Structural Characterization and Aquatic Toxicity Prediction of Esters①

LI Jian-Feng[a] LIAO Li-Min[a, b②]

[a] (*College of Chemistry and Chemical Engineering, Neijiang Normal University, Neijiang* 641100, *China*)

[b] (*Key Laboratory of Fruit Waste Treatment and Resource Recycling of Sichuan Provincial College, Neijiang* 641100, *China*)

ABSTRACT Based on the three-dimensional structures of the compounds, the structures of 48 ester compounds were expressed parametrically. Through multiple linear regression and partial least-squares regression, the relationship models between ester compound structures and aquatic toxicity log(1/IGC_{50}) were established. The correlation coefficients (R^2) of the models were 0.9974 and 0.9940, and the standard deviations (SD) were 0.0469 and 0.0646, respectively. The stability of the models was evaluated by the leave-one-out internal cross-test. The correlation coefficients (R_{CV}^2) of the models of interactive tests were 0.9939 and 0.8952, and the standard deviation (SD_{CV}) was 0.0715 and 0.0925, respectively. The external samples were used to test the predictive ability of the models, and the correlation coefficients (R_{test}^2) of the external predictions were 0.9955 and 0.9955, and the standard deviations (SD_{test}) were 0.0720 and 0.0716, respectively. The molecular structure descriptors could successfully represent the structural characteristics of the compounds, and the built models had good fitting effects, strong stability and high prediction accuracy. The present study has a good reference value for the study of the structure-toxicity relationship of toxic compounds in the environment.

Keywords: ester compounds, structural characterization, aquatic toxicity, simulation prediction;

DOI: 10.14102/j.cnki.0254–5861.2011–3032

1 INTRODUCTION

Esters are one of the important high-yield compounds. They are often used in the production of plastics. They are usually found in plastic pipes, furniture, floors, car interiors, insect repellents and cosmetics. Methyl *p*-hydroxybenzoate is the methyl ester of *p*-hydroxybenzoic acid (PHBA), which is widely used in cosmetics, toothpaste, hair care products, moisturizers and deodorants. Due to the wide applications of esters, more and more ester compounds enter the water environment and cause harm to living animals and plants[1-3]. Comprehensive acquisition of various property parameters of organic compounds is of great significance for standardizing their production and application[4, 5]. At present, the toxicity of ester compounds is mainly determined by experiments, which wastes resources such as chemical reagents and time. Moreover, the number of such compounds is huge, and it is difficult to measure various parameters only by experimental means. The study of the relationship between the structures and properties of compounds is of great significance for analyzing and evaluating various properties or environmental behaviors of compounds, and assisting in the identification of compounds. The parameterized characterization of structures of compounds is one of the key steps to establish the relationships between compound structures and properties. At present, two-dimensional structure characterization methods[6-8] and three-dimensional structure characterization methods[9-11] are widely used. The two-dimensional structure characterization methods are simple and fast, but they are difficult to reflect the three-dimensional structure characteristics of the compounds, and cannot distinguish phenomena such as *cis-trans* isomerisms. The three-dimensional structure characterization methods are relatively complicated, but they can be calculated based on the three-dimensional structures of compound molecules and can distinguish various isomerism phenomena. In the present

Received 9 November 2020; accepted 8 December 2020

① This project was supported by the Youth Foundation of Sichuan Provincial Department of Education (18ZB0323)

② Corresponding author. Liao Li-Min, male, born in 1981, professor, majoring in quantitative structure-activity relationship. E-mail: liaolimin523@126.com

study, three-dimensional structure descriptors were used to characterize the structures of some ester compounds, and then the multiple linear regression (MLR) and the partial least-squares regression (PLS) were used to establish the models of relationship between compound structures and toxicity, and the structural factors affecting compound toxicity were analyzed. This paper can provide a reference for the study of the structure-property relationship of ester compounds.

2 MATERIALS AND METHODS

2. 1 Experimental materials

In the present study, two QSAR models for the modeling and predicting aquatic toxicity $\log(1/IGC_{50})$ of 48 aliphatic esters were proposed. The experimental toxic activities which show toxic effects on the *Tetrahymena pyriformis* protozoa ciliate were taken from literature[12]. The samples were divided into the training and test sets, and the test set samples were marked with "*".

