Eco-Environmental Knowledge Web

生态毒理学报
Asian Journal of Ecotoxicology

第 14 卷 第 4 期 2019 年 8 月
Vol. 14, No.4 Aug. 2019

DOI: 10.7524/AJE.1673-5897.20190409002
席越, 杨先海, 张红雨, 等. 基于形态修正的描述符构建可电离化合物对大型溞急性毒性的 QSAR 模型[J]. 生态毒理学报,2019, 14(4): 183-191
Xi Y, Yang X H, Zhang H Y, et al. Development of acute toxicity of *Daphnia magna* QSAR models for ionogenic organic chemicals based on chemical form adjusted descriptors [J]. Asian Journal of Ecotoxicology, 2019, 14(4): 183-191 (in Chinese)

基于形态修正的描述符构建可电离化合物对大型溞急性毒性的 QSAR 模型

席越，杨先海*，张红雨，刘会会

南京理工大学环境与生物工程学院,江苏省化工污染控制与资源化高校重点实验室,南京 210094

收稿日期:2019-04-09　　**录用日期**:2019-05-28

摘要：在环境水体中,可电离有机化合物(IOCs)可解离为分子和离子形态。研究表明,IOCs 离子形态的环境行为、毒性效应等都与其分子形态存在较大差异,因而在研究 IOCs 环境行为、毒性效应时不应忽略离子化的影响。在构建 IOCs 相关预测模型时如何表征离子化的影响是当前研究的重要内容之一。探讨了采用基于形态修正的描述符构建 IOCs 水生毒性预测模型的可行性。具体而言,采用逐步多元线性回归(MLR)方法,构建了可预测 63 种取代酚、取代苯甲酸和取代苯胺等 IOCs 对大型溞急性毒性的定量结构-活性关系(QSAR)模型。与仅采用分子形态描述符的模型相比,使用基于形态修正描述符的模型决定系数(R^2)、去一法交叉验证系数(Q^2_{LOO})、外部验证系数(Q^2_{EXT})等参数从 0.622 ~ 0.705 提高到了 0.840 ~ 0.875,表明基于形态修正描述符的模型具有更好的拟合优度、稳健性和预测能力。因此,在将来的研究中,可采用基于形态修正的描述符构建 IOCs 水生毒性效应预测模型。

关键词：可电离有机化合物;大型溞;急性毒性;基于形态修正的描述符;定量结构-活性关系
文章编号：1673-5897(2019)4-183-09　　**中图分类号**：X171.5　　**文献标识码**：A

Development of Acute Toxicity of *Daphnia magna* QSAR Models for Ionogenic Organic Chemicals Based on Chemical Form Adjusted Descriptors

Xi Yue, Yang Xianhai* , Zhang Hongyu, Liu Huihui

Jiangsu Key Laboratory of Chemical Pollution Control and Resources Reuse, School of Environmental and Biological Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

Received 9 April 2019　　**accepted** 28 May 2019

Abstract: Ionogenic organic chemicals (IOCs) may ionize to form anion and/or cation in aquatic environment. It had been elucidated that the environmental behavior, toxic effects of ionic form differ greatly from that of neutral form. Thus, ionization is nonnegligible in performing the research of the environmental behavior, toxic effects of IOCs. To date, how to characterize the ionization is one of the critical issues in developing the predictive models for IOCs. Here, the feasibility of using chemical form adjusted descriptors as predictive variable to derive model for the endpoint of aquatic toxicity was studied. In this regard, the acute toxicity data of 63 substituted phenols, anilines and benzoic acids to *Daphnia magna* were collected firstly. Then, the quantitative structure-activity relationship (QSAR) model was developed by stepwise multiple linear regressions (MLR) analysis. The modeling results indica-

基金项目:国家自然科学基金(No.21507038, 41671489, 21507061)
作者简介:席越(1995-),男,工学学士,硕士研究生,研究方向为生态毒理学,E-mail: 1224689052@qq.com
* **通讯作者**(Corresponding author), E-mail: xhyang@njust.edu.cn

ted that the values of coefficient determination (R^2), cross validated Q^2 leave-one-out (Q^2_{LOO}), external validation coefficient (Q^2_{EXT}) for the QSAR model based on chemical form adjusted descriptors was significantly improved from 0.622-0.705 to 0.840-0.875, compared with that of the model constructed from neutral form descriptors only. The results indicated that the QSAR model based on chemical form adjusted descriptors had high goodness-of-fit, robustness, and predictive ability. Thus, the predictive models of IOCs for aquatic toxicity could be developed by employing chemical form adjusted descriptors in the future QSAR modeling.

Keywords: ionogenic organic chemicals; *Daphnia magna*; acute toxicity; chemical form adjusted descriptors; quantitative structure-activity relationship

在商用化学品中,可电离有机化合物(ionogenic organic chemicals, IOCs)往往占有较大比例[1],例如,在欧盟登记注册的 14 万余种化学品中,约 50% 为 IOCs[2];此外,> 60% 的药物为 IOCs[3],大部分个人护理用品也属于 IOCs[4]。随着 IOCs 的大量生产、使用,可能导致 IOCs 通过多种途径进入水环境。据估计,每年约有 3 亿吨合成化学物质被排放进入水环境[5],导致水环境中 IOCs 等化合物的环境检出率和检出浓度越来越高。进入水环境的 IOCs,会对各种水生生物产生持续暴露,进而引发各种生态危害效应。因此,有必要筛选评估具有毒性效应的 IOCs,并对其进行管控,以减少其对水生生物的危害。

在环境水体中,IOCs 会解离为不同比例的分子和离子形态。各形态存在比例取决于 IOCs 自身的酸碱解离常数(pK_a)和环境 pH 条件,一元酸碱的解离程度可采用下式计算:

$$\delta_M = \frac{1}{1 + 10^{(pH - pK_a)\cdot I_{ab}}},$$

$$\delta_I = \frac{10^{(pH - pK_a)\cdot I_{ab}}}{1 + 10^{(pH - pK_a)\cdot I_{ab}}} \quad (1)$$

