Bull Environ Contam Toxicol (2012) 89:316–321
DOI 10.1007/s00128-012-0696-6

# 2D and 3D-QSBR Study on Biodegradation of Phenol Derivatives

**Fuyang Wang · Jiaqi Shi**

Received: 13 March 2012 / Accepted: 24 May 2012 / Published online: 15 June 2012


**Abstract** Eighteen phenol derivatives were optimized with density function theory (DFT) and comparative molecular similarity indices analysis (CoMSIA) respectively. Corresponding 2D and 3D descriptors were obtained to establish QSBR models. The biodegradation of them is mainly related to $\alpha$ and $S^{o}$ according to the 2D QSBR, and influenced by the hydrophobicity and hydrogen bonding properties following the 3D QSBR. The 2D model performs better in stability and predictive ability than the 3D one. To some degrees, the two models verify and supplement each other. They can be used in predicting biodegradation of chlorinated, amine, nitro, nitroso and methyl phenol.

**Keywords** Quantitative structure—biodegradation relationship (QSBR) · Phenol derivative · Production of carbon dioxide (PCD) · Density function theory (DFT) · Comparative molecular similarity indices analysis (CoMSIA)

Phenol derivatives have been widely used in chemical production and are released to the environment as a kind of harmful pollutants. Biodegradation is an important way to remove them from the environment, and has attracted the attention of many scientists (Goi et al. 2004; Hsu et al. 2005; Suarez-Ojeda et al. 2007, 2008; Hao et al. 2009). They mainly concern with their intermediate products and factors affecting the degradation. Various substituents have influences on the biodegradation. Liu et al. (2003) found that the relative biodegradability of 51 substituted benzenes were mainly affected by some substructures such as –CH3, –CH=, >C=, $–NH_2$, $–NO_2$, –OH, $–SO_3H$, and –Cl. Chen et al. (1997) investigated the aerobic biodegradation of 32 kinds of aromatic compounds and found that the degradation of mono-substitutional phenol weakened in turn when the ortho or para position was replaced with –OH, $–NO_2$, $–NH_2$, $–CH_3$, –Cl, respectively. The degradation of di-substitutional phenol weakened in turn when they were replaced with $–CH_3$, –Cl and $–NO_2$, respectively. Moreover, it is harder to degrade molecules with increasing number of substituents. Pagga (1997) proposed that ultimate aerobic biodegradation the basic formula test substance (DOC) + $O_2$ → $CO_2$ + $H_2O$+ biomass demonstrated the possibilities of measuring biodegradation. To determine ultimate degradability (mineralization), the production of carbon dioxide (PCD) is frequently used. The information obtained is very useful for most cases.

Quantitative structure—activity relationship (QSAR) is a method which predicts activity of chemicals based on their structure. It has been used in the study of phenol derivatives on many properties, including the catalysis of estrogen enzyme on degradation, the relationship between biodegradation rate constant (ln $K$) and substituting groups, the acute toxicity to mice connective tissue fibroblasts L929 and human hepatocellular carcinoma cells HepG2, estrogenic activity, and so on (Tabak and Rakesh 1993; Jiang et al. 2004; Cui et al. 2006; Mao and Gao 2008). We can call it quantitative structure—biodegradation relationship (QSBR) when the activity is biodegradability. Structural and thermodynamic descriptors calculated based on density function theory (DFT) have been successful in establishing 2D models (Yang et al. 2010; Shi et al. 2011). Comparative molecular similarity indices analysis (CoMSIA) has been widely used to build 3D model which takes 3D structure as descriptors and overcomes the limitations of conventional

F. Wang (✉) · J. Shi
Jinling School, Nanjing University, 8 Xuefu Road, Pukou District, Nanjing 210089, People's Republic of China
e-mail: wfyayz@163.com

2D model in characterizing the relation between activity and structure.

The PCD of phenols has not been studied by the combination of 2D and 3D methods to this day, and we do not know how molecular structure influences their degradation. In the present study, molecules are optimized on different levels with Gaussian 03, to build the best 2D-QSBR model of PCD. 3D-QSBR based on CoMSIA of selected phenol derivatives is also established. The comparative analysis of 2D and 3D models are performed. The object is to obtain information about the biodegradability of the homologous series of phenol compounds. It is hoped that this work is conducive to gaining the PCD data of phenol derivatives, and provides reference for their biological degradation.

