without transfer learning had to "learn" from scratch. With transfer learning, the CNN had already acquired the ability to effectively extract key features from images, so the predictive performance of the models for training, validation and test datasets were all considerably improved.

3.2. The predictive performance of QSAR models obtained by molecular image-CNN with transfer learning and data augmentation versus by molecular fingerprints-based models

We trained CNN with transfer learning and data augmentation to develop molecular image-CNN QSAR models and listed the results in Table 1. All the 5 models for 5 groups showed similar predictive performance in terms of MAE, RMSE, and R^2 for training, validation and test datasets, indicating that the performance of the models was independent of data splitting. We previously used the same dataset and molecular fingerprints to develop QSAR models and compared them with recently reported molecular descriptor-based QSARs (details can be found in our previous paper) [12]. Table S2 lists a comparison between the molecular image-CNN model and other molecular descriptor-based and molecular fingerprint-based models. The molecular image-CNN models had $RMSE_{test}$ ranging from 0.28 to 0.34, which was slightly better than that of molecular fingerprint-based models ($RMSE_{test}$:

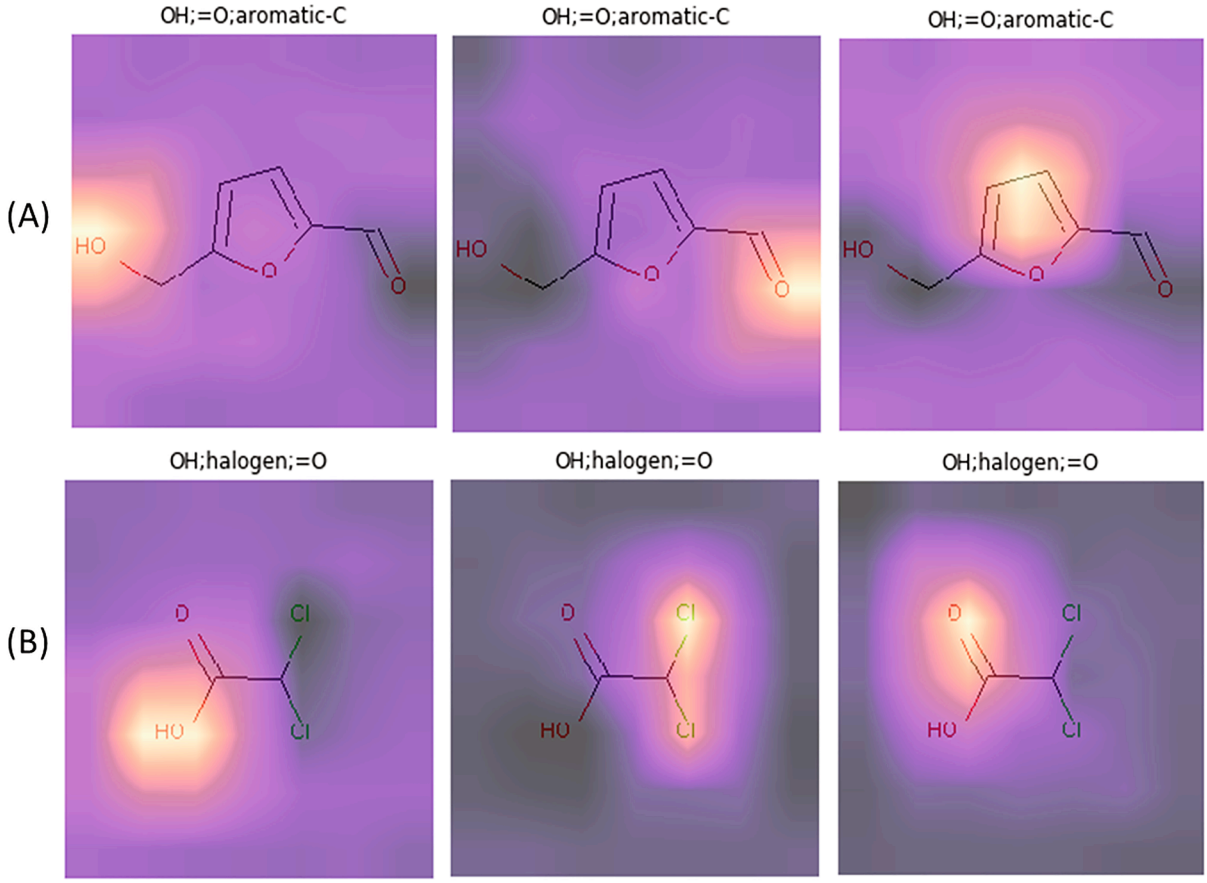

Fig. 5. The highlighted areas (heat maps) were obtained by Grad-CAM as the identified features for the corresponding functional groups in (A) 5-(hydroxymethyl) furan-2-carbaldhyde and (B) 2,2-dichloroacetic acid. The true labels are on the top of each image. The heat maps of all other compounds are provided in the SI (file name: *Classifiction_Grad_CAM_MIs.pdf)*.