Table 1. Compounds and Their Toxicity Values

No.	Compound	$\log(1/IGC_{50})$	Cal.1	Err.1	Cal.2	Err.2
1	Methyl propionate	–1.6092	–1.5590	0.0502	–1.4715	0.1377
2	Methyl acetate	–1.5954	–1.5562	0.0392	–1.5655	0.0299
3	Methyl formate	–1.4982	–1.4494	0.0488	–1.5288	–0.0306
4	Isobutyl formate	–1.3081	–1.2827	0.0254	–1.2795	0.0286
5	Ethyl acetate	–1.2968	–1.3663	–0.0695	–1.4073	–0.1105
6*	Methyl butyrate	–1.2463	–1.1420	0.1043	–1.2100	0.0363
7	Propyl acetate	–1.2382	–1.2242	0.0140	–1.1971	0.0411
8	Propargyl acetate	–1.1664	–1.1194	0.0470	–1.0973	0.0691
9	Methyl-2-methylbutyrate	–1.1650	–1.1455	0.0195	–1.1315	0.0335
10	Propyl formate	–1.0221	–1.0157	0.0064	–1.0858	–0.0637
11	Ethyl propionate	–0.9450	–0.9949	–0.0499	–0.9595	–0.0145
12*	Butyl formate	–0.9336	–1.0219	–0.0883	–1.0213	–0.0877
13	2-Butynyl-acetate	–0.8834	–0.9355	–0.0521	–0.9029	–0.0195
14	Allyl propionate	–0.8791	–0.8433	0.0358	–0.8546	0.0245
15	Vinyl acetate	–0.8595	–0.9203	–0.0608	–0.9553	–0.0958
16	Methyl valerate	–0.8448	–0.9251	–0.0803	–0.8212	0.0236
17	Propyl propionate	–0.8148	–0.801	0.0138	–0.7775	0.0373
18*	n-Amyl formate	–0.7826	–0.7211	0.0615	–0.7747	0.0079
19	Ethyl isovalerate	–0.7231	–0.7248	–0.0017	–0.6337	0.0894
20	Isobutyl propionate	–0.6935	–0.682	0.0115	–0.7108	–0.0173
21	sec-Butyl acetate	–0.6794	–0.6654	0.0140	–0.6543	0.0251
22	Propargyl propionate	–0.6554	–0.6309	0.0245	–0.5868	0.0686
23	Vinyl propionate	–0.6530	–0.6674	–0.0144	–0.6660	–0.0130
24*	Allyl butyrate	–0.6355	–0.6213	0.0142	–0.6575	–0.0220
25	Methyl hexanoate	–0.5611	–0.6319	–0.0708	–0.5988	–0.0377
26	Ethyl butyrate	–0.4903	–0.4893	0.0010	–0.5039	–0.0136
27	Butyl acetate	–0.4864	–0.5145	–0.0281	–0.5736	–0.0872
28	Propyl butyrate	–0.4138	–0.4264	–0.0126	–0.4703	–0.0565
29	Tert butyl propionate	–0.4095	–0.4075	0.0020	–0.4722	–0.0627
30*	Vinyl butyrate	–0.3825	–0.3511	0.0314	–0.3094	0.0731
31	n-Hexyl formate	–0.3824	–0.3565	0.0259	–0.3952	–0.0128
32	Ethyl valerate	–0.3580	–0.3624	–0.0044	–0.4302	–0.0722
33	2-Ethylbutyl acetate	–0.1202	–0.1015	0.0187	–0.1263	–0.0061
34	Amyl propionate	–0.0431	–0.0526	–0.0095	–0.0444	–0.0013
35	Hexyl acetate	–0.0087	–0.0081	0.0006	–0.0144	–0.0057
36*	Propyl valerate	0.0094	0.0230	0.0136	–0.0083	–0.0177
37	Ethyl hexanoate	0.0637	0.0715	0.0078	0.0607	–0.0030
38	Methyl heptanoate	0.1039	0.0982	–0.0057	0.1446	0.0407
39	Allyl hexanoate	0.2128	0.2714	0.0586	0.2558	0.0430
40	Methyl octanoate	0.5358	0.5063	–0.0295	0.5348	–0.0010

To be continued

41	Allyl heptanoate	0.7282	0.8174	0.0892	0.8435	0.1153
42*	Methyl nonanoate	1.0419	1.1275	0.0856	1.1337	0.0918
43	Vinyl 2-ethylhexanoate	1.0462	0.9630	–0.0832	0.9030	–0.1432
44	Octyl acetate	1.0570	1.0820	0.0250	1.1547	0.0977
45	Tert butyl formate	1.3719	1.3398	–0.0321	1.4095	0.0376
46	Methyl decanoate	1.3778	1.3225	–0.0553	1.2409	–0.1369
47	Methyl undecanoate	1.4248	1.5058	0.0810	1.4757	0.0509
48*	Decyl acetate	1.8794	1.8081	–0.0713	1.7688	–0.1106

2. 2 Experimental methods

2. 2. 1 Characterization of the compound structure

The 3*D* holographic vector of atomic interaction field (3*D*-HoVAIF)[13-15] started from the two spatial invariants of the three-dimensional structures of molecules—the relative distance of atoms and the properties of the atoms themselves based on three classical non-bonding interaction modes between atoms, such as electrostatic, stereo and hydrophobic interactions. It provided three-dimensional vector descriptors for characterizing the molecular structures of compounds without any experimental parameters. The molecules of common organic compounds usually include hydrogen, carbon, nitrogen, phosphorus, oxygen, sulfur, fluorine, chlorine, bromine and iodine. They belong to five main groups in the periodic table, such as IA, IVA, VA, VIA and VIIA. Based on this, these atoms could be divided into 5 categories. At the same time, in order to characterize the microenvironment of the molecular structure more accurately, according to the above classification, the atoms in different main groups were further subdivided into 10 categories according to their hybrid state (1.H, 2. $C_{(sp^3)}$, 3. $C_{(sp^2)}$, 4. $C_{(sp)}$, 5. $N_{(sp^3)}$, $P_{(sp^3)}$, 6. $N_{(sp^2)}$, $P_{(sp^2)}$, 7. $N_{(sp)}$, $P_{(sp)}$, 8. $O_{(sp^3)}$, $S_{(sp^3)}$, 9. $O_{(sp^2)}$, $S_{(sp^2)}$ and 10. F, Cl, Br, I). The interaction between various atoms in a compound molecule could be up to 10×(10+1)/2 = 55 items. 3*D*-HoVAIF used three potential energies (electrostatic, stereo and hydrophobic) to express different forms of action. Therefore, for an organic compound molecule, there were at most 3×55 = 165 atomic action terms to characterize the molecular structure information. Although the atomic interaction mode in 3*D*-HoVAIF was not a direct manifestation of the compound, in most cases, the 3*D*-HoVAIF descriptors contained a wealth of information on the potential energy distribution of organic compounds, which could well characterize the microenvironment of the molecules.