## Materials and Methods

The experimental PCD values of selected phenol derivatives come from Chen et al. (1997) and listed in Table 1. They were determined after 12 days of biodegradation.

The total experimental data set (n = 18) was divided into training set (n = 15) and test set (n = 3). The experimental data were ordered according to the compound names and compounds of No. 6, 12, and 18 were picked as the training set. They were marked with asterisk (*) in Table 1. The template (No. 2) during the 3D modeling was kept in the training set. All the steps of the test set molecules in calculation of descriptors were the same as those of the training set molecules, and PCD was predicted using the model derived from the training set.

The 18 kinds of phenol derivatives were fully optimized at the B3LYP/midix, 6-31G* and 6-311G** levels respectively with Gaussian 03 program. Frequency calculations were also performed to ensure they were at the minimal potential energy surface. There is no negative frequency in all the calculation results of vibration analysis. Structural and thermodynamic descriptors were therefore obtained. Structural descriptors include: dipole moments ($\mu$), energy of the highest occupied molecular orbital ($E_{HOMO}$), energy of the lowest unoccupied molecular orbital ($E_{LUMO}$), the most negative atomic net charges of the molecular ($q^-$), the most positive atomic net charges of the molecular on the hydrogen ($qH^+$), molecular volume ($V_m$) and molecular average polarizability ($\alpha$). Thermodynamic parameters include: total energy (*TE*), zero-point vibrational energy (*ZPE*), enthalpy ($H^o$), free energy ($G^o$), correction value of thermal energy ($E_{th}$) (i.e., the sum of molecular vibrational energy, rotational energy and translational energy), heat capacity at constant volume ($C_V^o$) and entropy ($S^o$). The correlations of PCD with the descriptors were obtained using multiple linear regression method of the SPSS 12.0 for windows program.

The energy optimization was made for molecular structures of 18 compounds with the Tripos standard molecular force

**Table 1** The experimental and predicted production of carbon dioxide (PCD) of phenol derivates

| No | Chemical | PCD (mmol $L^{-1}$) | | | | | Descriptor (on 6-311G** level) | |
|---|---|---|---|---|---|---|---|---|
| | | Exp. | Eq. (7) | | CoMSIA | | $\alpha$ (Debye) | $S^o$ (J $mol^{-1}$ $K^{-1}$) |
| | | (Chen et al. 1997) | Pred. | Res. | Pred. | Res. | | |
| 1 | Catechol | 16.30 | 16.242 | −0.058 | 15.440 | −0.860 | 68.075 | 336.810 |
| 2 | Resorcinol | 17.94 | 15.320 | −2.620 | 14.425 | −3.515 | 68.444 | 335.472 |
| 3 | Phloroglucinol | 15.32 | 17.609 | 2.289 | 17.144 | 1.824 | 72.796 | 358.661 |
| 4 | *o*-aminophenol | 9.50 | 9.902 | 0.402 | 8.704 | −0.796 | 73.744 | 339.436 |
| 5 | *m*-aminophenol | 10.44 | 9.742 | −0.698 | 10.355 | −0.085 | 73.974 | 339.834 |
| 6* | *o*-chlorophenol | 5.36 | 8.614 | 3.254 | 4.018 | −1.342 | 75.370 | 341.757 |
| 7 | *p*-chlorophenol | 7.27 | 8.506 | 1.236 | 8.956 | 1.686 | 75.701 | 342.690 |
| 8 | *o*-nitrodiphenol | 12.72 | 13.077 | 0.357 | 12.721 | 0.001 | 80.041 | 372.583 |
| 9 | *p*-nitrodiphenol | 11.02 | 9.976 | −1.044 | 12.146 | 1.126 | 82.085 | 371.119 |
| 10 | 2,3-dichlorophenol | 2.00 | 3.692 | 1.692 | 2.764 | 0.764 | 86.737 | 370.078 |
| 11 | 2,4-dichlorophenol | 1.69 | 2.903 | 1.213 | 1.610 | −0.080 | 87.861 | 371.981 |
| 12* | 2,5-dichlorophenol | 2.06 | 2.572 | 0.512 | 7.719 | 5.659 | 88.123 | 371.989 |
| 13 | 3,5-dihydroxytoluene | 13.58 | 14.539 | 0.959 | 15.435 | 1.855 | 81.621 | 382.870 |
| 14 | 2,4-dimethylphenol | 3.01 | 5.177 | 2.167 | 1.961 | −1.049 | 90.381 | 388.215 |
| 15 | 2,6-dimethylphenol | 4.67 | 3.379 | −1.291 | 5.350 | 0.680 | 89.574 | 379.851 |
| 16 | 3,5-dimethylphenol | 14.22 | 12.916 | −1.304 | 12.572 | −1.648 | 90.398 | 411.178 |
| 17 | 5-nitrosoorthocresol | 2.94 | 1.000 | −1.940 | 3.043 | 0.103 | 93.350 | 387.057 |
| 18* | 2,4,6-trichlorophenol | −2.78 | −3.838 | −1.058 | 2.583 | 5.363 | 100.378 | 399.256 |