式中,δ_M 和 δ_I 分别是分子和离子态的比例分数;酸碱化合物的 I_{ab} 分别取值 1 和−1。前人研究结果表明,化合物的分子和离子态具有不同的环境行为、生物富集能力和毒性效应。例如,在 IOCs 对大型溞的毒性研究中,发现随着溶液 pH 的增加,酚类、苯甲酸类 IOCs 对大型溞的毒性作用降低,而苯胺类 IOCs 对大型溞的毒性效应则相反,说明酸碱化合物的分子态具有更强的水生急性毒性[6]。而在 IOCs 与运甲状腺素转运蛋白的相互作用过程中,IOCs 的离子态具有更重要的贡献[7-8]。因此,在研究 IOCs 的环境行为、健康与生态毒性效应时,不能忽视离子化的影响。

虽然各种健康和生态毒性效应测试体系已建立数十年,但由于实验成本高、耗时长,难以对所有 14 万多种商用化学品进行一一测试,导致仅有少部分化合物具有完整的毒性数据[9]。为了应对该挑战,美国、欧盟、经济合作与发展组织(OECD)和世界卫生组织(WHO)等国家或组织都大力倡导应用定量结构-活性关系(QSAR)等计算毒理学技术填补缺失的化学品数据[10-12]。那么在构建 QSAR 等预测模型时,如何表征 IOCs 离子化的影响就成为需要重点解决的问题。在前人的研究中,一般通过以下几种方式表征离子化的影响:(1)采用形态修正的正辛醇-水分配系数($\log K_{OW}$),即正辛醇-水分布系数($\log D_{OW}$)[13],但是该参数只能用于评估与分配相关的过程;(2)采用酸碱解离常数(pK_a)、分子态和离子态的比例分数(δ_M 和 δ_I)[14];(3)采用引入离子参数的多参数线性自由能关系(PP-LFER)[15],该方法仅能适用于部分有离子参数的化合物;(4)采用基于形态修正的描述符,其计算方法如下:

$$X_{修正} = \delta_M \cdot X_M + \sum_{i=1}^{n} \delta_{I\text{-}i} \cdot X_{I\text{-}i} \quad (2)$$

式中,X_M 和 $X_{I\text{-}i}$ 分别是化合物分子态和第 i 种解离态的描述符值;$\delta_{I\text{-}i}$ 是化合物第 i 种解离态的比例分数。从定义式可以看出,该方法的本质是通过考虑目标化合物在给定条件下所有存在形态的贡献而计算一个表观值。近年来,笔者所在课题组采用该方法,计算了 10 多种基于形态修正的量化描述符,并成功使用这些参数构建了 IOCs 与运甲状腺素转运蛋白[7-8,16]、血清蛋白[17]、结构蛋白[18]和磷脂膜[19]相互作用的预测模型。在本研究中,我们将进一步探索采用基于形态修正的描述符来构建 IOCs 对水生毒性效应的预测模型的可行性。基于此,本研究将构建 2 类模型:(1)仅采用 $\log K_{OW}$ 和分子形态计算的量化描述符构建预测模型;(2)采用 $\log D_{OW}$ 和基于形态修正的量化描述符构建预测模型,进而通过模型表征,比较 2 类模型预测性能的差异。

1　材料与方法(Materials and methods)

1.1　数据集

数据集包含 63 个取代苯酚、苯胺和苯甲酸类 IOCs 对大型溞的 24 h 急性毒性数据(表 1)。实验数据来源于包信等[20]的研究。原始文献测定和整理了 pH = 6.0、7.8 和 9.0 共 3 个条件的毒性数据,本研究选取 pH = 7.8 的数据作为代表进行研究。所有化合物信息及其效应值列于表 1。

表 1　模型化合物信息、大型溞急性毒性实验及预测数据

Table 1　Information of model compounds, their observed and predicted acute toxicity data of *Daphnia magna*