2. 2. 1. 1 Electrostatic interaction

The electrical effect of atoms is proportional to the charge and inversely proportional to the distance between atoms. As an important form of non-bonding interaction, electrostatic interaction was expressed by the classic Coulomb theorem (Eq. (1)). Among them, r_{ij} (nm) was the Euclid distance between atoms; e was the unit charge amount of 1.6021892 × 10^{-19} C; ε_0 was the dielectric constant in vacuum 8.85418782 × 10^{-12} C^2/J m; Z was the net charge of the atom, the electron as the unit; m and n were the types of atoms. The electrostatic potential between all atoms in the molecule was calculated by this formula, and count into 55 electrostatic interaction terms according to their type.

$$EE(m-n)=\sum_{i\in m,\,j\in n}\frac{e^2}{4\pi\varepsilon_0}\cdot\frac{Z_i\cdot Z_j}{r_{ij}} \qquad (1\leqslant m\leqslant 10,\leqslant n\leqslant 10) \tag{1}$$

2. 2. 1. 2 Steric interaction

The steric interaction is the nondipole-dipole or dipole induced interaction between atoms in space. The Lennard-Jones equation was used to describe this mode of action (Eq. (2)). In the formula, $\varepsilon_{ij}=(\varepsilon_{ii}\cdot\varepsilon_{jj})^{1/2}$ was the depth of the atom-pair potential energy well, which was taken from the literature[16]; D was the empirically derived interatomic interaction energy correction constant taken as 0.01[17]; $R_{ij}^{*}=(C_h R_{ii}^{*}+ C_h R_{jj}^{*})/2$, which was the corrected atom pair van der Waals radius, the correction factor C_h was 1.00 of sp^3 hybridization, 0.95 of sp^2 hybridization, and 0.90 of sp hybridization[17].

$$ES(m-n)=\sum_{i\in m,\,j\in n}\varepsilon_{ij}\cdot D\cdot\left[\left(\frac{R_{ij}^{*}}{r_{ij}}\right)^{12}-2\cdot\left(\frac{R_{ij}^{*}}{r_{ij}}\right)^{6}\right]$$

$$(1\leqslant m\ 10,\ m\leqslant n\leqslant 10) \tag{2}$$

2. 2. 1. 3 Hydrophobic interaction

Hydrophobic interaction is one of the factors that affect on the properties of compounds. Considering that the 3*D*-HoVAIF descriptors required to express the interaction between atoms in the molecule, the hint method proposed by Kellogg *et al*.[18-22] was used to express this type of potential field. A simple expression for calculating the hydrophobic interaction between two atoms was defined in the hint (Eq. (3)). In the formula, S was the solvent accessible surface area

of atom (SASA), which was the surface area formed by water molecules (van der Waals radius of 0.14 nm) as the probe rolls its sphere on the surface of the atom[23]. T was a binary discriminant function of the action form to indicate the direction of the entropy effect of the hydrophobic interaction of different types of atoms[18-22], and a was the atomic hydrophobicity constant, taking the literature value[24].

$$EH(m-n)=10^{-3}\sum_{i\in m,\,j\in n}S_i\cdot a_i\cdot S_j\cdot a_j\cdot e^{-r_{ij}}\cdot T_{ij}$$

$$(1\leqslant m\leqslant 10,\ m\leqslant n\leqslant 10) \qquad (3)$$

The Chemoffice 2006 was used to construct the molecular three-dimensional structures of the studied samples, and the MOPAC semi-empirical quantum chemistry software that comes with Chem3*D* was used to optimize the molecular structures and get the position coordinates of the atoms in the molecules at the AM1 level, and the Mulliken layout analysis method was used to calculate the net charge e of the atom in a single-point form (*e.g.*, ethyl acetate dimensional structure is shown in Fig. 1. The position coordinates of each atom and the net charge quantity e are shown in Table 2). The space position coordinates of each atom in the molecule were used to calculate the distance r_{ij} between atoms, and finally the 165 3*D*-HoVAIF descriptors were obtained by formulas (1), (2) and (3).

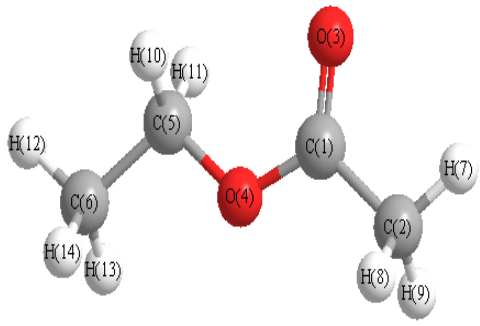

Fig. 1. Three-dimensional structure diagram of ethyl acetate

Table 2. Partial Charges and Coordinates of Each Atom of Ethyl Acetate

Atom	*e*	*x*	*y*	*z*
C(1)	0.355728	0.2291	–1.2842	–0.0000
C(2)	–0.357127	1.5360	–2.0388	–0.0000
O(3)	–0.397118	–0.8170	–1.8882	–0.0000
O(4)	–0.322514	0.2292	0.0538	–0.0000
C(5)	–0.098051	–0.9849	0.7548	–0.0000
C(6)	–0.35570	–0.7074	2.2523	–0.0000
H(7)	0.152265	2.3823	–1.3160	–0.0000
H(8)	0.170018	1.5947	–2.6780	0.9092
H(9)	0.170076	1.5948	–2.6787	–0.9088
H(10)	0.137276	–1.5679	0.4860	0.9092
H(11)	0.137362	–1.5684	0.4858	–0.9088
H(12)	0.137378	–1.6713	2.8088	–0.0000
H(13)	0.135164	–0.1245	2.5210	–0.9093
H(14)	0.135243	–0.1239	2.5213	0.9088

2. 2. 2 Modeling and evaluation

The stepwise regression (SMR) is a commonly used method for variable screening, so it was used to screen the original descriptors. Multiple linear regression (MLR) and partial least-squares regression (PLS) are commonly used methods for modeling, and therefore multiple linear regression (MLR) and partial least-squares regression (PLS) were used to build models. An excellent model must meet the following requirements: 1) Modeling correlation coefficient (R^2) $\geqslant$ 0.81, "Leave one method" cross-test correlation coefficient (R_{CV}^2) $\geqslant$ 0.64 and external prediction correlation coefficient (R_{test}^2) $\geqslant$ 0.64, which are all higher than the standards mentioned in the literature[25]; 2) The ratio of various standard deviations (*SD*) to the value range (V_r) should be less than or equal to 10%[26]; 3) The absolute value of the prediction error for above 80% samples should be less than or equal to 2 times that of the standard deviation (2*SD*). The external prediction correlation coefficient (R_{test}^2) and

standard deviation (SD_{test}) were calculated according to Eqs. (4) and (5), respectively.

$$R_{test}^{2} = 1 - \frac{\sum_{i=1}^{test}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{test}(y_i - \bar{y}_i)^2} \quad (4)$$

$$SD_{test} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{test}(y_i - \hat{y}_i)^2} \quad (5)$$

In equations (4) and (5), both y_i and $\hat{y}_i$ were the experimental and predicted values of the test set samples, respectively. $\bar{y}_i$ was the average of the experimental values of the test set samples.