field. The energy convergence standard was 0.05 kcal $mol^{-1}$ $Å^{-1}$. The atom net charge in the molecule was calculated with Gasteriger-Hückel method to obtain minimum energy conformation. The benzene ring was used as the superposition skeleton since each compound contained it and resorcinol (No. 2) compound with the largest PCD was used as the superposition template for superposition. CoMSIA analysis was conducted by the SYBYL 7.3 software package (SYBYL software, Version 7.3, 2006). Five fields, namely steric, electrostatic, hydrophobic, hydrogen bond donor and acceptor fields were calculated at each lattice intersection of a regularly spaced grid of 2.0 Å. Energy cutoff values of 30 kcal $mol^{-1}$ were selected for the fields. The *sp*3 hybrid C+ was used as the probe atom to calculate the value and distribution of five field energies at the peripheral grid of superpositioned molecule. Other parameters of the fields were set as default.

Statistical analysis was carried out with the partial least-squares regression (PLS) method. To improve the signal–noise ratio for CoMSIA, a minimum column filtering of 2.0 kcal $mol^{-1}$ was used. By the leave-one-out (LOO) cross validation method, the optimum main ingredient number $n$ and cross validation correlation coefficient ($q^2$) were determined. Then, non-cross validation was made and CoMSIA model was established.

The goodness-of-fit of the models was mainly tested by the conventional correlation coefficient ($R^2$), the standard error of estimate (*SEE*) and the Fisher test value $F$.

The Leave-One-Out (LOO) cross-validation (CV) was performed. That the cross-validation coefficient ($Q^2_{LOO} > 0.7$) and the root-mean-square error of cross-validation ($RMSE_{CV} < 1$) was used to verify robustness and internal predictivity of each model (Gramatica 2007; Puzyn et al. 2008).

Then the predictive correlation ($R^2_{pred}$) based on the test set molecules was computed using Eq. (1)

$$R^2_{pred} = (SD - PRESS)/SD \quad (1)$$

where *SD* is the sum of the squared deviations between the experimental values of the test set and the mean values of the training set compounds, and *PRESS* is the sum of the squared deviation between the predicted and actual values for each molecule in the test set. $R^2_{pred}$ reflects the external predictivity of models. The scatter plot of experimental versus predicted values for the training and test set was also given to confirm the predictivity of the models.

## Results and Discussion

The relationship between PCD and the descriptors was obtained with multiple linear stepwise regression. The obtained 2D-QSBR models were all listed in Table 2, and the descriptors at three levels were all listed in the supplement material.

We can see that the model does better with increasing number of descriptors. Equation (7) performs the best and is taken as the model in this work. The $\alpha$ and $S^e$ enter the model and their values are listed in Table 1. The predicted values of 18 chemicals are also shown in Table 1, and the scatter plots of experimental versus predicted values for the training and test set is given in Fig. 1. The biggest residual is 3.254 of *o*-chlorophenol and the mean absolute error is 1.339. It can be seen that the 2D model has quite a good stability and predictive ability.

In this work, variance inflation factor (*VIF*) is adopted to evaluate the collinearity of descriptors in Eq. (7). *VIF* is defined as $VIF = 1/(1 - r^2)$, where $r$ represents the multiple regression correlation coefficient between one independent variable and others. If *VIF* is 1.0, there is no self-correlation among the variables; that *VIF* is between 1.0 and 5.0 represents the correlation equation can be accepted; and $VIF > 10$ states the regression equation is unstable and shall be retested.

