Consensus Hologram QSAR Model Studying on the Aqueous Hydroxyl Radical Oxidation Reaction Rate Constants of Organic Micropollutants①

JIAO Long[a②] LEI Bin[a] QU Le[b]
LI Rui[b, c] YAN Chun-Hua[a] LI Hong[a]

[a] (*College of Chemistry and Chemical Engineering, Xi'an Shiyou University, Xi'an* 710065, *China*)

[b] (*Shaanxi Cooperative Innovation Center of Unconventional Oil and Gas Exploration and Development, Xi'an Shiyou University, Xi'an* 710065, *China*)

[c] (*College of Petroleum Engineering, Xi'an Shiyou University, Xi'an* 710065, *China*)

ABSTRACT The combination of hologram quantitative structure-activity relationship (HQSAR) and consensus modeling was employed to study the quantitative structure-property relationship (QSPR) model for calculating the aqueous hydroxyl radical oxidation reaction rate constants (k_{OH}) of organic micropollutants (OMPs). Firstly, individual HQSAR model were established by using standard HQSAR method. The optimal individual HQSAR model was obtained while setting the parameter of *fragment distinction* and *fragment size* to "B" and "3~6" respectively. Secondly, consensus HQSAR model was established by building the regression model between the k_{OH} and the hologram descriptors with consensus partial least-squares (cPLS) approach. The obtained individual and consensus HQSAR model were validated with a randomly selected external test set. The result of external test set validation demonstrates that both individual and consensus HQSAR model are available for predicting the k_{OH} of OMPs. Compared with the optimal individual HQSAR model, the established consensus HQSAR model shows higher prediction accuracy and robustness. It is shown that the combination of HQSAR and consensus modeling is a practicable and promising method for studying and predicting the k_{OH} of OMPs.

Keywords: QSPR, hologram QSAR, consensus modeling, organic micropollutants, hydroxyl radical, rate constant; DOI: 10.14102/j.cnki.0254–5861.2011–3083

1 INTRODUCTION

Hologram quantitative structure-activity relationship (HQSAR) is an ingenious and efficient quantitative structure-property relationship (QSPR) technique, which proposes a specialized fragment fingerprint, known as molecular hologram (MH), as the structural descriptor to build a QSPR model. Because the HQSAR model is easier-to-built than many other QSPR models, as well as possessing comparable prediction accuracy, it has been successfully applied to a number of QSPR researches in the fields of biology[1, 2], pharmacology[3-6], chemistry[7, 8], environmental science[9, 10], etc. Traditionally, HQSAR method builds individual regression models between molecular properties and hologram descriptors by using partial least-squares (PLS) regression. As is known to all, individual regression models tend to underfitting or overfitting[11]. By contrast, consensus modeling method can overcome this shortcoming to a great extent through integrating several individual models[12-14]. The predictive accuracy and robustness of regression models could be improved by consensus modeling. As a powerful and reliable modeling strategy, consensus modeling has been successfully applied to lots of research fields, such as QSPR

Received 30 December 2020; accepted 20 May 2021

① This research was supported by the National Natural Science Foundation of China (No. 21775118), Shaanxi Natural Science Basic Research Project (No. 2018JM2018), Youth Innovation Team of Shaanxi Universities (No. 2019.21), Young Outstanding Talent Support Program of Shaanxi Universities, Xi'an Shiyou University Youth Research and Innovation Team Construction Plan (No. 2019QNKYCXTD17), and Xi'an Shiyou University Graduate Innovation and Practice Ability Training Project (No. YCS19211016)

② Corresponding author. E-mail: mop@xsyu.edu.cn

modeling, spectral analysis, machine learning, artificial intelligence and so on[15-17]. Obviously, it is necessary and advisable to introduce consensus modeling into HQSAR modeling in order to build more accurate and robust models.

Organic micropollutants (OMPs) as a group of compounds that cover a wide array of physical-chemical properties have been identified as emerging contaminants due to the possible threats to ecological environments. In recent decades, contamination of OMPs on surface water has received increasingly scientific and public awareness[18]. OMPs has the characteristics of low concentration and high toxicity, which can cause direct or potential harm to aquatic ecosystems and human health. Therefore, with the progress of science and the enhancing attention about human health, technology is needed to remove these pollutants from wastewater effluents prior to the discharge of wastewater to the environment. Recently, ozone has been used to process OMPs in wastewater. According to Hoigne and Bader[19], there are two ways for ozone reacting with organic pollutants in water: (1) direct reactions; (2) indirect reactions of hydroxyl radicals produced by the process of ozone decomposition. The rate constants of direct reaction could be easily determined by experiments[20]. However, due to the complexity of analytical methods, experimentally determining the aqueous oxidation reaction rate constants of hydroxyl radical with OMPs is always a time-consuming, costly and hard task[21-25]. Hence, the QSPR method has been extensively used to predict the hydroxyl radical rate constant of the contaminants by relating the properties (rate constant) of contaminants with their molecular structures[26-30]. Several 2*D*-QSPR models have been proposed in many literatures for studying the rate constant of hydroxyl radical on the basis of quantum chemical or topological descriptors[31-33]. However, the modeling processes of these models are always time-consuming and complex, and it is always meaningful to improve the accuracy and robustness of these models. Thus, the QSPR model of the aqueous hydroxyl radical oxidation reaction rate constant of OMPs (k_{OH}) was studied in this work, based on the HQSAR and consensus modeling method.