序号 No.	名称 Chemical name	CAS 号 CAS No.	$-\log EC_{50}$		
			实验值 Observed	模型 I Model I	模型 II Model II
1	苯酚 Phenol	000108-95-2	3.37	3.77	3.57
2	邻甲酚 2-Methylphenol	000095-48-7	3.68	3.29	3.86
3	3-甲基苯酚 3-Methylphenol	000108-39-4	3.57	3.38	4.01
4	4-甲基苯酚 4-Methylphenol*	000106-44-5	3.72	3.55	4.03
5	2,3-二甲酚 2,3-Dimethylphenol	000526-75-0	4.06	3.21	4.55
6	2-氯苯酚 2-Chlorophenol	000095-57-8	4.14	3.51	4.26
7	3-氯苯酚 3-Chlorophenol	000108-43-0	4.41	3.88	4.04
8	4-氯苯酚 4-Chlorophenol*	000106-48-9	4.37	4.83	4.05
9	2-溴苯酚 2-Bromophenol	000095-56-7	4.33	3.03	4.24
10	3-溴苯酚 3-Bromophenol	000591-20-8	4.60	4.10	4.27
11	4-溴苯酚 4-Bromophenol	000106-41-2	4.44	4.79	4.31
12	2,4-二氯苯酚 2,4-Dichlorophenol*	000120-83-2	4.75	4.81	4.99
13	2,4,6-三氯酚 2,4,6-Trichlorophenol	000088-06-2	5.05	3.8	5.22
14	五氯苯酚 Pentachlorophenol	000087-86-5	6.20	4.91	6.03
15	2-硝基苯酚 2-Nitrophenol	000088-75-5	3.65	3.13	3.74
16	3-硝基苯酚 3-Nitrophenol*	000554-84-7	3.71	3.87	4.00
17	4-硝基苯酚 4-Nitrophenol	000100-02-7	3.67	3.48	4.77
18	2,4-二硝基苯酚 2,4-Dinitrophenol	000051-28-5	4.72	4.35	4.25
19	间苯二酚 Resorcinol	000108-46-3	3.59	3.74	3.4
20	苯胺 Aniline*	000062-53-3	4.06	4.64	4.00
21	2-甲基苯胺 2-Methylaniline	000095-53-4	4.21	4.63	4.47
22	3-甲基苯胺 3-Methylaniline	000108-44-1	4.31	4.26	4.50
23	4-甲基苯胺 4-Methylaniline	000106-49-0	4.52	3.96	4.57
24	2-氯苯胺 2-Chloroaniline*	000095-51-2	4.29	5.07	4.29
25	3-氯苯胺 3-Chloroaniline	000108-42-9	4.45	4.65	4.33
26	4-氯苯胺 4-Chloroaniline	000106-47-8	4.42	4.41	4.34
27	3-溴苯胺 3-Bromoaniline	000591-19-5	4.73	4.91	4.60
28	4-溴苯胺 4-Bromoaniline*	000106-40-1	4.63	5.37	4.59
29	2,3-二氯苯胺 2,3-Dichloroaniline	000608-27-5	4.8	4.57	4.63
30	2,4-二氯苯胺 2,4-Dichloroaniline	000554-00-7	4.66	5.68	4.75
31	2,5-二氯苯胺 2,5-Dichloroaniline	000095-82-9	4.77	5.80	4.67
32	2,4,6-三溴苯胺 2,4,6-Tribromoaniline*	000147-82-0	4.92	5.65	5.61
33	2-硝基苯胺 2-Nitroaniline	000088-74-4	4.22	4.05	4.07
34	3-硝基苯胺 3-Nitroaniline	000099-09-2	4.39	4.93	4.06
35	4-硝基苯胺 4-Nitroaniline	000100-01-6	3.63	3.46	3.78
36	2,4-二硝基苯胺 2,4-Dinitroaniline*	000097-02-9	4.37	4.87	4.17
37	4-氨基苯酚 4-Aminophenol	000123-30-8	4.42	4.11	3.92
38	3-氨基苯酚 3-Aminophenol	000591-27-5	3.74	3.63	3.89
39	2-氨基苯酚 2-Aminophenol	000095-55-6	5.01	4.29	4.22
40	苯甲酸 Benzoic acid*	000065-85-0	2.24	2.19	2.40

续表1

序号 No.	名称 Chemical name	CAS 号 CAS No.	$-\log EC_{50}$		
			实验值 Observed	模型 I Model I	模型 II Model II
41	邻氟苯甲酸 2-Fluorobenzoic acid	000445-29-4	1.89	2.39	2.02
42	对氟苯甲酸 4-Fluorobenzoic acid	000456-22-4	2.15	2.57	2.54
43	2-氯苯甲酸 2-Chlorobenzoic acid	000118-91-2	2.32	2.67	2.20
44	3-氯苯甲酸 3-Chlorobenzoic acid*	000535-80-8	2.75	2.44	2.91
45	4-氯苯甲酸 4-Chlorobenzoic acid	000074-11-3	2.41	2.64	2.91
46	2-溴苯甲酸 2-Bromobenzoic acid	000088-65-3	2.49	2.58	2.50
47	3-溴苯甲酸 3-Bromobenzoic acid	000585-76-2	3.11	2.76	3.14
48	4-溴苯甲酸 4-Bromobenzoic acid*	000586-76-5	2.55	2.93	3.15
49	2,4-二氯苯甲酸 2,4-Dichlorobenzoic acid	000050-84-0	2.56	3.04	2.74
50	2,5-二氯苯甲酸 2,5-Dichlorobenzoic acid	000050-79-3	2.99	3.04	2.67
51	2,4,6-三氯苯甲酸 2,4,6-Trichlorobenzoic acid	000050-43-1	3.25	2.87	3.18
52	2,3,4,5-四氯苯甲酸 2,3,4,5-Tetrachlorobenzoic acid*	000050-74-8	3.43	3.59	3.80
53	2-氨基苯甲酸 2-Aminobenzoic acid	000118-92-3	2.88	2.95	3.08
54	3-氨基苯甲酸 3-Aminobenzoic acid	000099-05-8	2.45	3.33	2.59
55	4-氨基苯甲酸 4-Aminobenzoic acid	000150-13-0	2.40	3.11	2.62
56	4-羟基苯甲酸 4-Hydroxybenzoic acid*	000099-96-7	2.31	2.73	2.68
57	3-羟基苯甲酸 3-Hydroxybenzoic acid	000099-06-9	2.01	2.9	2.53
58	2-羟基苯甲酸 2-Hydroxybenzoic acid	000069-72-7	2.69	2.97	2.57
59	2,4-二羟基苯甲酸 2,4-Dihydroxybenzoic acid	000089-86-1	3.02	3.20	2.98
60	2,5-二羟基苯甲酸 2,5-Dihydroxybenzoic acid*	000490-79-9	3.25	4.10	2.87
61	3,4,5-三羟基苯甲酸 3,4,5-Trihydroxybenzoic acid	000149-91-7	3.88	3.31	2.46
62	邻苯二甲酸 2-Phthalic acid	000088-99-3	1.53	1.81	1.70
63	间苯二甲酸 Isophthalic acid	000121-91-5	1.44	2.23	1.47

注：*验证集化合物。

Note: *Compounds selected as the external validation set.