According to Table 3, *VIF* of Eq. (7) is 4.607 (less than 5.0). When the confidence is 95 %, the standard $t$ value ($t_{\alpha/2}$) is 2.145, the $t$ values of $\alpha$ and $S^e$ are −11.459 and 8.200 respectively, whose absolute values are greater than the standard $t$ value. It indicates that the relevance among the parameters in Eq. (7) is small, the equation has great statistical significance and fine stability, and this equation can be accepted.

The standard regression coefficeints of $\alpha$ and $S^e$ are −1.941 and 1.389 respectively. The most important factor influencing PCD is $\alpha$, which is the measure of the change in a molecule's electron distribution in response to an applied electric field. Chen et al. (2010) has investigated the relation between biodegradation rate constant ($K_b$) of chlorophenol compounds and theoretical descriptors. The equation obtained is $K_b{*}100 = 549.016 + 3.686\ \alpha - 3.605\ S^e + 3.789\ C^e_V$, with the $R^2$ of 0.894 and $Q^2_{LOO}$ of 0.796. The $\alpha$ entered into the equation firstly and then $S^e$ (Chen et al. 2010). The $\alpha$ determines the dynamical response of a bound system to external fields, and provides insight into a molecule's internal structure. The greater the amount of electrons and the distance of electrons from nuclear charge, the less control the nuclear charge has on charge distribution, and thus the increased $\alpha$ of the atom. The $\alpha$ has an inverse relationship with PCD. The greater $\alpha$ is, the smaller PCD is. That means that the chemical with greater $\alpha$ is more difficult to biodegrade. In addition, Eq. (8) shows that $\alpha$ has a good correlation with molecular volume.

$$\begin{aligned} &\alpha = -13.333 + 0.984 M_v \\ &n = 18, R^2 = 0.962, SEE = 1.848 \end{aligned} \quad (8)$$

The $S^e$ is a thermodynamic property used to determine the energy not available for work in a thermodynamic process. When $S^e$ comes into the model, the degree of

**Table 2** 2D-QSBR models at three levels

| Basis set | No | Equation | n | $R^2$ | *SEE* | *F* | $Q^2_{LOO}$ | $RMSE_{CV}$ | $R^2_{pred}$ |
|---|---|---|---|---|---|---|---|---|---|
| Midix | 2 | PCD = 49.185(±12.021) − 0.541(±0.163)*$\alpha$ | 15 | 0.458 | 4.256 | 10.985 | 0.321 | 4.534 | 0.985 |
| | 3 | PCD = −2.019(±11.712 − 1.313(± 0.171)*$\alpha$ + 0.298(± 0.056)*$S^o$ | 15 | 0.841 | 2.403 | 31.621 | 0.755 | 2.739 | 0.991 |
| 6-31G* | 4 | PCD = 46.583(±10.175) − 0.488(±0.133)*$\alpha$ | 15 | 0.508 | 4.056 | 13.419 | 0.381 | 0.431 | 0.986 |
| | 5 | PCD = −7.612(±12.980) − 1.153(±0.162)*$\alpha$ + 0.290(±0.061)*$S^o$ | 15 | 0.830 | 2.484 | 29.214 | 0.768 | 2.622 | 0.989 |
| 6-311G** | 6 | PCD = 47.379(±10.416) − 0.468(±0.128)*$\alpha$ | 15 | 0.507 | 4.061 | 13.354 | 0.374 | 4.344 | 0.986 |
| | 7 | PCD = −10.804(±8.255) − 1.275(±0.111)*$\alpha$ + 0.338 (±0.041)*$S^o$ | 15 | 0.925 | 1.645 | 74.321 | 0.863 | 2.024 | 0.992 |