2 EXPERIMENTAL

2. 1 Data set and software

The experimental aqueous hydroxyl radical oxidation reaction rate constants of the investigated 83 OMPs was collected from reference[34]. The 83 OMPs were randomly divided into two sample sets, training set and test set, in the light of 2:1. The training set, which was used to establish and optimize the HQSAR model, includes 55 samples. The test set, which was utilized to assess the prediction performance of the developed QSPR models, of course comprises the other 28 samples.

All the computations were carried out in an *i*5-4258U/4G-RAM personal computer. The computations related to HQSAR modeling were performed in SYBYL-X 2.0 software (Certara, U.S.). Other computations were performed with the program developed by our research team.

2. 2 Model assessment

Several statistical indices, including root mean square error (*RMSE*), squared correlation coefficient of cross validation (q_{cv}^2), squared correlation coefficient of external (q_{ext}^2), concordance correlation coefficient (*CCC*), predictive squared correlation coefficient (Q_{F2}^2 and Q_{F3}^2), $\bar{r}_m^2$ and Δr_m^2 [35-39], were jointly used to assess the prediction performance of the generated models.

Eqs. 1 and 2 show the definition of *CCC*, Q_{F2}^2 and Q_{F3}^2 :

$$CCC = \frac{2\sum_{i=1}^{n_{EXT}}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sum_{i=1}^{n_{EXT}}(y_i-\bar{y})^2+\sum_{i=1}^{n_{EXT}}(\hat{y}_i-\bar{\hat{y}})^2+n_{EXT}(\bar{y}-\bar{\hat{y}})^2} \quad (1)$$

$$Q_{F2}^2 = 1-\frac{\sum_{i=1}^{n_{EXT}}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n_{EXT}}(y_i-\bar{y}_{EXT})^2} \quad (2a)$$

$$Q_{F3}^2 = 1-\frac{\left[\sum_{i=1}^{n_{EXT}}(\hat{y}_i-y_i)^2\right]/n_{EXT}}{\left[\sum_{i=1}^{n_{TR}}(y_i-\bar{y}_i)^2\right]/n_{TR}} = 1-\frac{PRESS/n_{EXT}}{TSS/n_{TR}} \quad (2b)$$

In these equations, y_i represents the experimental value of each sample, $\hat{y}_i$ denotes the predicted value of each sample, $\bar{y}$ means the average of all experimental values, $\bar{\hat{y}}$ is the average of all predicted values, n_{TR} is described as the number of samples in training set, n_{EXT} stands for the number of samples in external test set, *PRESS* denotes the predictive error sum of squares, and *TSS* is regarded as the sum of squares of prediction errors of all the samples. Chirico et al.[36-39] suggested that for an acceptable QSPR model, the value of *CCC* should be higher than 0.85, Q_{F2}^2 and Q_{F3}^2 higher than 0.70, q_{cv}^2 and q_{ext}^2 higher than 0.50.

The definition of $\bar{r}_m^2$ and Δr_m^2 is shown in Eq. 3:

$$r_m^2 = r^2(1-\sqrt{r^2-r_0^2}) \quad (3a)$$

$$r'^2_m = r'^2(1-\sqrt{r'^2-r'^2_0}) \quad (3b)$$

$$\bar{r}_m^2 = \frac{r_m^2 + r_m'^2}{2} \quad (3c)$$

$$\Delta r_m^2 = \left| r_m^2 - r_m'^2 \right| \quad (3d)$$

Here, r^2 and r_0^2 stand for the determination coefficients of the regression equation between the predicted and experimental values of test set, compelling respectively the origin of the axis (r_0^2) or not (r^2). When calculating (r_m^2, r^2, r_0^2) and ($r_m'^2$, r'^2, $r_0'^2$), the experimental values were used as ordinate and abscissa, respectively. Roy et al.[40] suggested that the value of $\bar{r}_m^2$ should be higher than 0.65 and Δr_m^2 should be lower than 0.2, for an acceptable QSPR model.

2. 3 HQSAR

HQSAR is an excellent 2.5*D*-QSPR approach proposed by Hurst et al.[41, 42] which contains the advantages of both 2*D*-QSPR methods and 3*D*-QSPR methods. The notable advantage of HQSAR is it can rapidly and automatically process large data set with high prediction accuracy and statistical quality. Compared with 3*D*-QSPR methods, conformation optimization or alignment of molecules is not required in HQSAR. HQSAR is an ingenious and successful combination of molecular hologram descriptors and PLS regression methods.