3 RESULTS AND DISCUSSION

The research samples contained only six types of atoms: H, $C_{(sp^3)}$, $C_{(sp^2)}$, $C_{(sp)}$, $O_{(sp^3)}$, and $O_{(sp^2)}$, thereby producing a total of 63 structural descriptors, including 21 electrostatic interaction terms, 21 stereoscopic interaction terms, and 21 hydrophobic interaction terms. Because there were too many structural descriptors, some structural descriptors may have little correlation with compound toxicity, so it was necessary to screen variables before modeling. The stepwise regression was used to screen variables which were introduced into the model for significance. By observing the changes of model correlation coefficient (R^2), standard deviation (SD), cross-test correlation coefficient (R_{CV}^2), and standard deviation (SD_{CV}), we selected the best combination of variables to build the model. When 7 variables x_1, x_{18}, x_{33}, x_{72}, x_{80}, x_{118} and x_{127} (listed in Table 3) were selected, the correlation coefficient (R^2), standard deviation (SD), cross-test correlation coefficient (R_{CV}^2) and standard deviation (SD_{CV}) achieved ideal values at the same time. Among the selected variables, x_1, x_{18} and x_{33} were electrostatic interaction terms, x_{72} and x_{80} were steric interaction terms, and x_{118} and x_{127} were hydrophobic interaction terms.

Table 3. Structural Descriptors Selected out by SMR for Modeling

No.	x_1	x_{18}	x_{33}	x_{72}	x_{80}	x_{118}	x_{127}
1	0.2630	0.3081	0.0000	9.4512	13.3881	2.5058	–39.5061
2	0.1657	0.1615	0.0000	9.4438	13.0792	1.8963	–38.3331
3	0.0698	0.0245	0.0000	9.4438	12.1325	1.6753	–49.8270
4	0.4206	0.2568	0.0000	9.4469	15.3234	1.6745	–49.7144
5	0.2680	0.1514	0.0000	9.4438	14.3071	1.6298	–45.1418
6*	0.4174	0.1914	0.0000	9.4469	19.9730	2.1067	–49.8103
7	0.3896	0.2575	0.0000	9.4438	14.5846	2.0875	–48.7584
8	0.1927	0.1684	0.0512	9.4438	10.3336	1.6976	–8.5252
9	0.5591	0.3334	0.0000	9.4395	19.2187	2.1008	–49.4027
10	0.2594	0.0676	0.0000	9.4469	13.5881	1.9412	–34.6710
11	0.4152	0.1653	0.0000	-0.0003	5.3267	2.1228	–61.1421
12*	0.3820	0.1936	0.0000	9.4469	13.7263	2.0957	–52.7547
13	0.1583	0.1061	0.0845	6.8120	3.1219	0.8706	–29.4123
14	0.4586	0.1943	0.0000	9.4457	14.9213	2.1556	–64.1180
15	0.2164	0.0710	0.0000	17.5580	9.1665	0.8051	–45.6454
16	0.5381	0.3592	0.0000	9.4438	14.0289	2.6582	–44.0449
17	0.5532	0.2670	0.0000	9.4469	14.8435	2.0998	–49.3409
18*	0.5207	0.1782	0.0000	9.4938	16.7599	1.9268	–10.8316
19	0.6781	0.3838	0.0000	9.4438	15.2826	2.4261	–51.0394
20	0.6598	0.2529	0.0000	9.4438	16.5778	1.5201	–31.6317
21	0.6572	0.2955	0.0000	9.4395	17.1701	2.1375	–40.1943
22	0.3111	0.1929	0.0814	9.4469	10.5921	1.9374	–38.9479
23	0.3240	0.1195	0.0000	17.5559	9.7247	1.1304	–5.9053
24*	0.6202	0.2145	0.0000	9.4457	15.9782	1.6753	–49.8270
25	0.6974	0.3212	0.0000	9.4438	15.6485	1.9338	–38.8962
26	0.5271	0.1581	0.0000	9.4438	12.1632	1.9373	–38.9466
27	0.5677	0.1454	0.0000	9.4438	14.7229	1.6755	–49.8563

To be continued

28	0.7204	0.2281	0.0000	9.4395	16.9000	1.6298	–45.1417
29	0.7477	0.2267	0.0000	9.4438	18.8592	1.6757	–49.8691
30*	0.4712	0.1849	0.0000	17.5606	10.7128	1.7189	–62.1318
31	0.6710	0.2094	0.0000	9.4469	17.7720	2.2757	–50.5241
32	0.7266	0.2111	0.0000	9.4395	18.2193	1.7073	–8.5694
33	0.8912	0.2863	0.0000	9.4438	17.5452	1.7101	–8.5826
34	0.8137	0.2597	0.0000	9.4438	14.9115	2.1008	–49.3987
35	0.8998	0.2751	0.0000	9.4469	14.7182	1.5201	–31.6316
36*	0.8989	0.2102	0.0000	9.4438	19.7167	1.6299	–45.1410
37	0.8239	0.2281	0.0000	9.4438	14.7305	2.0523	–35.8980
38	0.8415	0.2828	0.0000	9.4438	13.5559	2.4284	–53.8515
39	0.9012	0.2218	0.0000	9.4499	13.0857	1.6625	–8.3638
40	1.0125	0.2792	0.0000	9.4438	13.5567	2.1621	–60.6598
41	1.0157	0.2399	0.0000	9.4499	13.5891	2.7693	–45.2683
42*	1.2021	0.2633	0.0000	9.4438	15.6672	2.6267	–83.5691
43	1.4141	0.3793	0.0000	9.4438	26.0487	2.3941	–14.7822
44	1.2429	0.3617	0.0000	9.4469	15.9724	3.0923	–63.8925
45	0.7323	0.1013	0.0000	9.4438	13.8681	5.9578	–68.1012
46	1.5014	0.2745	0.0000	9.4438	15.6676	0.3897	–18.0973
47	1.5707	0.3163	0.0000	9.4438	13.5576	0.5714	–44.5666
48*	1.7261	0.3412	0.0000	9.4469	14.7726	0.4994	–31.6074