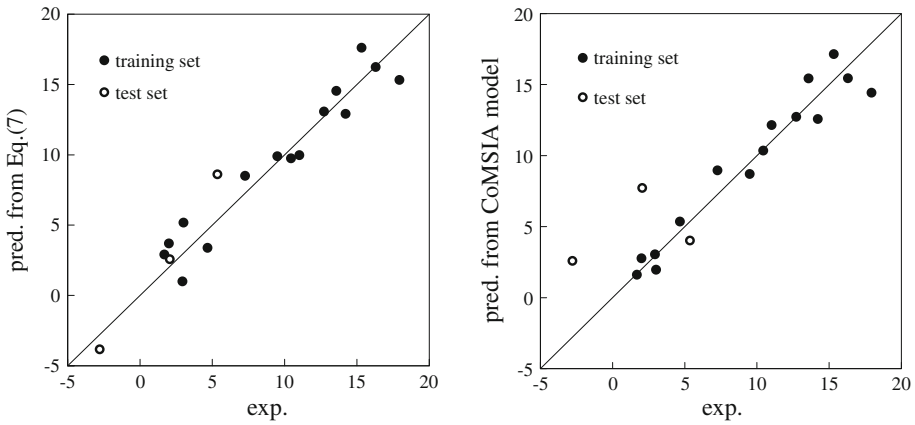


**Fig. 1** The *scatter plot* of experimental versus predicted values for the training and test set

**Table 3** *VIF*, *SR* and *t* test value of descriptors in 2D model

| Descriptor | VIF | SR | $t$ ($t_{\alpha/2}$ = 2.145) $\alpha$ = 0.05 |
|---|---|---|---|
| $\alpha$ | 4.607 | −1.941 | −11.459 |
| $S^o$ | 4.607 | 1.389 | 8.200 |

fitting is much better than $\alpha$ alone. The PCD is greater with a bigger $S^o$ value.

Figure 2 shows molecular superposition results in CoMSIA analysis. It indicates that all molecules are in good superposition.

The summary of the statistical results obtained for CoMSIA study is shown in Table 4.

The series of measurement data of CoMSIA model shows good stability. The scatter plot of experimental versus predicted values for the training and test set is also given in Fig. 1. The 2,5-dichlorophenol (No. 12) and 2,4,6-trichlorophenol (No. 18) in the test set have the greatest difference of predictive value of 5.659 and 5.363 respectively, while others have predictive values close to the experimental ones. That indicates that the 3D model has some flaws in predicting high chlorine replaced phenols. The mean absolute error of predictive values of 18 chemicals is 1.580. The predictive ability is acceptable on the whole. Based on the contribution values of field energy, the hydrophobic, hydrogen bond donor and acceptor descriptors played more significant roles (26.0, 26.3 and 29.5 % of contribution respectively) than descriptors such as steric

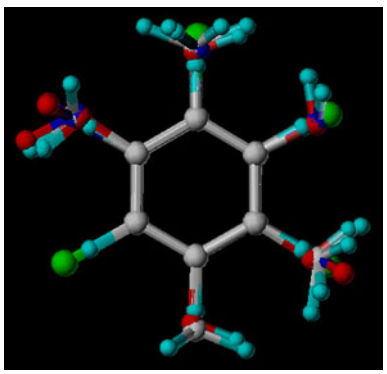

**Fig. 2** Superposition of substituted phenols

**Table 4** Statistical parameters of the CoMSIA model of PCD

| | *n* | $R^2$ | *SEE* | *F* | *S* | *E* | *H* | *D* | *A* |
|---|---|---|---|---|---|---|---|---|---|
| PCD | 2 | 0.871 | 2.162 | 40.477 | 0.089 | 0.094 | 0.260 | 0.263 | 0.295 |

*S*, *E*, *H*, *D* and *A* represent the contribution values of steric, electrostatic, hydrophobic, hydrogen bond donor and acceptor field, respectively

(8.9 %) and electrostatic (9.4 %) in the prediction of PCD. The previous QSBR models based on novel geometrical chemical descriptors of the protein–ligand interfaces have revealed that the models tend to rely on chemically diverse descriptor types that capture major intermolecular binding interactions such as hydrophobic effect and hydrogen bonds (Zhang et al. 2006).