1.2 分子描述符计算

根据试验 pH = 7.8 和各 IOCs 的 pK_a 值，分析目标化合物在该条件下的存在形态。化合物分子态和离子态初始 3D 结构采用 ChemBioOffice 2010 软件绘制。然后采用 Gaussian 16 软件进行分子结构优化[21]。从化合物分子态 Gaussian 16 输出文件提取了 19 个量化描述符，包括偶极矩(*Dipole*-$_M$)、分子极化率(*Polar*-$_M$)、分子体积(*Volume*-$_M$)、分子最高占据轨道能($E_{HOMO\text{-}M}$)、分子最低未占据轨道能($E_{LUMO\text{-}M}$)、分子中氢原子的最正净电荷(qH_M^+)、分子中氧原子、卤素和电子供体原子的最负净电荷(qO^-_M, qX^-_M, qD^-_M)、电负性指数(ω-$_M$)、化学势(μ-$_M$)、化学硬度(η-$_M$)、分子表面上的最正和最负静电势($V_{s,max\text{-}M}$, $V_{s,min\text{-}M}$)、分子表面上正静电势和负静电势的平均值(V_s^+-$_M$, V_s^--$_M$)、分子表面上静电势平均值($V_{s\text{-}M}$)、分子表面静电势的分散度(Π-$_M$)、静电势的平衡参数(τ-$_M$)。从化合物离子态 Gaussian 16 输出文件提取了 13 个量化描述符，包括偶极矩(*Dipole*-$_I$)、分子极化率(*Polar*-$_I$)、分子体积(*Volume*-$_I$)、分子最高占据轨道能(E_{HOMO}-$_I$)、分子最低未占据轨道能(E_{LUMO}-$_I$)、分子中氧原子、卤素、电子供体原子的最负净电荷(qO^--$_I$, qX^--$_I$, qD^--$_I$)、电负性指数(ω-$_I$)、化学势(μ-$_I$)、化学硬度(η-$_I$)、分子表面上静电势平均值(V_s-$_I$)、分子表面静电势的分散度(Π-$_I$)。然后采用式 2 计算了 13 种基于形态修正的量化描述符：*Dipole*-$_{adj}$, *Polar*-$_{adj}$, *Volume*-$_{adj}$, E_{HOMO}-$_{adj}$, E_{LUMO}-$_{adj}$, qO^--$_{adj}$, qX^--$_{adj}$, qD^--$_{adj}$, ω-$_{adj}$, μ-$_{adj}$, η-$_{adj}$, V_s-$_{adj}$, Π-$_{adj}$。

其次，从 EPI Suit 4.10 软件查询了 $\log K_{OW}$ 实验值，其中 4 个无实验值的化合物采用预测的 $\log K_{OW}$ 数据。采用 MarvinSketch (ChemAxon 15.6.29.0, http://www.chemaxon.com) 软件计算 pH = 7.8 的 $\log D_{OW}$, δ_M, δ_I。其中，D_{OW} 在该软件中的定义为：

$$D_{OW} = \frac{\sum_{i=1}^{i=n} m_{i,\text{正辛醇相}}}{\sum_{i=1}^{i=n} m_{i,\text{水相}}} \quad (3)$$

式中,$m_{i,\text{正辛醇相}}$和 $m_{i,\text{水相}}$分别是化合物第 i 种存在形态在正辛醇相和水相中的浓度。

1.3　QSAR 模型构建与表征

将数据集以 3∶1 的比例划分为训练集(48 个化合物)和验证集(15 个化合物)。采用 IBM SPSS Statistics 19.0 软件中的逐步回归方法构建多元线性回归(MLR)模型。根据 OECD 关于 QSAR 模型构建与验证的导则[22],对模型进行内部和外部预测能力表征。采用决定系数(R^2)、去一法交叉验证系数(Q^2_{LOO})、Bootstrapping 验证系数(Q^2_{BOOT})、外部验证系数(Q^2_{EXT})表征模型拟合优度、稳健性和预测能力[22-24]。此外,还计算了均方根误差(RMSE)和模型的平均绝对误差(MAE),以评估预测方法的可靠性。采用方差膨胀因子(VIF)评估多重共线性,若 VIF > 10,则说明存在严重多重共线性[25]。采用 Y-随机检验评估模型是否存在偶然相关性,若 Y-随机检验的 R^2_{Y} 截距值< 0.3,Q^2_{Y} 截距值< 0.05,则模型不存在偶然相关性[26]。

采用基于杠杆值(leverage)的 Williams 图和欧几里德距离图定义模型的应用域[27]。

2　结果与讨论(Results and discussion)

2.1　大型溞急性毒性与 $\log K_{\text{OW}}$, $\log D_{\text{OW}}$ 的关系

$\log K_{\text{OW}}$ 表征了化合物分子态在正辛醇相和水相间的分配能力。不可电离化合物的水生毒性效应往往与 $\log K_{\text{OW}}$ 存在较好的线性相关性[28]。对 IOCs 该关系是否依然存在呢?从图 1 可以看出,对所研究的 63 种取代苯酚、苯胺和苯甲酸类化合物对大型溞的 24 h 急性毒性而言,$\log K_{\text{OW}}$ 与 $-\log \text{EC}_{50}$ 的 Pearson 相关系数仅为 0.265,虽然仍具有显著相关性,但相关性较差。通过引入考虑解离态贡献的 $\log D_{\text{OW}}$ 后,$\log D_{\text{OW}}$ 与 $-\log \text{EC}_{50}$ 的 Pearson 相关系数增加到 0.848。这说明在构建 IOCs 的水生毒性效应预测模型时采用 $\log D_{\text{OW}}$ 要优于 $\log K_{\text{OW}}$。

2.2　最优模型及其表征结果

仅采用化合物分子态描述符构建的最优模型(模型 I)为:

$$-\log \text{EC}_{50} = 15.7 - 9.59\ qH^{+}_{\text{-M}} - 9.12\ \tau_{\text{-M}} + 474\ V_{\text{s-M}} + 25.6\ E_{\text{HOMO-M}} \quad (4)$$

$n_{\text{Train}} = 48$, $R^2_{\text{Train}} = 0.705$, $Q^2_{\text{LOO}} = 0.622$, $Q^2_{\text{BOOT}} = 0.777$, $R^2_{\text{Y}} = 0.0826$, $Q^2_{\text{Y}} = -0.154$, $\text{RMSE}_{\text{Train}} = 0.569$, $s_{\text{Train}} = 0.601$, $\text{MAE}_{\text{Train}} = 0.461$, $P < 0.0001$

$n_{\text{EXT}} = 15$, $Q^2_{\text{EXT}} = 0.651$, $\text{RMSE}_{\text{EXT}} = 0.497$, $s_{\text{EXT}} = 0.609$, $\text{MAE}_{\text{EXT}} = 0.423$