7-variable multiple linear regression model (M1), as in Eq. (6).

$$\log(1/IGC_{50}) = -2.0424 + 2.7984x_1 - 3.0006x_{18} + 6.2540x_{33} + 0.0491x_{72} - 0.0369x_{80} + 0.2897x_{118} + 0.0006x_{127} \quad (6)$$

$N = 40$, $R_1^2 = 0.9974$, $SD_1 = 0.0469$, $F_1 = 1748.8673$; $R_{CV1}^2 = 0.9939$, $SD_{CV1} = 0.0715$, $F_{CV1} = 749.2740$; $R_{test1}^2 = 0.9955$, $SD_{test1} = 0.0720$

N was the number of regression points, R_1^2 the correlation coefficient, SD_1 the standard deviation, F_1 the significance test value; R_{CV1}^2 the correlation coefficient of the cross-test, SD_{CV1} the standard deviation of the cross-test, F_{CV1} the significance test value of the cross-test, R_{test1}^2 the external test correlation and SD_{test1} the standard deviation of the external test. The correlation coefficient (R_1^2) of the above model was as high as 0.9974, much greater than the 0.81 standard, indicating that the model fit well; the value range (V_r) of the research samples was 1.8794 – (–1.6092) = 3.4886, and the standard deviation (SD_1) was 0.0469, (0.0469/3.4886) × 100% = 1.3444%, much lower than the 10% standard, which meat the model fitting errors were small. The cross-test correlation coefficient (R_{CV1}^2) was 0.9939 and much larger than the 0.64 standard; the cross-test standard deviation (SD_{CV1}) was 0.0715, (0.0715/3.4886) × 100% = 2.0495%, which was much lower than the 10% standard, suggesting that the model was stable. The external test correlation coefficient (R_{test1}^2) was 0.9955 and much greater than 0.64; the external test standard deviation (SD_{test1}) was 0.0720, (0.0720/3.4886) × 100% = 2.0639%, which was greatly lower than the 10% standard, indicating strong predictive ability and small prediction errors of the model.

In order to further understand the influence of variables on compound toxicity, the structural descriptors in Table 2 were used as the independent variables X, and the compound toxicity value $\log(1/IGC_{50})$ as the dependent variable Y. The partial least-squares regression was used to establish a model (M2). The change of the correlation coefficients (R^2/R_{CV}^2) with the number of principal components is shown in Fig. 2. When the number of the principal components reached 3, the correlation coefficient (R^2) of the model got the maximum value, and the cross-test correlation coefficient (R_{CV}^2) was close to the maximum value. Thereafter, 3 principal components were chosen to build the model.

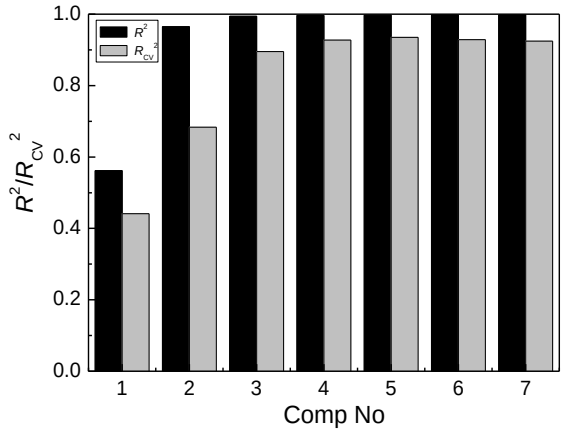

Fig. 2. Correlation coefficient (R^2/R_{CV}^2) changes with the number of principal components

The distribution of the scores of the 40 training set samples in the top 2 principal components of the PLS space is plotted in Fig. 3. The scores of most of the studied samples (97.5%) fell within the 95% confidence elliptical confidence circle. There was only one abnormal point (compound No. 13), which reflected that the structural descriptors could represent the molecular structure characteristics of ester compounds and got the correct performance in the statistical model. The abnormal point in Fig. 3 is compound No. 13 "2-butynyl-acetate", which contained a "triple bond" and had a certain degree of particularity.