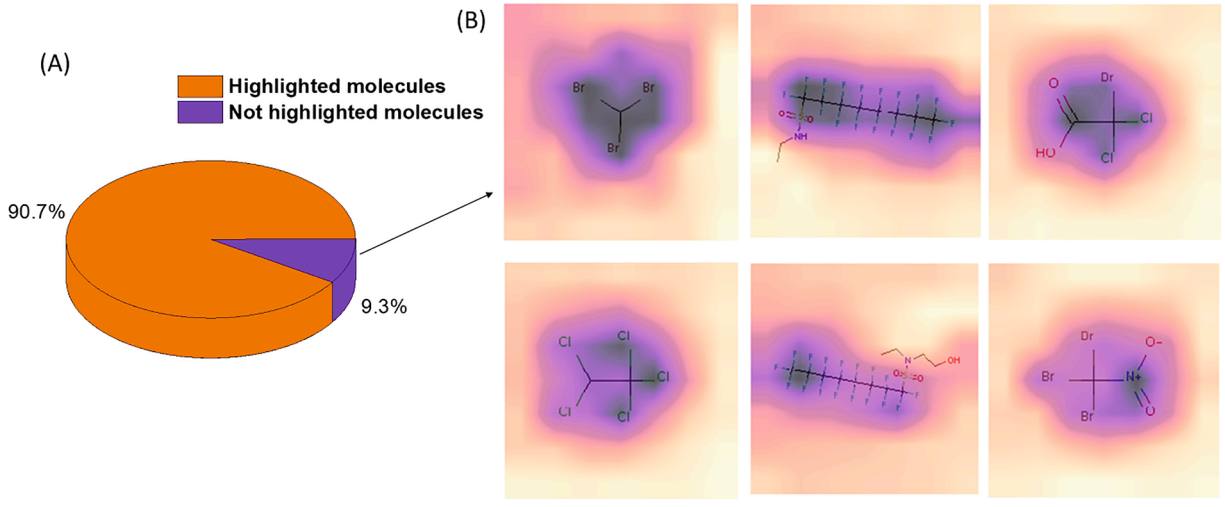

Fig. 6. (A) The percentages of images in which the molecules were or were not highlighted (in part or in whole) by Grad-CAM; (B) The heat maps of some compounds whose structures were not highlighted by Grad-CAM, in which the bright yellow color is the highlighted areas while the dark purple color is not the highlighted areas (The heat maps of all the compounds are in the SI with the file name *Regression MIs and Grad_CAM.pdf*). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

0.30–0.35) [12]. Fig. 3 shows the scatter plots of the experimental versus predicted $\log k_{HO\bullet}$ by the molecular image-CNN models. All the data points centered on the perfect fitting line (dotted line) with similar RMSE, MAE and R^2 for the training, validation and test datasets, respectively.

Our previous molecular fingerprint-based models showed particularly less accurate prediction for compounds with experimental $\log k_{HO\bullet}$ values below 9, as shown in Fig. 4 (black squares). This was because there is only a small number of compounds with experimental $\log k_{HO\bullet}$ values <9. Although the accuracy could be improved if we obtain more data points for compounds with $\log k_{HO\bullet}$ below 9, which is experimentally demanding, we can more easily improve the accuracy by applying data augmentation, as shown in Fig. 4 (red circles). After applying data augmentation, the molecular image-CNN models showed similar accurate predictions for compounds with experimental $\log k_{HO\bullet}$ below 9 and over 9.

3.3. Functional group recognition and Grad-CAM interpretation

The above results showed that the molecular image-CNN models had been well trained and ready to be interpreted. We first examined a classification application to test the reliability of Grad-CAM toward feature identification. Based on the molecular image-CNN models, we trained 5 molecular image-CNN classifiers for five random functional groups, namely, –OH, –CN, halogen, =O and aromatic carbon, in compounds from the same 5 groups as above. This classification test was conducted because it can easily tell if molecular image-CNN classifiers choose the most relevant parts of the molecular images to predict the correct functional groups.