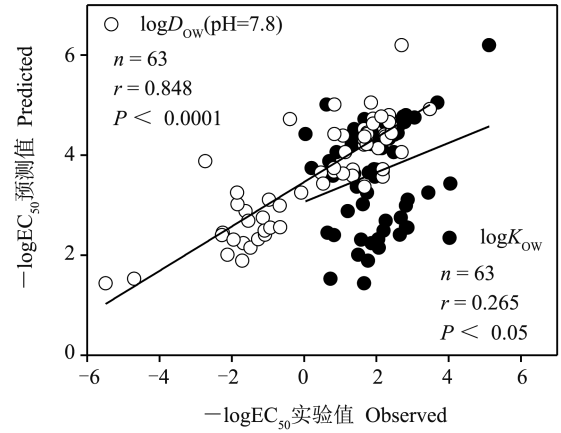

图 1　$-\log \text{EC}_{50}$ 与 $\log K_{\text{OW}}$, $\log D_{\text{OW}}$ 的关系

注:$\log K_{\text{OW}}$ 表示正辛醇-水分配系数,$\log D_{\text{OW}}$ 表示正辛醇-水分布系数。

Fig. 1　Relationship between $-\log \text{EC}_{50}$ and $\log K_{\text{OW}}$, $\log D_{\text{OW}}$

Note: $\log K_{\text{OW}}$ is n-octanol/water partition coefficient; $\log D_{\text{OW}}$ is n-octanol/water distribution coefficient.

采用基于形态修正的描述符构建的最优模型(模型 II)为:

$$-\log \text{EC}_{50} = -0.906 + 0.426\ \log D_{\text{OW}} - 4.87\ qD^{-}_{\text{-adj}} + 0.014\ Polar_{\text{-adj}} - 32.9\ \Pi_{\text{-adj}} \quad (5)$$

$n_{\text{Train}} = 48$, $R^2_{\text{Train}} = 0.875$, $Q^2_{\text{LOO}} = 0.840$, $Q^2_{\text{BOOT}} = 0.778$, $R^2_{\text{Y}} = 0.0838$, $Q^2_{\text{Y}} = -0.156$, $\text{RMSE}_{\text{Train}} = 0.370$, $s_{\text{Train}} = 0.391$, $\text{MAE}_{\text{Train}} = 0.259$, $P < 0.0001$

$n_{\text{EXT}} = 15$, $Q^2_{\text{EXT}} = 0.851$, $\text{RMSE}_{\text{EXT}} = 0.336$, $s_{\text{EXT}} = 0.411$, $\text{MAE}_{\text{EXT}} = 0.279$

2 个模型均包含 4 个预测变量,且其 VIF 值都小于 5(表 2),说明描述符之间不存在严重的多重相关性。从模型 I 和模型 II 的内部和外部验证参数可以看出,当采用基于形态修正的描述符后,模型预测

表 2　模型所选描述符的 t, P, VIF 值

Table 2　Values of t, P, VIF for selected descriptors

模型 Model	描述符 Descriptor	t	P	VIF
模型 I Model I	$qH^{+}_{\text{-M}}$	−2.75	< 0.01	1.80
	$\tau_{\text{-M}}$	−4.85	< 0.0001	1.11
	$V_{\text{s-M}}$	4.97	< 0.0001	1.79
	$E_{\text{HOMO-M}}$	4.48	< 0.0001	2.60
模型 II Model II	$\log D_{\text{OW}}$	6.97	< 0.0001	4.58
	$qD^{-}_{\text{-adj}}$	−5.23	< 0.0001	1.64
	$Polar_{\text{-adj}}$	4.91	< 0.0001	1.21
	$\Pi_{\text{-adj}}$	−3.62	< 0.001	4.17

能力显著提升,其 R^2_{Train}、Q^2_{LOO} 和 Q^2_{EXT} 等参数从 0.622 ~0.705 提高到了 0.840 ~0.875,表明基于形态修正描述符的模型具有更好的拟合优度、稳健性和预测能力。在下文中,我们仅讨论模型 II。

模型 II 的 R^2_{Train} = 0.875, Q^2_{LOO} = 0.840, Q^2_{BOOT} = 0.778, Q^2_{EXT} = 0.851,表明该模型具有较好的稳健性、拟合优度和预测能力,能够用于预测模型应用域内其他物质对大型溞的急性毒性效应。模型 II 的 R^2_Y 和 Q^2_Y 分别小于 0.3 和 0.05,说明模型不存在偶然相关性。图 2 显示了模型 I 和模型 II 中 $-\log EC_{50}$ 实验值与预测值的关系。

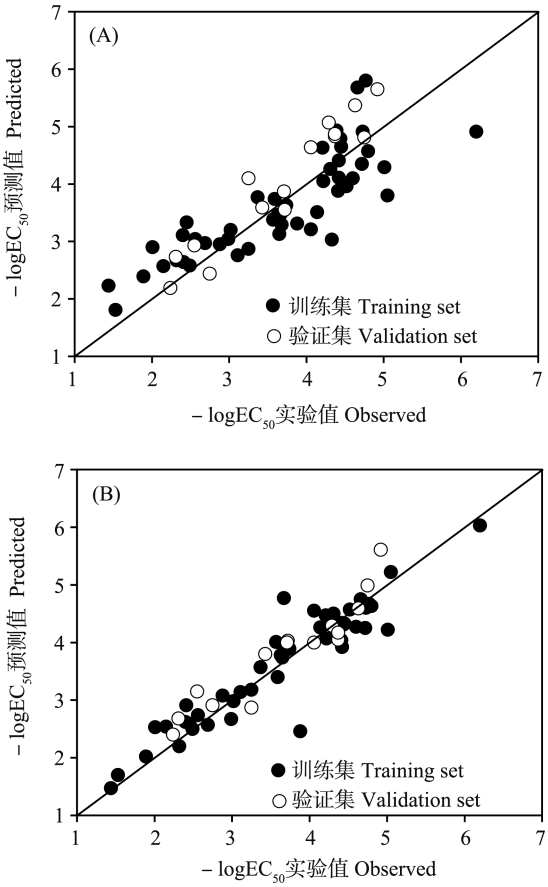

图 2 模型 I (A)和模型 II (B)中 $-\log EC_{50}$ 实验值与预测值的关系

Fig. 2 Plots of the observed versus predicted $-\log EC_{50}$ for the model I (A) and model II (B)