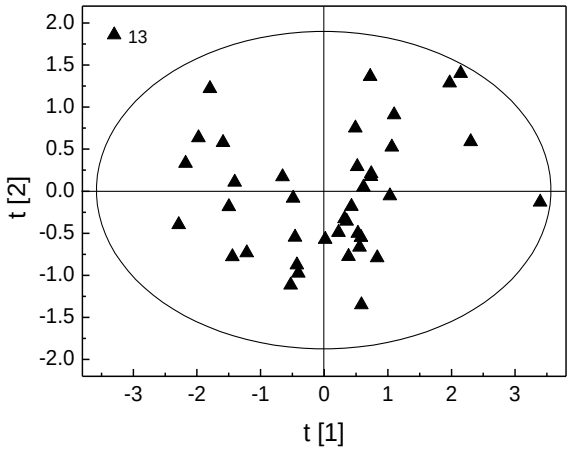

Fig. 3. PLS scores of samples in the top 2 principal components

At this time, the model's R_2^2 = 0.9940, SD_2 = 0.0646; R_{CV2}^2 = 0.8952, SD_{CV2} = 0.0925; R_{test2}^2 = 0.9955, SD_{test2} = 0.0716. The correlation coefficient (R_2^2) of the model was as high as 0.9940, which was much larger than the 0.81 standard, indicating that the model fit well; the standard deviation (SD_2) was 0.0646, (0.0646/3.4886) × 100% = 1.8517%, which was much lower than the 10% standard, so the model fitting errors were small. The cross test correlation coefficient (R_{CV2}^2), 0.8952, was much larger than the 0.64 standard; the cross test standard deviation (SD_{CV2}) was 0.0925, (0.0925/3.4886) ×100% = 2.6515% and greatly lower than the 10% standard, which suggested stability for the model. The external test correlation coefficient (R_{test2}^2) of 0.9955 was remarkably greater than 0.64; the external test standard deviation (SD_{test2}) was 0.0716, (0.0716/3.4886) × 100% = 2.0524%. It was significantly lower than the 10% standard, also showing that the model had strong predictive ability and the prediction errors were small.

In order to verify whether the excellent model results were accidental, the model was verified by random sorting of the *Y* vector 20 times. The correlation coefficients of the *Y* original vector and the randomly sorted *Y* vector are plotted on the model R^2 and R_{CV}^2 in Fig. 4. According to the judgment criteria proposed by Andersson et al.[27], the intercepts of R^2 and R_{CV}^2 on the vertical axis should not exceed 0.300 and 0.050, respectively. From Fig. 4, it can be found that the

intercepts of R^2 and $R_{CV}{}^2$ of the PLS model built in this paper were 0.072 and –0.400, respectively. Therefore, it could be considered that the excellent results of the model built in this paper were not accidental, so our model could be used to analyze the structures of ester compounds.

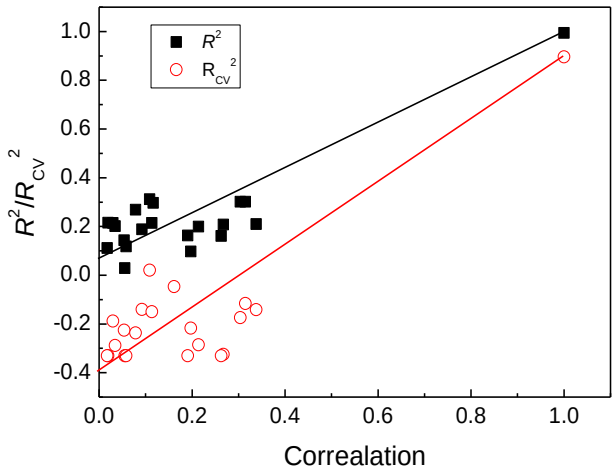

Fig. 4. Plot of *Y* random permutations test

In order to further study the influence of each variable on the compound toxicity log(1/IGC_{50})(*Y*), the load distribution of the samples in PLS is plotted in Fig. 5, in which x_1, x_{72}, x_{118}, and x_{127} are in the upper right. It means that they are positively correlated with *Y* in the first and second principal components, and the distance between x_1 and the origin is relatively large, which reflects that it has a relatively larger correlation with *Y*. x_{18} and x_{80} are at the bottom right of the figure, indicating that they are positively correlated with *Y* in the first principal component, and negatively correlated with *Y* in the second principal component. x_{33} is at the upper left of the figure, which suggests that it is negatively correlated with *Y* in the first principal component, and positively correlated with *Y* in the second principal component.

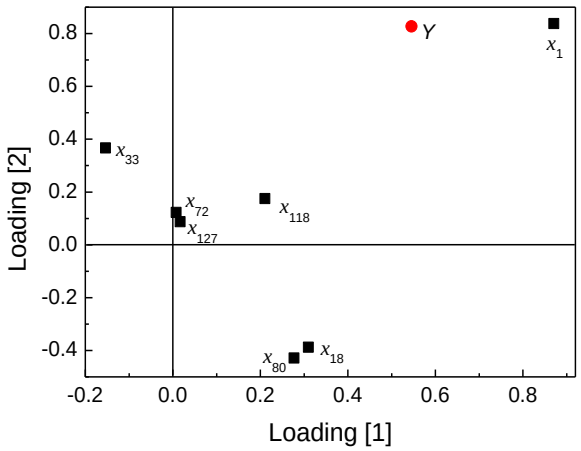

Fig. 5. Plot of PLS loadings plot of the samples

The importance of a variable can reflect the degree of correlation between the variable and *Y*. It is generally considered that variables with variable importance projection (VIP) values greater than 1 are highly correlated with the toxicity log(1/IGC_{50}) of ester compounds. The variable importance projection is shown in Fig. 6. Fig. 6 shows that the VIP values of the three variables x_1, x_{18}, and x_{80} were greater than 1, indicated that these three variables were highly correlated with the toxicity log(1/IGC_{50}) of ester compounds. x_1 corresponding to the electrostatic interaction of hydrogen atoms, described that the more hydrogen atoms in the compound, the higher the toxicity log(1/IGC_{50}) value of the ester compound may be. x_{18} corresponding to the electrostatic effect of $C_{(sp^3)}$ and $O_{(sp^2)}$, and x_{80} corresponding to the stereoscopic interaction effect of $C_{(sp^2)}$ and $O_{(sp^3)}$. The above shows that oxygen atoms had a greater influence on the toxicity value.