Table 2 shows the predictive performance of the 5 molecular image-CNN classifiers on the test dataset in terms of accuracy, F1 score and AUC-ROC. The obtained high accuracy (>91.7%), F1 scores (>0.977) and AUC-ROC values (>0.98) for all the five classification models on the corresponding test datasets indicated that the performance of the models was independent of data splitting and the models had been well trained and ready to be interpreted. We then applied Grad-CAM to interpret one of the classification models (Model-C-4) by checking if the model had chosen the most relevant feature(s) in the images as the target functional groups. Fig. 5 shows two examples of the Grad-CAM results (the Grad-CAM results for all the compounds are supplied in the file named "Classifiction_Grad_CAM_MIs.pdf" in the SI), in which the highlighted areas represent the features, as unveiled by the Grad-CAM method, that the CNN relied on to classify as the target functional groups. There are three labeled functional groups (–OH, =O and aromatic-C) in 5-(hydroxymethyl) furan-2-carbaldhyde (Fig. 5A), all of which have been successfully identified by the model. Likewise, the three functional groups of –OH, halogen and =O in 2,2-dichloroacetic acid have also been chosen correctly (Fig. 5B). These results indicate that Grad-CAM is a reliable method to highlight what features of the images are chosen by CNN to make predictions.

3.4. Molecular image-CNN QSAR model interpretation

After validating the reliability of the Grad-CAM method, we can now interpret the molecular image-CNN QSAR models. Firstly, we should check if the models chose structural features rather than blank areas in the molecular images when making prediction because this should be a prerequisite for any meaningful prediction. Otherwise, we should not trust the model. Fig. 6A shows that for over 90% of the compounds their molecular images were highlighted in part or in whole, indicating that the molecular image-CNN models were at least not based on blank areas in the images to make prediction.

We then carefully visited the compounds whose structures were not highlighted in the images and found that, interestingly, all of them are recalcitrant to oxidation by HO• radicals, such as halogenated compounds (examples in Fig. 6B). For these compounds, electron-withdrawing groups largely decrease their electron density, and hence lower their reactivity toward HO• radicals which mainly attack electron-rich molecules. When developing QSARs, CNN darkened these groups so that other blank areas were highlighted instead. We thus can identify structural features in molecular images that can decrease $\log k_{HO\bullet}$ based on these darkened areas in the heat maps. Based on this rule, we found that the CNN can accurately identify groups that can reduce the reactivity (i.e., $\log k_{HO\bullet}$), as examples shown in Figure S5. The groups of –CN, $–NO_2$, $–SO_2$, –F, and –COOH are all well-known electron-withdrawing groups when attached to aromatic rings and can decrease the $\log k_{HO\bullet}$. Table 4 shows that 100% of common aromatic substitutes with negative electronic effects were correctly identified by the molecular image-CNN.

We next evaluate if the highlighted features by the CNN are linked to higher reactivity of the compounds containing these features. Table 5 lists 10 common classes of organic compounds; interestingly, the highlighted sites for all 10 classes except for nitriles are indeed all known to increase the reactivity. For nitriles, –CN is an electron-withdrawing group whose presence can decrease the $\log k_{HO\bullet}$ and, thus, should not be highlighted. The CNN highlighted the blank areas in the molecular images instead, which is correct in that the –CN is the least reactive feature on the images (less reactive than no structural features).

We have to recognize that molecular image-CNN models, similar to conventional molecular descriptor-based QSARs, cannot directly reflect reaction mechanisms between organic compounds and OH radicals, such as elementary reactions and rate-limiting steps. This is understandable because reaction mechanisms are not directly related to the reactivity. Rather, conventional QSARs rely on selected molecular descriptors to represent properties that would affect the reactivity, e.g., Hammett constants for electronic properties, whereas the obtained CNN model identifies key structural features that can enhance or inhibit the reaction toward the OH radical. Because the main purpose of QSAR models is to predict $\log k_{HO\bullet}$, correctly using the knowledge of how functional groups affect $\log k_{HO\bullet}$ is relevant to making predictions, as indeed correctly utilized by our CNN-model.