MH is an extended form of molecular fingerprint, a kind of fragment-based descriptor which translates chemical structure representations into binary bit strings. It can code more structural information than traditional 2*D* molecular fingerprint, such as stereo-chemical structure, branching and cyclic fragments. All possible molecular fragments, including linear, branched, cyclic, and overlapping features within a molecule, could be contained in MH. MH is actually an array containing counts of molecular fragments. In MH, the molecular fragments are described with Sybyl Line Notation (SLN), a specification for explicitly characterizing molecular fragments, structures, structural libraries, reactions, formulations, molecular and reaction queries by using short ASCII strings.

Two parameters, *fragment distinction* and *fragment size*, are used to set the type and length of MH descriptors. The parameter of *fragment distinction* defines the type of fragments, including atoms (A), bonds (B), connections (C), hydrogen atoms (H), chirality (Ch), and donor and acceptor atoms (DA)[43, 44]. Different types of fragments could be combined. For example, the default setting of *fragment distinction* is "A/B/C" in SYBYL. The parameter of *fragment size* is used to specify the length of fragments. All the possible fragments are generated with *S* atoms[45, 46]. Here, *S* is an integer between *M* and *N*. The value of *M* should be larger than 2 and smaller than *N*. The values of *N* is usually larger than 12 and does not exceed the number of atoms in the molecule. This parameter is set to "4~7" by default in SYBYL. After setting the parameter of *fragment distinction* and *fragment size*, each fragment was assigned to a unique integer in the range of 0～231 using a cyclic redundancy check (CRC) algorithm[47]. Each integer corresponds to a bin in an integer array of fixed length *L*, which represents the length of MH. In the HQSAR module of SYBYL software, *L* usually is one of the 12 prime numbers ranging from 53 to 401. The initial setting of *L* is 97, 151, 199, 257, 307 and 353. The terms of molecular bit string fingerprint involve "0", which usually does not have any useful information. In the subsequent PLS modeling step, the computation time may be dramatically increased with the increase of fingerprint length. More importantly, these null values may hinder the follow-up computation of PLS model. Therefore, it is necessary to adopt effective method for reducing the length of fingerprint. This reduction is achieved through the process called "hashing", which allocates multiple fragments to the same location in a fingerprint[48].

In general, HQSAR method consists of three main steps: (1) generating sub-structure fragments of each molecule in the data set; (2) encoding these fragments with MH descriptors; (3) establishing the quantitative relationship model between the MH descriptors and the properties of compounds by using PLS. The PLS model with the highest value of q_{cv}^2 is usually considered as the established best HQSAR model.

2. 4 Consensus modeling

The idea behind consensus modeling is building a series of models, namely member models, with different training subsets, which consists of different samples randomly selected from one training set, and combining the eligible member models according to the consensus rules. A consensus model always contains member models with different prediction characteristics. The most significant advantage of consensus modeling is that it is able to resist underfitting and overfitting to a certain extent, and thus can improve the robustness and predictability of a regression model. The flow chart of consensus modeling is shown in Fig. 1.

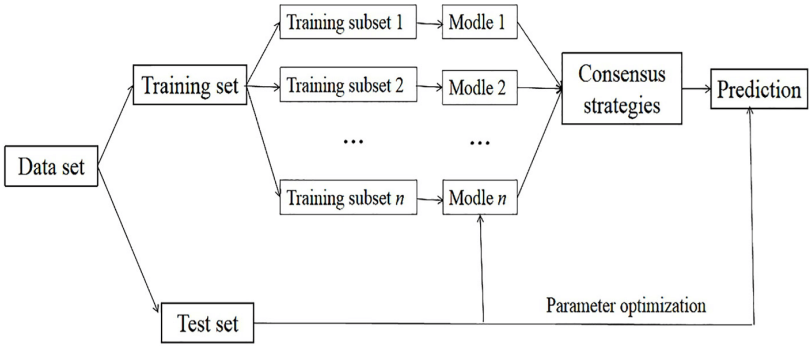


Fig. 1. Flow chart of consensus modeling

Krogh and Vedelsby[49] proposed the prediction error decomposition theory of consensus models and expressed the theory as follows:

$$E(\bar{x}) = \bar{E}(\bar{x}) - \bar{A}(\bar{x}) \quad (4a)$$

$$\bar{E}(\bar{x}) = \frac{1}{N_m}\sum_{i=1}^{N_m}(\hat{y}_i - y)^2 \quad (4b)$$

$$\bar{A}(\bar{x}) = \frac{1}{N_m}\sum_{i=1}^{N_m}(\hat{y}_i - \hat{y})^2 \quad (4c)$$

$$\hat{y} = \frac{1}{N_m}\sum_{i=1}^{N_m}\hat{y}_i \quad (5)$$

In these equations, $\bar{E}(\bar{x})$ denotes the average error of all the member models, $\bar{A}(\bar{x})$ means the variance of the member models relative to the consensus model, N_m is the total number of member models, y represents the dependent variable (namely the property value) of each sample, $\hat{y}_i$ and $\hat{y}$ indicate the predicted value of member models and consensus model respectively, and $\bar{x}$ represents the independent variables (namely the structural descriptor in QSAR models) of each sample. The consensus rule used in this study is arithmetic mean that means the predicted values of the N_m member models are averaged, as shown in Eq. 5.