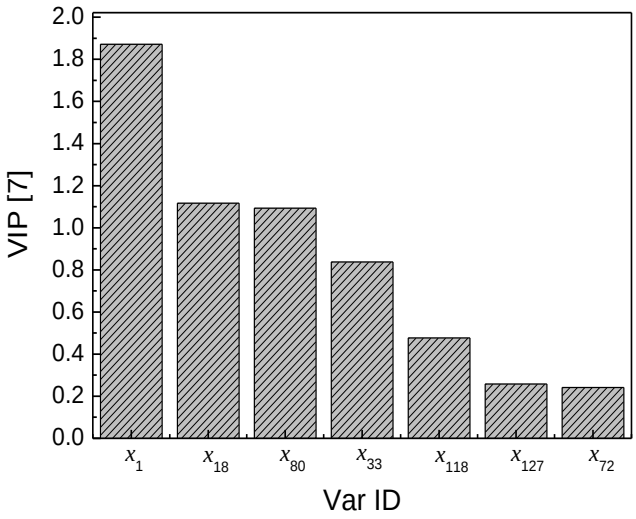

Fig. 6. Variable importance projection

The calculated values of the toxicity log(1/IGC_{50}) of the two models for the compounds are listed in Table 1 as Cal.1 and Cal.2, Err.1 and Err.2 are the errors, respectively. For the convenience of observation, the correlation between the calculated log(1/IGC_{50}) of the model's toxicity to the compound and the experimental values is plotted in Fig. 7, and the corresponding errors are plotted in Fig. 8. Fig. 7 shows that most of the sample points were near the 45° diagonal, indicating that the calculated values of the model's toxicity log(1/IGC_{50}) for the compounds were highly correlated with the experimental values. The two values were close in size. The toxicity log(1/IGC_{50}) could be predicted accurately, which once again showed the model's good predictive ability and excellent predictive results.

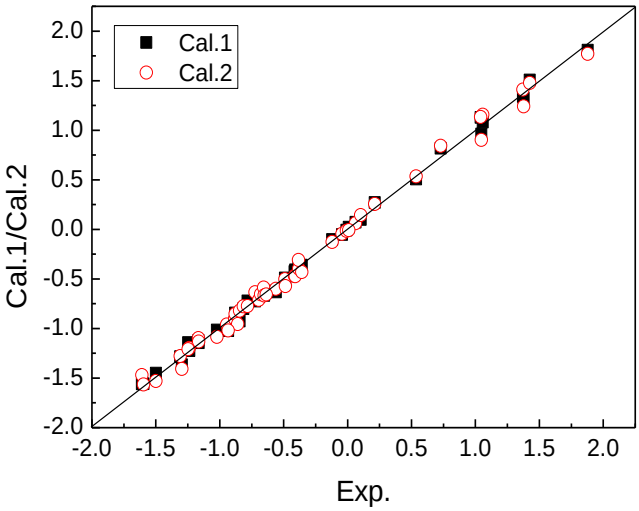

Fig. 7. Plot of the predicted values *vs*. the experimental ones

A good prediction model usually requires the prediction errors of most samples not exceeding plus or minus 2 times that of the standard deviation (*ie* ± 2*SD*). It can be found in Fig. 8 that most of the samples' errors were within ± 2*SD* of the model. For model M1, only 1 sample (No. 1) had a prediction error exceeding ± 2SD_1; for model M2, only 3 samples (Nos. 1, 43, 46) had prediction errors larger than ± 2SD_2. This shows that the model was accurate in predicting the toxicity log(1/IGC_{50}) of the compounds, and the prediction errors were in an acceptable range. The model could be used to predict the toxicity log(1/IGC_{50}) of ester compounds. At the same time, the existence of large error samples indicated that some special structural information of compounds had not been fully expressed, and the molecular structure characterization method needed further improvement.

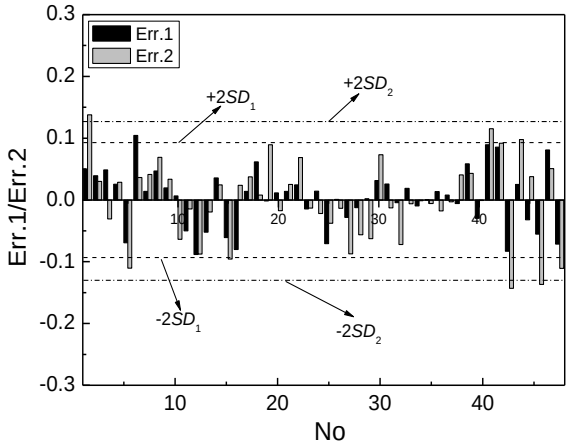

Fig. 8. Plot of the predicted residuals scattered

4 CONCLUSION

By classifying the atoms in the compound, the electrostatic interaction, steric interaction and hydrophobic interaction between the atoms were calculated as structural descriptors on the three-dimensional structure of the compound, and then the structures of 48 ester compounds were expressed parametrically. The relationship models between compound structures and toxicity $\log(1/IGC_{50})$ were established through multiple linear regression (MLR) and partial least-squares regression (PLS), and it was found that the toxicity of ester compounds $\log(1/IGC_{50})$ was closely related to the molecular structures of the compounds. The constructed structure-toxicity $\log(1/IGC_{50})$ relationship models can be used to predict the toxicity $\log(1/IGC_{50})$ of ester compounds. Due to the slightly larger prediction errors of individual samples, there is still a lot of room for improvement in the molecular structure characterization method, and related researches are underway. This paper has certain reference value for the quantitative structure-toxicity relationship study of toxic compounds in environment.

REFERENCES

(1) Lin, L.; Dong, L.; Meng, X. Y.; Li, Q. Y.; Huang, Z.; Li, C.; Li, R.; Yang, W. J.; Crittenden, J. Distribution and sources of polycyclic aromatic hydrocarbons and phthalic acid esters in water and surface sediment from the Three Gorges Reservoir. *J. Environ. Sci.* **2018**, 69, 271–280.

(2) Sun, Y. S.; Zhang, K. Y.; Liu, Y. Determination of ester compounds in water by purge and trap/gas chromatography-mass spectrometry. *Inner Mongolia Environ. Sci.* **2019**, 31, 85–86.

(3) Zhu, Y. Y. Determination of acetic acid esters in surface water by headspace gas chromatography. *Environ. Monit. Forewarning* **2017**, 9, 32–34.