3.5. Applicability domain (AD)

We finally defined the applicability domain (AD) of our models to evaluate if a reliable prediction can be made for a given compound. A reliable prediction can be made if the new compound is structurally similar to the ones used in the training dataset [31–33]. Here, we used model-C-3 to determine its AD and compare it with that of the molecular fingerprint-based models in our previous study [12]. Table 6 shows that with increasing threshold values for both the maximum and mean similarity, more compounds in the test dataset were outside the AD, but the recalculated $RMSE_{test}$ first decreased and then increased. The threshold value determines if a new compound is outside the AD or not. For example, if the threshold value of the maximum similarity is set as 0.86, a new compound that has a maximum similarity of over 0.86 is inside the AD, otherwise, it is outside the AD. Based on the maximum similarity, a threshold of 0.85 led to a minimal $RMSE_{test}$ (0.334) with only one compound outside the AD, while based on the mean similarity a threshold of 0.81 yielded a minimal $RMSE_{test}$ (0.335) with two compounds outside the AD. Based on the principle that the optimum threshold value is the one that leads to the minimum $RMSE_{test}$ with the least number of compounds determined as outside the AD, the optimum similarity metric and threshold value are the maximum similarity and 0.85, respectively. We can thus conclude that if the maximum structural similarity of a given compound to the ones in the training dataset is higher than 0.85, our molecular image-CNN models can make a reliable prediction.

We also compared the AD of the molecular image-CNN models with that of the molecular fingerprint-based models in our previous study [12]. There were three compounds determined as outside the AD for the molecular fingerprint-based models [12], which is more than that of the

Table 4
Percentages of common aromatic substituents with negative electronic effects that have been correctly identified by molecular image-CNN.

Functional group	No. of compounds with the FG	No. of correctly identified compounds	Percentage (%)
$–NO_2$	36	36	100
–CN	23	23	100
–CHO	11	11	100
$–COCH_3$	9	9	100
$–CONH_2$	6	6	100
–CONHR	9	9	100
$–CONR_2$	1	1	100
–N=O	2	2	100
Halogen	266	266	100

Table 5
The known features of 10 classes of compounds that are beneficial to the reactivity and the highlighted features by the CNN.

Compound class	Compound No.[a]	Highlighted features
Alkane	41–46	$–CH_2–/–CH_3$
Aldehyde	0–5	$–CH_2–/–CH_3$
Nitrile	810–813	Blank area
Primary amine	285–302	$–CH_2–/–CH_3$
Cycloalkane	667–671	$–CH_2–$
Alkene	49–52	–C=C
Alcohol	850–853	$–CH_2–/–CH_3$
Thiol	1016–1018	–SH
Ether	736–751	$–CH_2–/–CH_3$
Thioether	973–981	–S–

[a] The indices of the compounds in the dataset (file name *dataset.xlsx* in the SI).

Table 6
The thresholds of similarity, the number of compounds outside the AD for each threshold value, and the corresponding $RMSE_{test}$.

Similarity metrics	Threshold value	# of Compounds outside the AD	Recalculated $RMSE_{test}$
Maximum similarity	0.84	0	0.340
	0.85	***1***	***0.334***
	0.86	2	0.335
	0.87	3	0.336
Mean similarity	0.80	0	0.340
	0.805	1	0.341
	0.81	2	0.335
	0.83	5	0.339
	0.84	12	0.347

molecular image-CNN models. This result indicated that the molecular image-CNN models can be applied to a slightly broader range of compounds than the molecular fingerprint-based models. This may result from the benefits of transfer learning and data augmentation. Fig. 7 shows the specific compounds that are outside the ADs for the molecular image-CNN and molecular fingerprint-based models. Different models excluded different compounds because they used different chemical representations so that the similarity comparison between compounds is different. Molecular image-CNN chose one compound with a more complex structure than those of the molecular fingerprint-based models.

4. Conclusions

This study developed molecular image-CNN models with techniques of transfer learning and data augmentation. These two techniques can largely enhance the robustness and predictive performance of models. The obtained models showed a better prediction performance than molecular fingerprint-based models. The Grad-CAM method demonstrated that the models correctly chose the relevant molecular features and identified the key functional groups in the molecular images, indicating that we can trust the molecular image-CNN-based models. Compared with molecular fingerprint-based models, the molecular image-CNN models had a broader AD, which can be applied to a wider range of compounds. This study not only offered a new, easy way to develop QSARs for environmental applications but also evaluated the trustworthiness of the models, which, as far as we know, should be a mandatory step when "black box" machine learning algorithms are employed. The application of this new modeling approach is not limited to HO• radicals but can be extended to any applications that involve many organic compounds. Given continuous needs for QSARs in various environmental applications and the fact that more and more contaminants may arise in the future [34], the molecular image-CNN-based approach will show great applications in the environmental field. Moreover, we encourage researchers targeting other environmental issues to be more creative in data analysis by, for example, considering converting their own data to images as the inputs for CNN so that the advantages of CNN, including transfer learning, data augmentation, and intuitive interpretation, can be used to generate more robust, easily interpretable models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