According to Eq. 4a, two factors is concerned to the prediction error of a consensus model: $\bar{E}(\bar{x})$, the average relative error of all member models and $\bar{A}(\bar{x})$, the variance of the member models relative to the consensus model. $\bar{E}(\bar{x})$ reflects the prediction accuracy of the model, and $\bar{A}(\bar{x})$ shows the correlation between the member models. An excellent consensus model should comprise member models with high prediction accuracy (small prediction error, $\bar{E}(\bar{x})$) and uncorrelated member models (large prediction variance, $\bar{A}(\bar{x})$). Raising the accuracy and robustness of a consensus model need to decrease the difference between $\bar{E}(\bar{x})$ and $\bar{A}(\bar{x})$. The best model will be obtained when the value of $\bar{E}(\bar{x})$ is very close to that of $\bar{A}(\bar{x})$. Therefore, choosing appropriate member models is of great importance in consensus modeling.

Consensus partial least squares (cPLS) is a commonly used consensus modeling method[12, 14, 50, 51]. Its basic idea is disturbing the training set by random sampling, establishing a series of individual PLS models, and selecting appropriate member models from these individual PLS models to jointly predict the unknown samples. The main steps of cPLS includes:

(1) Setting the training subset and inspection set;

(2) Setting the total number of individual PLS models;

(3) Building the individual PLS models with the training subset, and predicting the inspection set with the obtained individual PLS models;

(4) Determining whether the individual PLS model established in step (3) could be accepted as member model of the consensus model, according to the prediction results of inspection set;

(5) Repeating steps (2)～(4) to find enough eligible member models;

(6) Combining the prediction result of all the member models according to the fusion criteria, such as calculating the mean value, to build the cPLS model.

3 RESULTS AND DISCUSSION

3. 1 Individual HQSAR model

The two key parameters, *fragment distinction* and *fragment size*, were optimized in order to build acceptable HQSAR models. The first step is optimizing the *fragment distinction* parameter. A series of candidate HQSAR models were developed by setting *fragment distinction* to different values at the default *fragment size* of "4~7". Training set was used to establish these candidate models and the statistics of

these models were calculated. The key statistics of the best eight models are listed in Table 1. The optimum model, of which the q_{cv}^2 R^2 and standard error value are 0.816, 0.875 and 1.754, respectively, was obtained by setting *fragment distinction* to "B". Thus, we set *fragment distinction* to "B" in the followed models. Secondly, the parameter of *fragment size* was optimized. A series of HQSAR models were built by setting *fragment size* to different values, while setting the *fragment distinction* to "B". These models were still built up by using the 55 samples of training set. The key statistics of the best eight models were calculated and listed in Table 2. The optimum model was obtained by setting *fragment size* to "3~6", and the statistical parameters q_{cv}^2, R^2 and standard error are 0.854, 0.915 and 1.451, respectively. This is the optimal individual HQSAR model of k_{OH}. As shown in Table 2, the optimal individual HQSAR model of k_{OH} could be built when setting *fragment distinction*, *fragment size*, *fragment length* and *principal components* to "B", "3~6", "151" and "6" respectively.

Table 1. Statistics of the k_{OH} Models with Different *Fragment Distinctions*

No.	*Fragment distinction**	q_{cv}^2	R^2	Standard error	Best length	PCs
1	**B**	**0.816**	**0.875**	**1.754**	**199**	**6**
2	A, B, H, CH, DA	0.743	0.846	1.912	307	4
3	A, B, H, DA	0.722	0.846	1.911	199	4
4	C, DA	0.716	0.900	1.570	151	6
5	A, C, DA	0.699	0.870	1.775	151	5
6	B, C	0.697	0.802	2.164	97	4
7	A, CH, DA	0.636	0.862	1.844	151	6
8	A	0.630	0.822	2.077	307	5

* In this column, A, B, C, H, Ch and DA respectively represent "atoms", "bonds", "connections", "hydrogen atoms", "chirality" and "donor and acceptor atoms".