2.3 应用域表征

模型应用域表征结果如图 3 所示。从图 3A 可以看出,仅 1 个验证集化合物(2,4,6-三溴苯胺)处于训练集化合物定义的结构域外。在 Williams 图中,若化合物的标准残差 δ^* 落在±3.0 以外时,认为该点是离群点。从图 3B 可见,仅一个化合物(3,4,5-三羟基苯甲酸)的标准残差 δ^* 落于±3.0 以外。由于其类似物如 4-羟基苯甲酸、3-羟基苯甲酸、2-羟基苯甲酸、2,4-二羟基苯甲酸、2,5-二羟基苯甲酸的标准残差 δ^* 均落于±3.0 以内,说明模型能够正确预测该类化合物的毒性效应。导致 3,4,5-三羟基苯甲酸离群的原因可能是实验高估了其对大型溞的急性毒性。在图 3B 中有 2 个物质(间苯二甲酸和 2,4,6-三溴苯胺)的杠杆值均大于警戒值 h^*。但是模型较好地预测了间苯二甲酸和 2,4,6-三溴苯胺对大型溞的急性毒性,说明模型具有较好的延展性[29]。

2.4 机理解释

模型 II 包含 4 个描述符,即 $\log D_{OW}$、qD^-_{adj}、$Polar_{adj}$ 和 Π_{adj}。$\log D_{OW}$ 是考虑离子化影响的正辛醇-水分配系数,其能够较好表征具有麻醉作用模式的水生生物急性毒性[28,30]。在本研究中,$\log D_{OW}$ 与 $-\log EC_{50}$ 的线性相关系数为 0.848,一方面说明数据集中存在具有麻醉作用模式物质;同时 $\log D_{OW}$ 不能完全解释化合物对 $-\log EC_{50}$ 的贡献,也意味着存在其他作用模式的物质,需要引入其他参数来表征其

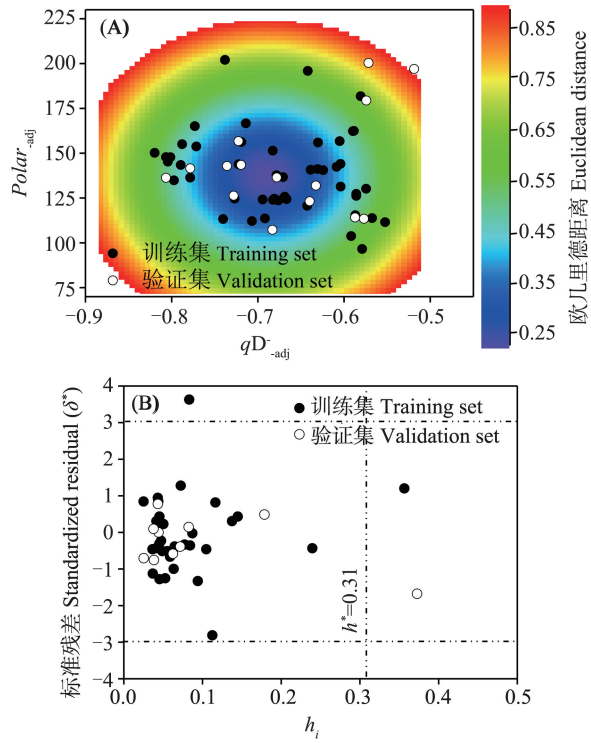

图 3 基于欧几里德距离方法(A)和 Williams 图(B)表征的模型 II 应用域模

Fig. 3 Characterization of application domain for model II based on the Euclidean distance (A) and Williams plot (B)

作用的贡献。qD^{-}_{adj} 是基于形态修正的分子中电子供体原子(氧、氮和卤素)的最负净电荷[31],它表征了形成氢键的能力,其值越负,形成氢键的能力越强。在模型 II 中,其具有负的系数,说明其数值越负,$-\log EC_{50}$ 值越大,也即毒性越强,说明分子结构中引入氢键基团能提高其毒性效应;$Polar_{adj}$ 是基于形态修正的分子极化率,$Polar_{adj}$ 与分子疏水性有关[32]。$Polar_{adj}$ 具有正的系数,说明其数值越大,$-\log EC_{50}$ 值越大,毒性越强。Π_{adj} 是基于形态修正的分子表面静电势的分散度,表征了分子中正负静电势的平衡性[33]。其值越大说明分子表面正负静电势分布越不均衡,反之则说明分子表面正负静电势分布均衡。在研究的数据集中,我们发现在实验条件下解离程度越大的物质,其 Π_{adj} 值越大,而解离程度越小的物质则 Π_{adj} 越小。取代苯甲酸类物质的解离程度最大,几乎全以阴离子态存在,而阴离子的分子表面以负电势为主,正电势分布较少,导致其正负静电势分布极不均衡。Π_{adj} 具有负的系数,说明其数值越小,$-\log EC_{50}$ 值越大,毒性越强。这说明解离程度越大的物质,会导致 Π_{adj} 越大,进而导致 $-\log EC_{50}$ 值越小,毒性也越小。

2.5　模型比较

包信等[20]分别构建了针对 19 种苯酚类、17 种苯胺类和 24 种苯甲酸类物质大型溞急性毒性的局域预测模型,从表 3 可以看出,针对 19 种苯酚类、17 种苯胺类物质的模型具有较好的预测能力,但是对 24 种苯甲酸类物质的模型预测能力较差,仅在删除部分苯甲酸类物质的情况下,才能得到预测能力较好的模型。他们构建的局域模型可以用于分别预测苯酚类、苯胺类和苯甲酸类物质对大型溞的急性毒性数据。本研究针对 19 种苯酚类、20 种苯胺类和 24 种苯甲酸类物质,构建了能同时预测上述 3 类物质对大型溞急性毒性的模型,所建模型具有较好的内部和外部预测能力,并进行了应用域表征。