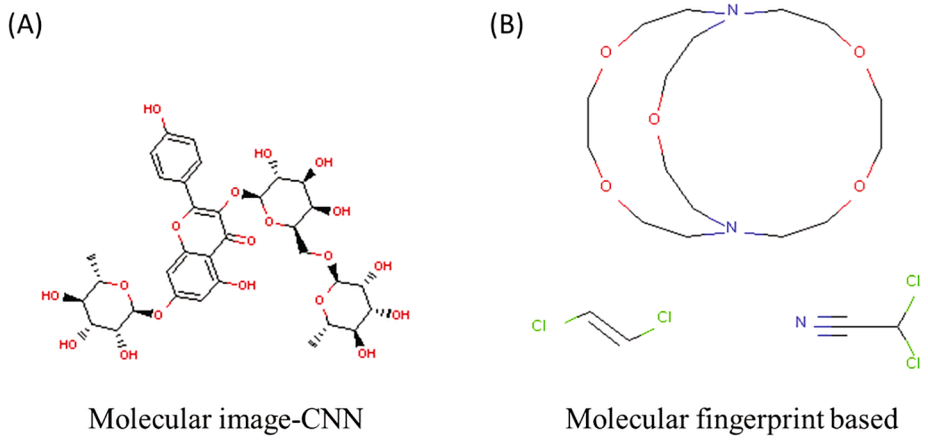

Fig. 7. The compounds that are determined as outside the AD for the molecular image-CNN models (A) and molecular fingerprint-based models (B).

Acknowledgment

This material is based upon work supported by the National Science Foundation under Grants CBET-1804708 and CHEM-1808406.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cej.2020.127998.