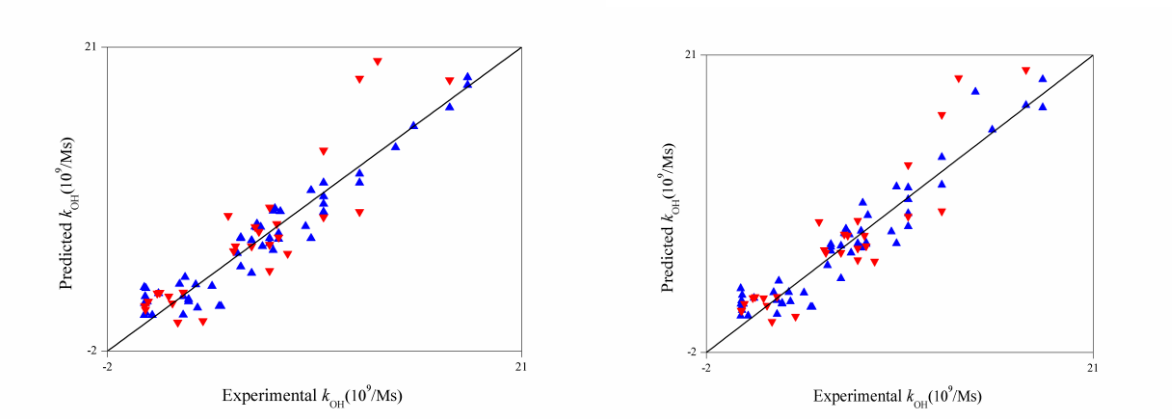
Table 2. Statistical Results of the k_{OH} Models with Different *Fragment Sizes*

No.	*Fragment distinction*	*Fragment size*	q_{cv}^2	R^2	Standard error	Best length	PCs
1	**B**	**3~6**	**0.854**	**0.915**	**1.451**	**151**	**6**
2	B	4~5	0.821	0.909	1.498	151	6
3	B	4~7	0.816	0.875	1.754	199	6
4	B	2~5	0.812	0.903	1.531	97	5
5	B	5~6	0.794	0.880	1.720	199	6
6	B	1~4	0.788	0.894	1.618	97	6
7	B	3~4	0.784	0.898	1.585	97	6
8	B	5~8	0.775	0.886	1.881	199	6

In order to evaluate the prediction ability of this model, external test set verification was carried out. The k_{OH} of the samples in test set was predicted by this model. The predicted k_{OH} was shown in Fig. 2, the corresponding statistical indices were calculated and listed in Table 3. The results of external test set validation show that the established individual HQSAR model is acceptable. That is to say, the hologram information is quantitatively relate to the k_{OH} value of the 83 OMPs. The developed HQSAR model is practicable for predicting the k_{OH} value of organic micropollutants. However, the q_{ext}^2 and Q_{F2}^2 of this model just reaches 0.659 and 0.657 respectively, building more accurate QSPR models is still meaningful.

Table 3. Statistics of the Individual and Consensus HQSAR Model

Parameters	Individual HQSAR model	Consensus HQSAR model
RMSEP	2.531	2.181
CCC	0.865	0.895
q^2_{ext}	0.659	0.747
Q^2_{F2}	0.657	0.746
Q^2_{F3}	0.703	0.779
$\bar{r}^2_m$	0.727	0.780
Δr^2_m	0.136	0.114



(a) **(b)**

Fig. 2. Predicted k_{OH} versus experimental k_{OH}: (a) individual HQSAR model, (b) consensus HQSAR model. "▲" indicates the samples of training set, and "▼" are those of the test set

3. 2 Consensus HQSAR model

In this section, cPLS was employed to build consensus HQSAR model. Correspondingly, the 55 samples of training set was randomly divided into a training subset and an inspection subset, according to the ratio of 2:1. The training subset, which was used to build the regression model, comprises of 37 samples and the inspection subset, which was used to optimize the number of member models, includes the rest 18 samples.

The consensus model was built with cPLS regression method. All the member HQSAR models of the consensus model were established by using the training subset as the calibration set of PLS, while setting the *fragment distinction* and *fragment size* to "B" and "3~6" respectively. Generally, $q^2_{cv} > 0.80$ is commonly used as the acceptance standard of member models in cPLS, by considering the diversity and prediction ability of the member models. Thus, we chose the PLS regression model whose q^2_{cv} is larger than 0.80 as the member models. The cPLS method should be built by proper member models, which have positive effects on the model stability and accuracy. It is necessary to optimize the number of member models of a cPLS model. A series of cPLS models were built with different number and combination of the member PLS models, of which q^2_{cv} is greater than 0.80. The k_{OH} of the samples in inspection set was then predicted with these cPLS models and the *RMSE* for the inspection set was calculated. The prediction result demonstrates that when the number of models less than 60, the *RMSE* of the inspection set is unstable and continuously decreasing; when the number of member models is larger than 60, the *RMSE* of the inspection set has tended to stable and decreased to 2.03. Consequently, the consensus HQSAR model was established with 60 member models. Namely, the predicted k_{OH} of OMPs of the developed cPLS model is actually the average value of the 60 selected member models. Then, the k_{OH} of the samples in test set were predicted to assess the prediction ability of the consensus model. The prediction result is shown in Fig. 2 and the corresponding statistical indices are listed Table 3. Fig. 2 and the indices listed in Table 3 demonstrate that this model is more accurate and robust than the individual HQSAR method. It is demonstrated that the QSPR model for the k_{OH} of OMPs can be established through HQSAR method, and the robustness

and predictability of the model can be improved through the consensus modeling.