综上,本研究探索了采用基于形态修正的描述符构建 IOCs 水生毒性指标预测模型的可行性。研究结果表明,使用基于形态修正的描述符构建的 IOCs 大型溞急性毒性模型预测能力要优于仅采用分子形态描述符的模型。因此,在将来构建 IOCs 水生毒性效应预测模型时,可考虑引入基于形态修

表 3　本研究与文献模型比较

Table 3　Comparison of the current model with previous QSAR models

化合物类别 Chemical classes	模型及表征结果 Models and model statistics	应用域 Applicability domain	文献来源 References
苯酚类 Phenols	$\log 1/IC_{50} = 0.667\log K_{OW} + 2.65$ $m = 1,\ n_{Train} = 19,\ R^2_{Train} = 0.790,\ s_{Train} = 0.322$ $\log 1/IC_{50} = 0.561\log K_{OW} - 0.219\log F_0 + 2.79$ $m = 2,\ n_{Train} = 19,\ R^2_{Train} = 0.880,\ s_{Train} = 0.251$	否 No	[20]
苯胺类 Anilines	$\log 1/IC_{50} = 0.298\log K_{OW} + 3.90$ $m = 1,\ n_{Train} = 17,\ R^2_{Train} = 0.732,\ s_{Train} = 0.134$ $\log 1/IC_{50} = 0.333\log K_{OW} + 0.028\mathrm{p}K_a + 3.76$ $m = 2,\ n_{Train} = 17,\ R^2_{Train} = 0.786,\ s_{Train} = 0.124$		
苯甲酸类 Benzoic acids	$\log 1/IC_{50} = 0.206\log K_{OW} + 2.17$ $m = 1,\ n_{Train} = 24,\ R^2_{Train} = 0.091,\ s_{Train} = 0.568$ $\log 1/IC_{50} = 0.629\log K_{OW} + 1.01$ $m = 1,\ n_{Train} = 13,\ R^2_{Train} = 0.709,\ s_{Train} = 0.257$ $\log 1/IC_{50} = 0.429\log K_{OW} + 0.738 E_{HOMO} + 9.01$ $m = 2,\ n_{Train} = 24,\ R^2_{Train} = 0.373,\ s_{Train} = 0.483$		
苯酚类、苯胺类、苯甲酸类 Phenols, Anilines, Benzoic acids	$-\log EC_{50} = 15.7 - 9.59\ qH^{+}_{-M} - 9.12\ \tau_{-M} + 474\ V_{s\text{-}M} + 25.6\ E_{HOMO\text{-}M}$ $m = 4,\ n_{Train} = 48,\ R^2_{Train} = 0.705,\ s_{Train} = 0.601$ $n_{EXE} = 15,\ R^2_{EXE} = 0.651,\ s_{EXE} = 0.609$ $-\log EC_{50} = -0.906 + 0.426\ \log D_{OW} - 4.87\ qD^{-}_{adj} + 0.014\ Polar_{adj} - 32.9\ \Pi_{adj}$ $m = 4,\ n_{Train} = 48,\ R^2_{Train} = 0.875,\ s_{Train} = 0.391$ $n_{EXE} = 15,\ R^2_{EXE} = 0.851,\ s_{EXE} = 0.411$	是 Yes	本研究 This study

正的描述符。

通讯作者简介:杨先海(1985-),男,环境工程工学博士,讲师,主要研究方向为新兴有机污染物的生态与健康风险评估、面向化学品风险评估的计算毒理学方法与模型软件开发、内分泌干扰效应的分子机制探究与测试方法开发等,发表学术论文 50 余篇。

参考文献(References):

[1] Franco A, Ferranti A, Davidsen C, et al. An unexpected challenge: Ionizable compounds in the REACH chemical space [J]. The International Journal of Life Cycle Assessment, 2010, 15(4): 321-325

[2] Bittermann K, Spycher S, Goss K U. Comparison of different models predicting the phospholipid-membrane water partition coefficients of charged compounds [J]. Chemosphere, 2016, 144: 382-391

[3] Manallack D T. The p*K*(a) distribution of drugs: Application to drug discovery [J]. Perspectives in Medicinal Chemistry, 2007, 1: 25-38

[4] Boxall A B, Rudd M A, Brooks B W, et al. Pharmaceuticals and personal care products in the environment: What are the big questions? [J]. Environmental Health Perspectives, 2012, 120(9): 1221-1229

[5] Schwarzenbach R P, Escher B I, Fenner K, et al. The challenge of micropollutants in aquatic systems [J]. Science, 2006, 313(5790): 1072-1077

[6] Rendal C, Kusk K O, Trapp S. Optimal choice of pH for toxicity and bioaccumulation studies of ionizing organic chemicals [J]. Environmental Toxicology and Chemistry, 2011, 30(11): 2395-2406

[7] Yang X H, Xie H B, Chen J W, et al. Anionic phenolic compounds bind stronger with transthyretin than their neutral forms: Nonnegligible mechanisms in virtual screening of endocrine disrupting chemicals [J]. Chemical Research in Toxicology, 2013, 26(9): 1340-1347

[8] Yang X H, Lyakurwa F, Xie H B, et al. Different binding mechanisms of neutral and anionic poly-/perfluorinated chemicals to human transthyretin revealed by *in silico* models [J]. Chemosphere, 2017, 182: 574-583

[9] Card M L, Gomez-Alvarez V, Lee W H, et al. History of EPI Suite™ and future perspectives on chemical property estimation in US Toxic Substances Control Act new chemical risk assessments [J]. Environmental Science: Processes & Impacts, 2017, 19(3): 203-212