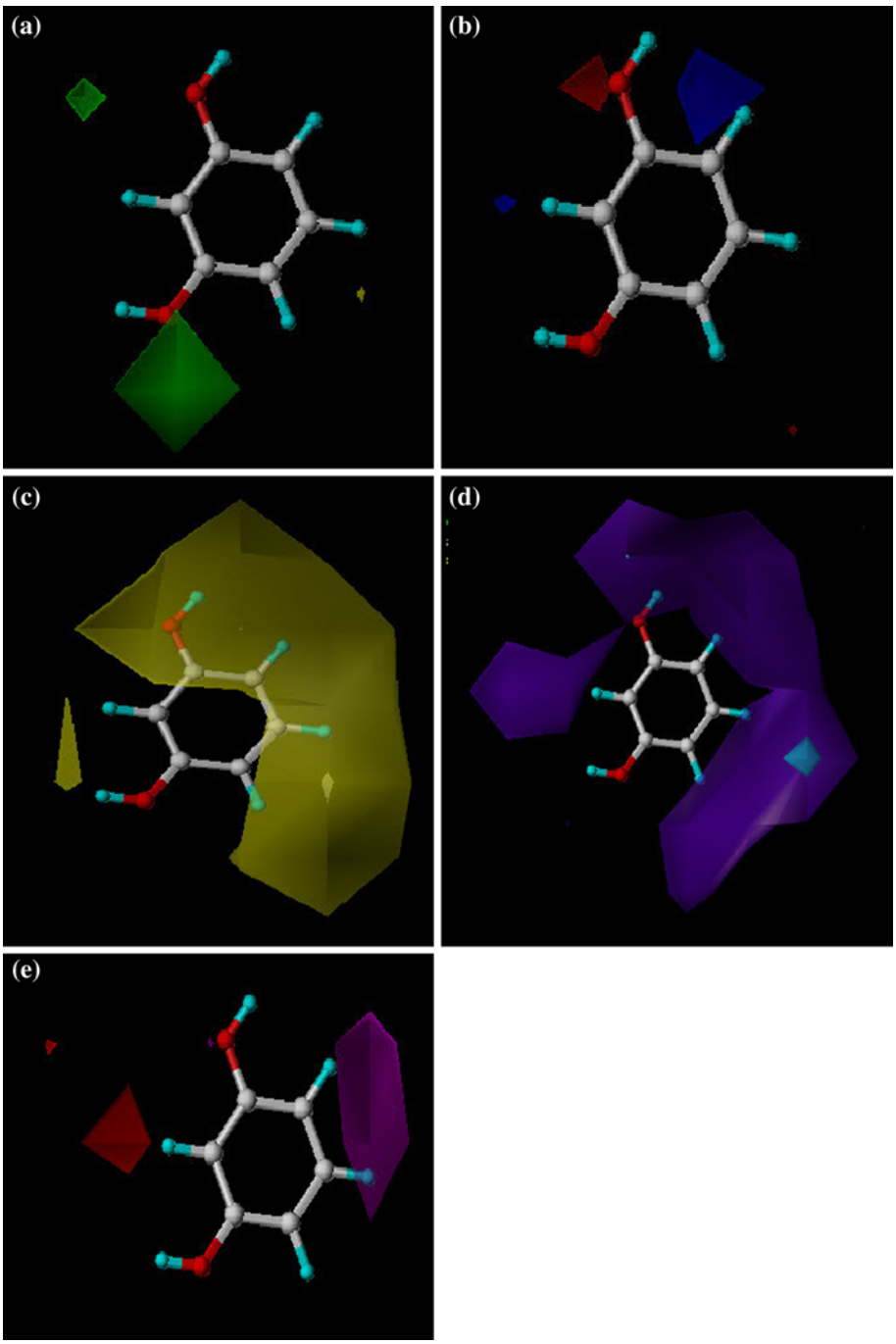


**Fig. 3** Contour map of CoMSIA **a** steric, **b** electrostatic, **c** hydrophobic, **d** donor and **e** acceptor fields of PCD

As hydrophobicity of a chemical is also relevant to its molecular volume, the 3D model proves the 2D-QSBR from a certain extent. Contour plots obtained from the CoMSIA method are useful to explore protein–ligand interactions, as shown in Fig. 3. For simplicity, only interaction between the compound 2 and the contour plots is shown. The green contour observed in Fig. 3a indicates that some bulky substitutions at these positions are favorable for biodegradation. The CoMSIA electrostatic contour is shown in Fig. 3b. The red contours indicate that negative charge in these regions is favored for biodegradation while the blue means opposite. The yellow contour in CoMSIA hydrophobic contour (Fig. 3c) around the molecule indicates that a hydrophilic group at these positions enhances PCD. That may be the reason why polysubstituted phenol tends to have higher PCD values. The purple contour in Fig. 3d around the molecule indicates that a hydrogen bonding donor group here degrades PCD. The purple red contour in Fig. 3e signifies that a hydrogen bonding acceptor group increases PCD while the red means opposite.