4 CONCLUSION

The QSPR models for predicting the aqueous oxidation reaction rate constants of organic micropollutants with hydroxyl radical were successfully established by using HQSAR approach combined with consensus modeling method. The result of external test set validation indicates that both individual HQSAR model and consensus HQSAR model is practicable for describing the quantitative relationship between the structural information and k_{OH} of the investigated organic micropollutants. Compared with individual HQSAR model, the established consensus HQSAR model has higher prediction accuracy and robustness. It is demonstrated that consensus HQSAR modeling is a practicable and promising approach for improving the accuracy and robustness of HQSAR model. And the established consensus HQSAR model is an easy-to-use and accurate model for studying and predicting the aqueous k_{OH} of the organic micropollutants oxidation reactions.

REFERENCES

(1) Tong, J. B.; Zhan, P.; Wang, X. S.; Wu, Y. J. Quionolone carboxylic acid derivatives as HIV-1 integrase inhibitors: docking-based HQSAR and topomer CoMFA analyses. *J. Chemometr*. **2017**, 31, e2934–13.

(2) Sun, J. Y.; He, Y. Q.; Du, H. R.; Liu, C. L.; Chen, A. Y.; Mei, H. In vitro anti-viral activities and structure-activity relationship studies of flavones and dihydroflavone derivatives as influenza virus potential neuraminidase inhibitors. *Chin. J. Struct. Chem.* **2015**, 34, 1641–1651.

(3) Ver ś simo, G. C.; Dutra, E. F. M.; Dias, A. L. T.; Fernandes, P. de. O.; Kronenberger, T.; Gomes, M. A.; Maltarollo, V. G. HQSAR and random forest-based QSAR models for anti-T. Vaginalis activities of nitroimidazoles derivatives. *J. Mol. Graph. Model*. **2019**, 90, 180–191.

(4) Cheng, Y. H.; Zhou, M.; Tung, C. H.; Ji, M. J.; Zhang, F. H. Studies on two types of PTP1B inhibitors for the treatment of type 2 diabetes: hologram QSAR for OBA and BBB analogues. *Bioorg. Med. Chem. Lett.* **2010**, 20, 3329–3337.

(5) Sun, J. Y.; Wang, J. C.; Hu, M. QSAR and pharmacophore studies of thiazolidine-4-carboxylic acid derivatives as novel influenza neuraminidase inhibitors using HQSAR, topomer CoMFA and CoMSIA. *Chin. J. Struc. Chem.* **2013**, 32, 744–750.

(6) Tong, J. B.; Feng, Y.; Wang, T. H.; Wu, L. Y. Topomer CoMFA, HQSAR studies and molecular docking of 2,5-diketopiperazine derivatives as oxytocin inhibitors. *Chin. J. Struct. Chem.* **2020**, 39, 1385–1394.

(7) Jiao, L.; Wang, Y.; Qu, L.; Xue, Z. W.; Ge, Y. Q.; Liu, H. H.; Lei, B.; Gao, Q.; Li, M. K. Hologram QSAR study on the critical micelle concentration of Gemini surfactants. *Colloid. Surface. A* **2020**, 586, 12422–8.

(8) Jiao, L.; Zhang, X. F.; Qin, Y. C.; Wang, X. F.; Li, H. Hologram QSAR study on the electrophoretic mobility of aromatic acids. *Chemometr. Intell. Lab. Syst.* **2016**, 157, 202–207.

(9) Zhao, X. H.; Wang, X. L.; Li Y. Combined HQSAR method and molecular docking study on genotoxicity mechanism of quinolones with higher genotoxicity. *Environ. Sci. Pollut. Res.* **2019**, 26, 34830–34853.

(10) Yang, J. W.; Gu, W. W.; Li, Y. Biological enrichment prediction of polychlorinated biphenyls and novel molecular design based on 3*D*-QSAR/HQSAR associated with molecule docking. *Biosci. Rep*. **2019**, 39, BSR20180409–20.

(11) Gadaleta, D.; Vuković, K.; Toma, C.; Lavado, G. J.; Karmaus, A. L.; Kamel, M.; Kleinstreuer, N.; Benfenati, E.; Roncaglioni, A. SAR and QSAR modeling of a large collection of LD50 rat acute oral toxicity data. *J. Cheminformatics* **2019**, 11, 1–16.

(12) Li, Y. K.; Shao, X. G.; Cai, W. S. Partial least-squares regression method based on consensus modeling for quantitative analysis of near-infrared spectra. *Chem. J. Chin. Univ.* **2007**, 28, 246–249.