(4) Li, J. F.; Liao, L. M. QSAR study on acute toxicity of benzene ring-containing compounds to luminescent bacteria. *Environ. Sci. Technol.* **2020**, 43, 14–19.

(5) Chi, Y. L.; Zhang, H. T.; Huang, Q. S.; Lin, Y.; Ye, G. Z.; Zhu, H. M. Dong, S. J. Environmental risk assessment of selected organic chemicals based on TOC test and QSAR estimation models. *J. Environ. Sci.* **2018**, 64, 23–31.

(6) Chourasiaa, S. S.; Kharkateb, S. K.; Kosec, T. D. Synthesis, QSAR modeling and antimicrobial studies of 1-(4-phenyl) substituted tetrahydro isoquinoline derivatives. *Mater. Today: Proc.* **2020**, 29, 956–963.

(7) Bitam, S.; Hamadache, M.; Salah, H. 2*D* QSAR studies on a series of (4S,5R)-5-[3,5-bis(trifluoromethyl)phenyl]-4-methyl-1,3-oxazolidin-2-one as CETP inhibitors. *SAR QSAR Environ. Res.* **2020**, 31, 423–438.

(8) Tawassl, T. H. H.; Abu Baker, M. O.; Ahmed, E. M. S. 2*D*-QSAR modeling and molecular docking studies on 1H-pyrazole-1-carbothioamide derivatives as EGFR kinase inhibitors. *ACS Omega* **2020**, 5, 18662−18674.

(9) Shu, M.; Wu, T.; Wang, B. W.; Li, J.; Xu, C. M.; Lin, Z. H. 3*D*-QSAR and Surflex docking studies of a series of alkaline phosphatase inhibitors. *Chin. J. Struct. Chem*. **2019**, 38, 7−16.

(10) Liu, W. G.; Hong, D. F.; Xi, C. C.; Ma, C.; Xiong, F.; Zhang, S. P. 3*D*-QSAR analysis of a series of dihydroquinolizinone derivatives as a hepatitis B virus expression inhibitor. *Chin. J. Struct. Chem*. **2020**, 39, 1615−1626.

(11) Shi, J. C.; Huang, X. Q.; Luo, M.; Huang, C. S. Identification of novel and potent curcuminoids inhibitors of tubulin with anticancer activities by 3*D*-QSAR and molecular docking. *Chin. J. Struct. Chem*. **2020**, 39, 1157−1166.

(12) http://www.vet.utk.edu/TETRaTOX/

(13) Li, M. P.; Zhang, S. W. Quantitative structure activity relationship studies of the herbicidal activity for 1,3,5-triazine derivatives by 3*D*-HoVAIF. *J. At. Mol. Phys.* **2017**, 34, 997−1002.

(14) Tong, J. B.; Li, K. N.; Wu, Y. J.; Zhan, P. QSAR studies of bitter tasting thresholds by 3*D*-HoVAIF. *J. At. Mol. Phys.* **2017**, 34, 155−160.

(15) Tong, J. B.; Zhong, L.; Zhao, X.; Liu, S. L.; Wang, P. Quantitative structure-activity relationship studies of diarylpyrimidine derivatives as anti-HIV drugs using new three-dimensional structure descriptors. *Med. Chem. Res.* **2014**, 23, 4883−4892.

(16) Levitt, M.; Perutz, M. F. Aromatic rings act as hydrogen bond acceptors. *J. Mol. Biol*. **1988**, 201, 751−754.

(17) Hahn, M.; Receptor surface models. 1. Definition and construction. *J. Med. Chem.* **1995**, 38, 2080−2090.

(18) Kellogg, G. E.; Semus, S. F.; Abraham, D. J. HINT-a new method of empirical hydrophobic field calculation for CoMFA. *J. Comput. Aided Mol. Des*. **1991**, 5, 545−552.

(19) Wireko, F. C.; Kellogg, G. E.; Abraham, D. J. Allosteric modifiers of hemoglobin. 2. Crystallographically determined binding sites and hydrophobic binding/interaction analysis of novel hemoglobin oxygen effectors. *J. Med. Chem*. **1991**, 34, 758−767.

(20) Kellogg, G. E.; Joshi, G. S.; Abraham, D. J. New tools for modeling and understanding hydrophobicity and hydrophobic interactions. *Med. Chem. Res*. **1992**, 1, 444−453.

(21) Kellogg, G. E.; Abraham, D. J. KEY, LOCK and LOCKSMITH: complementary hydrophobicity map predictions of drug structure from a known receptor-receptor structure from known drugs. *J. Mol. Graph.* **1992**, 10, 212−217.

(22) Nayak, V. R.; Kellogg, G. E. Cyclodextrin-barbiturate inclusion complexes: a CoMFA/HINT 3-*D* QSAR Study. *Med. Chem. Res.* **1994**, 3, 491−502.

(23) Hasel, W.; Hendrikson, T. F.; Still, W. C. A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahed. Comp. Method* **1988**, 1, 103−116.

(24) Pei, J.; Wang, Q.; Zhou, J.; Lai, L. Estimating protein-ligand binding free energy: atomic solvation parameters for partition coefficient and solvation free energy calculation. *Proteins* **2004**, 57, 651−664.

(25) Gu, Y. L.; Chen, X.; Jian, M. L. Study on the structure-toxicity relationship of aniline compounds by density functional theory. *Chem. Res. Appl*. **2015**, 27, 1139−1144.

(26) Sung, S. S.; Kaprlus, M. A comparative study of ligand-receptor complex binding affinity prediction methods based on glycogen phosphorylase inhibitors. *J. Compu. Aid. Mol. Des.* **1999**, 13, 243−258.

(27) Andersson, P. M.; Sjstrom, M.; Lundstedt, T. Preprocessing peptides sequences for multivariate sequence-property analysis. *Chemom. Intell. Lab. Syst.* **1998**, 42, 41−50.