[10] Kavlock R, Dix D. Computational toxicology as implemented by the U.S. EPA: Providing high throughput decision support tools for screening and assessing chemical exposure, hazard and risk [J]. Journal of Toxicology and Environmental Health, Part B, 2010, 13(2-4): 197-217

[11] 王中钰, 陈景文, 乔显亮, 等. 面向化学品风险评价的计算(预测)毒理学[J]. 中国科学: 化学, 2016, 46(2): 222-240
Wang Z Y, Chen J W, Qiao X L, et al. Computational toxicology: Oriented for chemicals risk assessment [J]. Scientia Sinica Chimica, 2016, 46(2): 222-240 (in Chinese)

[12] Tang W, Chen J W, Wang Z Y, et al. Deep learning for predicting toxicity of chemicals: A mini review [J]. Journal of Environmental Science and Health. Part C, 2018, 36(4): 252-271

[13] Kah M, Brown C D. LogD: Lipophilicity for ionisable compounds [J]. Chemosphere, 2008, 72(10): 1401-1408

[14] Zhao Y H, Yuan X, Su L M, et al. Classification of toxicity of phenols to tetrahymena pyriformis and subsequent derivation of QSARs from hydrophobic, ionization and electronic parameters [J]. Chemosphere, 2009, 75(7): 866-871

[15] Henneberger L, Goss K U, Endo S. Partitioning of organic ions to muscle protein: Experimental data, modeling, and implications for *in vivo* distribution of organic ions [J]. Environmental Science & Technology, 2016, 50(13): 7029-7036

[16] Yang X H, Ou W, Xi Y, et al. Emerging polar phenolic disinfection byproducts are high-affinity human transthyretin disruptors: An *in vitro* and *in silico* study [J]. Environmental Science & Technology, 2019, 53(12): 7019-7028

[17] Ding F, Yang X H, Chen G S, et al. Development of bovine serum albumin-water partition coefficients predictive models for ionogenic organic chemicals based on chemical form adjusted descriptors [J]. Ecotoxicology and Environmental Safety, 2017, 144: 131-137

[18] Ou W, Liu H H, He J Y, et al. Development of chicken and fish muscle protein—Water partition coefficients predictive models for ionogenic and neutral organic chemicals [J]. Ecotoxicology and Environmental Safety, 2018, 157: 128-133

[19] Lin S Y, Yang X H, Liu H H. Development of liposome/water partition coefficients predictive models for neutral and ionogenic organic chemicals [J]. Ecotoxicology and Environmental Safety, 2019, 179: 40-49

[20] 包信, 张栩嘉, 赵元慧. 可离子化有机污染物对大型溞的毒性及 QSAR 研究[J]. 生态毒理学报, 2016, 11(2): 720-731
Bao X, Zhang X J, Zhao Y H. The toxicities of ionized

organic compounds to *Daphnia magna* and QSAR study [J]. Asian Journal of Ecotoxicology, 2016, 11(2): 720-731 (in Chinese)

[21] Frisch M J, Trucks G W, Schlegel H B, et al. Gaussian 16, Revision A.03 [CP]. Wallingford, CT: Gaussian, Inc., 2016

[22] OECD. Guidance Document on the Validation of (Quantitative) Structure Activity Relationships [(Q)SAR]. Models Environment Health and Safety Publications Series on Testing and Assessment No. 69 [R]. Paris: OECD, 2007

[23] Schüürmann G, Ebert R U, Chen J W, et al. External validation and prediction employing the predictive squared correlation coefficient test set activity mean vs training set activity mean [J]. Journal of Chemical Information and Modeling, 2008, 48(11): 2140-2145

[24] 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]. Journal of Chemical Information and Modeling, 2012, 52(8): 2044-2058

[25] Liu H H, Yang X H, Lu R. Development of classification model and QSAR model for predicting binding affinity of endocrine disrupting chemicals to human sex hormone-binding globulin [J]. Chemosphere, 2016, 156: 1-7

[26] Eriksson L, Jaworska J, Worth A P, et al. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs [J]. Environmental Health Perspectives, 2003, 111(10): 1361-1375

[27] 杨先海, 刘会会, 杨倩, 等. 双酚 A 类似物雌激素干扰效应的定量结构-活性关系模型[J]. 生态毒理学报, 2016, 11(4): 69-78
Yang X H, Liu H H, Yang Q, et al. Predicting estrogenic activity of bisphenols using quantitative structure-activity relationships [J]. Asian Journal of Ecotoxicology, 2016, 11(4): 69-78 (in Chinese)

[28] Cronin M T. (Q)SARs to predict environmental toxicities: Current status and future needs [J]. Environmental Science: Processes & Impacts, 2017, 19(3): 213-220

[29] Wang Y, Chen J W, Yang X H, et al. *In silico* model for predicting soil organic carbon normalized sorption coefficient (K(OC)) of organic chemicals [J]. Chemosphere, 2015, 119: 438-444

[30] 刘羽晨, 乔显亮. 水生生物急性毒性 QSAR 模型研究进展[J]. 生态毒理学报, 2015, 10(2): 26-35
Liu Y C, Qiao X L. Progress in quantitative structure-activity relationship models for acute aquatic toxicity [J]. Asian Journal of Ecotoxicology, 2015, 10(2): 26-35 (in Chinese)

[31] Liu H H, Wei M B, Yang X H, et al. Development of TLSER model and QSAR model for predicting partition coefficients of hydrophobic organic chemicals between low density polyethylene film and water [J]. Science of the Total Environment, 2017, 574: 1371-1378

[32] Karelson M, Lobanov V S, Katritzky A R. Quantum-chemical descriptors in QSAR/QSPR studies [J]. Chemical Reviews, 1996, 96(3): 1027-1044

[33] Murray J S, Brinck T, Lane P, et al. Statistically-based interaction indices derived from molecular surface electrostatic potentials: A general interaction properties function (GIPF) [J]. Journal of Molecular Structure: THEOCHEM, 1994, 307: 55-64

◆