(13) Ouyang, L. H.; Zhou, D. Q.; Ma, Y. Z.; Tu, Y. L. Ensemble modeling based on 0-1 programming in micro-manufacturing process. *Comput. Ind. Eng.* **2018**, 123, 242–253.

(14) Su, Z. Q.; Tong, W. D.; Shi, L. M.; Shao, X. G.; Cai, W. S. A partial least-squares based consensus regression method for the analysis of near-infrared complex spectral data of plant samples. *Anal. Lett.* **2006**, 39, 2073–2083.

(15) Granitto, P. M.; Verdes, P. F.; Ceccatto, H. A. Neural network ensembles: evaluation of aggregation algorithms. *Artif. Intell*. **2005**, 163, 139–162.

(16) Jin, H. P.; Pan, B.; Chen, X. G.; Qian, B. Ensemble just-in-time learning framework through evolutionary multi-objective optimization for soft sensor development of nonlinear industrial processes. *Chemometr. Intell. Lab. Syst*. **2019**, 184, 153–166.

(17) Li, W. Z.; Miao, W.; Cui, J. X.; Fang, C.; Su, S. T.; Li, H. Z.; Hu, L. H.; Lu, Y. H.; Chen, G. H. Efficient corrections for DFT noncovalent interactions

based on ensemble learning models. *J. Chem. Inf. Model.* **2019**, 59, 1849–1857.

(18) Wert, E. C.; Rosario-Ortiz, F. L.; Snyder, S. A. Effect of ozone exposure on the oxidation of trace organic contaminants in wastewater. *Water Res.* **2009**, 43, 1005–1014.

(19) Hoigne, J.; Bader, H. Ozonation of water: kinetics of oxidation of ammonia by ozone and hydroxyl radicals. *Environ. Sci. Technol.* **1978**, 12, 79–84.

(20) Gunten, U. V. Ozonation of drinking water: part I. Oxidation kinetics and product formation. *Water Res.* **2003**, 37, 1443–1467.

(21) Zimmermann, S. G.; Schmukat, A.; Schulz, M.; Benner, J.; Gunten, V.; Ternes, T. A. Kinetic and mechanistic investigations of the oxidation of tramadol by ferrate and ozone. *Environ. Sci. Technol.* **2012**, 46, 876–884.

(22) Ternes, T. A.; Stüber, J.; Herrmann, N.; McDowell, D.; Ried, A.; Kampmann, M.; Teiser, B. Ozonation: a tool for removal of pharmaceuticals, contrast media and musk fragrances from wastewater? *Water Res.* **2003**, 37, 1976–1982.

(23) Nakada, N.; Shinohara, H.; Murata, A.; Kiri, K.; Managaki, S.; Sato, N.; Takada, H. Removal of selected pharmaceuticals and personal care products (PPCPs) and endocrine-disrupting chemicals (EDCs) during sand filtration and ozonation at a municipal sewage treatment plant. *Water Res.* **2007**, 41, 4373–4382.

(24) Huber, M. M.; Gobel, A.; Joss, A.; Hermann, N.; Löffler, D.; McArdell, C. S.; Ried, A.; Siegrist, H.; Ternes, T. A.; von Gunten, U. Oxidation of pharmaceuticals during ozonation of municipal wastewater effluents: a pilot study. *Environ. Sci. Technol.* **2005**, 39, 4290–4299.

(25) Hammes, F.; Salhi, E.; Köster, O.; Kaiser, H. P.; Egli, T.; von Gunten, U. Mechanistic and kinetic evaluation of organic disinfection by-product and assimilable organic carbon (AOC) formation during the ozonation of drinking water. *Water Res.* **2006**, 40, 2275–2286.

(26) Öberg, T. A QSAR for the hydroxyl radical reaction rate constant: validation, domain of application, and prediction. *Atmos. Environ.* **2005**, 39, 2189–2200.

(27) Li, X. H.; Zhao, W. X.; Li, J.; Jiang, J. Q.; Chen, J. J.; Chen, J. W. Development of a model for predicting reaction rate constants of organic chemicals with ozone at different temperatures. *Chemosphere* **2013**, 92, 1029–1034.

(28) Long, X.; Niu, J. Estimation of gas-phase reaction rate constants of alkylnaphthalenes with chlorine, hydroxyl and nitrate radicals. *Chemosphere* **2007**, 67, 2028–2034.

(29) Yang, Z. H.; Luo, S.; Wei, Z. S.; Ye, T. T.; Spinney, R.; Chen, D.; Xiao, R. Y. Rate constants of hydroxyl radical oxidation of polychlorinated biphenyls in the gas phase: a single-descriptor based QSAR and DFT study. *Environ. Pollut.* **2016**, 211, 157–164.