In general, the 2D model performs better in stability and predictive ability than the 3D one. It is better not to use the 3D model to predict high substituted phenols. To some degrees, the two models verify and supplement each other since they both related with the molecule volume. The models can be used in predicting biodegradation of chlorinated, amine, nitro, nitroso and methyl phenol.

## References


Chen YS, Chen LX, Yang J, Zhuang YY, Dai SG (1997) A study on biodegradability of 32 aromatic compounds. Environ Chem 16:43–48

Chen YJ, Wang ZY, Mao L, Gao SX (2010) QSBR study on the biodegradation rate constant of chloro-phenol compounds. Chin J Struct Chem 29:895–899

Cui SH, Liu SS, Yang J, Wang XD, Wang LS (2006) Quantitative structure-activity relationship of estrogen activities of bisphenol A analogs. Sci China Phys Mech 51:287–292

Goi A, Trapido M, Tuhkanen T (2004) A study of toxicity, biodegradability, and some by-products of ozonised nitrophenols. Adv Environ Res 8:303–311

Gramatica P (2007) Principles of QSAR models validation: internal and external. QSAR Comb Sci 26:694–701

Hao RX, Li JB, Zhou YW, Cheng SY, Yi Z (2009) Structure–biodegradability relationship of nonylphenol isomers during biological wastewater treatment process. Chemosphere 75:987–994

Hsu YC, Yang HC, Chen JH (2005) The effects of preozonation on the biodegradability of mixed phenolic solution using a new gas-inducing reactor. Chemosphere 59:1279–1287

Jiang J, Chen JN, Yu HX, Zhang F, Zhang JF, Wang LS (2004) Quantitative structure activity relationship and toxicity mechanisms of chlorophenols on cells in vitro. Chin Sci Bull 49:562–566

Liu Y, Liu SS, Cui SH, Cai SX (2003) A novel quantitative structure-biodegradability relationship (QSBR) of substituted benzenes based on MHDV descriptor. J Chin Chem Soc 50:319–324

Mao L, Gao SX (2008) Reactions of estrogenic phenolic chemicals mediated by horseradish peroxidase: quantitative structure-activity relationships. Acta Sci Circumst 28:2562–2567

Pagga U (1997) Testing biodegradability with standardized methods. Chemosphere 35:2953–2972

Puzyn T, Suzuki N, Haranczyk M (2008) How do the partitioning properties of polyhalogenated POPs change when chlorine is replaced with bromine? Environ Sci Technol 42:5189–5195

Shi JQ, Liu HX, Sun L, Hou HF, Xu Y, Wang ZY (2011) Theoretical study on hydrophilicity and thermodynamic properties of polyfluorinated dibenzofurans. Chemosphere 84:296–304

Suarez-Ojeda ME, Fabregat A, Stuber F, Fortuny A, Carrera J, Font J (2007) Catalytic wet air oxidation of substituted phenols: temperature and pressure effect on the pollutant removal, the catalyst preservation and the biodegradability enhancement. Chem Eng J 132:105–115

Suarez-Ojeda ME, Carrera J, Metcalfe IS, Font J (2008) Wet air oxidation (WAO) as a precursor to biological treatment of substituted phenols: refractory nature of the WAO intermediates. Chem Eng J 144:205–212

Tabak HH, Rakesh G (1993) Prediction of biodegradation kinetics using a nonlinear group contribution method. Environ Toxicol Chem 12:251–260

Yang X, Liu HX, Hou HF, Flamm A, Zhang XS, Wang ZY (2010) Studies of thermodynamic properties and relative stability of a series of polyfluorinated dibenzo-*p*-dioxins by density functional theory. J Hazard Mater 181:969–974

Zhang SX, Golbraikh A, Tropsha A (2006) The development of quantitative structure-binding affinity relationship (QSBR) models based on novel geometrical chemical descriptors of the protein-ligand interfaces. J Med Chem 49:2713–2724