References

[1] T. Borhani, M. Saniedanesh, M. Bagheri, J. Lim, QSPR prediction of the hydroxyl radical rate constant of water contaminants, Water Res. 98 (2016) 344–353.
[2] H. Su, C. Yu, Y. Zhou, L. Gong, Q. Li, P. Alvarez, M. Long, Quantitative structure–activity relationship for the oxidation of aromatic organic contaminants in water by TAML/H2O2, Water Res. 140 (2018) 354–363.
[3] S.M. Free, J.W. Wilson, A mathematical contribution to structure-activity studies, J. Med. Chem. 7 (4) (1964) 395–399.
[4] Y. Lee, U. von Gunten, Quantitative structure–activity relationships (QSARs) for the transformation of organic micropollutants during oxidative water treatment, Water Res. 46 (19) (2012) 6177–6195.
[5] S. Sudhakaran, G.L. Amy, 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) (2013) 1111–1122.
[6] T. Ye, Z. Wei, R. Spinney, D.D. Dionysiou, S. Luo, L. Chai, Z. Yang, R. Xiao, Quantitative structure–activity relationship for the apparent rate constants of aromatic contaminants oxidized by ferrate (VI), Chem. Eng. J. 317 (2017) 258–266.
[7] Z. Cheng, B. Yang, Q. Chen, Z. Shen, T. Yuan, 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 (2017) 534–540.
[8] R. Xiao, T. Ye, Z. Wei, S. Luo, Z. Yang, R. Spinney, Quantitative structure–activity relationship (QSAR) for the oxidation of trace organic contaminants by sulfate radical, Environ. Sci. Technol. 49 (22) (2015) 13394–13402.
[9] S. Luo, Z. Wei, R. Spinney, F.A. Villamena, D.D. Dionysiou, D. Chen, C.-J. Tang, L. Chai, R. Xiao, Quantitative structure–activity relationships for reactivities of sulfate and hydroxyl radicals with aromatic contaminants through single–electron transfer pathway, J. Hazard. Mater. 344 (2018) 1165–1173.
[10] C. Li, S. Zheng, T. Li, J. Chen, J. Zhou, L. Su, Y.-N. Zhang, J.C. Crittenden, D. Wang, S. Zhu, Y. Zhao, Quantitative Structure-Activity Relationship Models for Predicting Reaction Rate Constants of Organic Contaminants with Hydrated Electrons and Their Mechanistic Pathways, Water Res. 151 (2018) 468–477.
[11] S. Zhong, J. Hu, X. Fan, X. Yu, H. Zhang, A deep neural network combined with molecular fingerprints (DNN-MF) to develop predictive models for hydroxyl radical rate constants of water contaminants, J. Hazard. Mater. 383 (2020), 121141.
[12] S. Zhong, K. Zhang, D. Wang, H. Zhang, Shedding Light On “Black Box” Machine Learning Models for Predicting the Reactivity of HO• Radicals toward Organic Compounds, Chem. Eng. J. 126627 (2020).
[13] D. Minakata, K. Li, P. Westerhoff, J. Crittenden, Development of a Group Contribution Method To Predict Aqueous Phase Hydroxyl Radical (HO•) Reaction Rate Constants, Environ. Sci. Technol. 43 (16) (2009) 6220–6227.
[14] A. Monod, J. Doussin, Structure-activity relationship for the estimation of OH-oxidation rate constants of aliphatic organic compounds in the aqueous phase: alkanes, alcohols, organic acids and bases, Atmos. Environ. 42 (33) (2008) 7611–7622.
[15] E.S. Kwok, R. Atkinson, Estimation of hydroxyl radical reaction rate constants for gas-phase organic compounds using a structure-reactivity relationship: an update, Atmos. Environ. 29 (14) (1995) 1685–1695.
[16] Goh, G. B.; Siegel, C.; Vishnu, A.; Hodas, N. O.; Baker, N., Chemception: a deep neural network with minimal chemistry knowledge matches the performance of expert-developed QSAR/QSPR models. arXiv preprint arXiv:1706.06689 2017.
[17] M. Fernandez, F. Ban, G. Woo, M. Hsing, T. Yamazaki, E. LeBlanc, P.S. Rennie, W. J. Welch, A. Cherkasov, Toxic colors: the use of deep learning for predicting toxicity of compounds merely from their graphic images, J. Chem. Inf. Model. 58 (8) (2018) 1533–1543.
[18] T. Shi, Y. Yang, S. Huang, L. Chen, Z. Kuang, Y. Heng, H. Mei, Molecular image-based convolutional neural network for the prediction of ADMET properties, Chemometrics Intellig. Lab. Syst. 194 (2019), 103853.
[19] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2009) 1345–1359.
[20] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (1) (2019) 60.
[21] J.W. Deng, R. Dong, L.-J. Socher, K. Li, F.-F. Li, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, 2009, Ieee, 2009, pp. 248–255.
[22] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, Adv. Neural Inform. Process. Syst. 2018 (2018) 9505–9515.
[23] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision 2017, 2017, pp. 618–626.
[24] J. Hoigné, Inter-calibration of OH radical sources and water quality parameters, Water Sci. Technol. 35 (4) (1997) 1–8.
[25] M. Anbar, D. Meyerstein, P. Neta, The reactivity of aromatic compounds toward hydroxyl radicals, J. Phys. Chem. 70 (8) (1966) 2660–2662.
[26] D. Weininger, SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, J. Chem. Inf. Comput. Sci. 28 (1) (1988) 31–36.
[27] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, 2017,, pp. 4700–4708.
[28] X. Luo, X. Wei, J. Chen, Q. Xie, X. Yang, W.J.G.M. Peijnenburg, Rate constants of hydroxyl radicals reaction with different dissociation species of fluoroquinolones and sulfonamides: Combined experimental and QSAR studies, Water Res. 166 (2019), 115083.
[29] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[30] A. Azulay, Y. Weiss, Why do deep convolutional networks generalize so poorly to small image transformations? arXiv preprint arXiv:1805.12177 2018.
[31] D. Gadaleta, G.F. Mangiatordi, M. Catto, A. Carotti, O. Nicolotti, Applicability domain for QSAR models: where theory meets reality, Int. J. Quantitative Structure-Property Relationships (IJQSPR) 1 (1) (2016) 45–63.
[32] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of different approaches to define the applicability domain of QSAR models, Molecules 17 (5) (2012) 4791–4810.
[33] I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, A. Varnek, Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection, J. Chem. Inf. Model. 48 (9) (2008) 1733–1746.
[34] T.M. Nolte, A.M. Ragas, A review of quantitative structure–property relationships for the fate of ionizable organic chemicals in water matrices and identification of knowledge gaps, Environ. Sci. Processes Impacts 19 (3) (2017) 221–246.