(30) Toropov, A. A.; Toropova, A. P.; Rasulev, B. F.; Benfenati, E.; Gini, G.; Leszczynska, D.; Leszczynski, J. Coral: QSPR modeling of rate constants of reactions between organic aromatic pollutants and hydroxyl radical. *J. Comput. Chem.* **2012**, 33, 1902–1906.

(31) Yu, X. L.; Yi, B.; Wang, X. Y.; Chen, J. F. Predicting reaction rate constants of ozone with organic compounds from radical structures. *Atmos. Environ.* **2012**, 51, 124–130.

(32) Sudhakaran, S.; Calvin, J.; Amy, G. L. QSAR models for the removal of organic micropollutants in four different river water matrices. *Chemosphere* **2012**, 87, 144–150.

(33) Lee, Y.; von Gunten, U. Quantitative structure-activity relationships (QSARs) for the transformation of organic micropollutants during oxidative water treatment. *Water Res.* **2012**, 46, 6177–6195.

(34) Sudhakaran, S.; Amy, G. L. QSAR models for oxidation of organic micropollutants in water based on ozone and hydroxyl radical rate constants and their chemical classification. *Water Res.* **2013**, 47, 1111–1122.

(35) Consonni, V.; Ballabio, D.; Todeschini, R. Comments on the definition of the Q^2 parameter for QSAR validation. *J. Chem. Inf. Model.* **2009**, 49, 1669–1678.

(36) Chirico, N.; Gramatica, P. Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient. *J. Chem. Inf. Model.* **2011**, 51, 2320–2335.

(37) Bhhatarai, B.; Teetz, W.; Liu, T.; Öberg, T.; Jeliazkova, N.; Kochev, N.; Pukalov, O.; Tetko, I. V.; Kovarich, S.; Papa, E.; Gramatica, P. CADASTER QSPR models for predictions of melting and boiling points of perfluorinated chemicals. *Mol. Inform.* **2011**, 30, 189–204.

(38) Chirico, N.; Gramatica, P. Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection. *J. Chem. Inf. Model.* **2012**, 52, 2044–2058.

(39) Tropsha, A.; Gramatica, P.; Gombar, V. K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models. *Mol. Inform.* **2003**, 22, 69–77.

(40) Roy, K.; Mitra, I.; Kar, S.; Ojha, P. K.; Das, R. N.; Kabir, H. Comparative studies on some metrics for external validation of QSPR models. *J. Chem. Inf. Model.* **2012**, 52, 396–408.

(41) Hurst, J. R.; Heritage, T.W. *Patent, US005751605A* **1998**.

(42) Hurst, J. R.; Heritage, T.W. *Patent, US6208942B1* **2001**.

(43) Zhang, C. Q.; Du, C. M.; Feng, Z. W.; Zhu, J. Y.; Li, Y. Y. HQSAR, docking, and molecular dynamics studies of inhibitors for CXCR4. *Chem. Biol. Drug. Des.* **2014**, 85, 1877–1880.

(44) Guido, R. V. C.; Castilho, M. S.; Mota, S. G. R.; Oliva, G.; Andricopulo, A. D. Classical and hologram QSAR studies on a series of inhibitors of trypanosomatid glyceraldehyde-3-phosphate dehydrogenase. *Mol. Inform.* **2010**, 27, 768–781.

(45) Yu, S. L.; Yuan, J. T.; Zhang, Y.; Gao, S. F.; Gan, Y.; Han, M.; Chen, Y. W.; Zhou, Q. Q.; Shi, J. H. Combined HQSAR, topomer CoMFA, homology modeling and docking studies on triazole derivatives as SGLT2 inhibitors. *Future Med. Chem.* **2017**, 9, 847–858.

(46) Muñoz-Gutiérrez, C.; Caballero, J.; Morales-Bayuelo, A. HQSAR and molecular docking studies of furanyl derivatives as adenosine A_{2A} receptor antagonists. *Med. Chem. Res*. **2016**, 25, 1–13.

(47) Cheng, Y. H.; Zhou, M.; Tung, C. H.; Ji, M. J.; Zhang, F. H. Studies on two types of PTP1B inhibitors for the treatment of type 2 diabetes: hologram QSAR for OBA and BBB analogues. *Bioorg. Med. Chem. Lett.* **2010**, 20, 3329–3337.

(48) Flower, D. R. On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **1998**, 38, 379–386.

(49) Krogh, A.; Vedelsby, J. Neural network ensembles, Eross validation and active learning. *Adv. Neural. Inform. Process. Syst.* **1995**, 7, 231–238.

(50) Bian, X. H.; Diwu, P. Y.; Liu, Y. R.; Liu, P. Ensemble calibration for the spectral quantitative analysis of complex samples. *J. Chemometr*. **2017**, 32, e2940–13.

(51) Li, Y.; Jing, J. A consensus PLS method based on diverse wavelength variables models for analysis of near-infrared spectra. *Chemometr. Intell. Lab. Syst.* **2014**, 130, 45–49.