

有机磷酸酯类化合物的定量结构-活性关系及稳健性分析*

赵劲松 王 宇 王海燕 王晓栋 王连生**

(南京大学环境学院, 污染控制与资源化国家重点实验室, 南京, 210093)

摘 要 应用理论线性溶剂化能关系 (TLSER) 模型研究了 35 种有机磷酸酯类 (OP) 化合物的结构及其对家蝇急性毒性之间的关系, 得到了可以指示毒性作用机理的定量结构-活性关系. 在 OP 化合物与生物体的作用过程中, 分子的极性是影响活性的最重要因素, 表明 OP 化合物可能与生物体发生特异性反应. 采用 Bootstrap 方法分析模型的稳健性, 结果表明, 模型的系数、标准误差及相关系数均与 Bootstrap 估计接近, 模型具有较强的稳健性.

关键词 有机磷化合物, 定量结构-活性关系, 理论线性溶剂化能关系, Bootstrap 方法.

有机磷酸酯类化合物 (OP) 的广泛应用所带来的环境问题已不容忽视. 定量结构-活性关系 (QSAR) 在有机化合物的生态风险评价中有着广泛的应用, 然而有关 OP 化合物的 QSAR 研究非常少, 涉及到毒性研究的则更少^[1]. 交叉验证是 QSAR 研究中应用最为广泛的模型验证方法, 采用交叉验证相关系数 Q^2 表征模型的预测能力^[2-3], 然而研究表明, 交叉验证可能会高估模型的预测能力^[3]. 另外, 该方法不能给出 QSAR 模型各个参数分布的估计.

Bootstrap 是一种基于对原始样本采用有放回的重抽样的统计模拟方法. 在不需对未知总体的分布作任何假设的条件下, 通过大量的重抽样, 进行参数的偏置、方差及置信区间等统计量的估计^[4-6], 从而可用于模型稳健性的分析. 随着计算机技术的发展, 在化学计量领域已有相当广泛的应用^[6].

本研究的目的是基于理论线性溶剂化能关系 (TLSER), 建立 35 种 OP 对家蝇 (*Musca nobulo* L.) 急性毒性的 QSAR 模型, 并采用 Bootstrap 方法对模型的稳健性进行分析.

1 数据来源及 TLSER 描述符

35 种 OP 化合物对家蝇的急性毒性数据取自文献^[7] (见表 1), 所有化合物均为乙酰胆碱酯酶的直接抑制剂甲基和乙基对氧磷 (化合物 12 和 22) 的类似物^[8-9]. 化合物在结构与极性上有较大的差异, 苯环上的取代基 (X) 的电负性变化较大; 而烷基 (R) 的体积变化较大.

TLSER 是在线性溶剂化能关系 (LSER) 的基础上通过理论计算代替 LSER 经验性描述符而建立起来的一种 QSAR 研究方法. 根据文献^[10-11], 采用半经验量子化学软件包 MOPAC 6.0 中的 PM3 方法^[12]对 35 种 OP 化合物的结构进行优化之后, 计算了全部 12 种 TLSER 描述符 (见表 1).

2 统计方法

TLSER 模型的建立通常是采用多元分析中的逐步回归进行^[10-11]. 然而, 由于逐步回归对进入模型的变量是否存在自相关非常敏感, 当 TLSER 描述符之间存在共线性时, 采用逐步回归便难以得到统计意义显著的 TLSER 模型^[13]. 因此, 在本研究中采用最佳子集搜索, 即对所有描述符的组合进行遍历搜索, 并设定判断描述符共线性的变量膨胀因子 (VIF), 以得到最佳 TLSER 模型. 通常, VIF 值愈小表明变量间的共线性愈弱, 对于多元线性回归而言, $VIF < 5$ 是一个可接受的值. 在本研究中, 为减少描述符共线性对模型质量的影响, 设定 $VIF < 2$ ^[13]. 模型的拟合结果用预测能力由相关系数 R^2 , 标准偏差 SE; 预测能力由交叉验证的相关系数 Q^2 , 预测标准误差 SE_p 表征.

2005 年 3 月 1 日收稿.

* 国家重点基础研究发展计划项目 (973 计划) (2003CB415002)、国家自然科学基金项目 (20477018)、欧盟国际合作项目 (ICA4-CT-2001-10039)、国家高技术研究发展计划 (863 计划) (2001AA640601-4) 基金资助. ** 通讯联系人.

表 1 35 种 OP 化合物对家蝇的急性毒性及 TLSER 描述符
Table 1 The acute toxicity to housefly and TLSER descriptors of 35 OP compounds

R	No.	X	$\lg(1/LC_{50})$	V	π_1	ϵ_s	q^-	ϵ_s	qH^+	$q\sigma^-$	$q\sigma^+$	$q\sigma^-$	μ_{tot}	μ_{max}	μ_{ave}
CH ₃	1	H	2.75	157.78	0.0899	0.16669	0.8759	0.17947	0.139	29.2223	28.4380	9.5761	25.3087	4.7253	1.0545
	2	3-CH ₃	2.00	177.06	0.0870	0.16753	0.8759	0.17922	0.1401	25.8595	25.3101	9.6226	25.6391	4.7254	0.9496
	3	4-CH ₃	1.99	175.51	0.0887	0.16864	0.8762	0.17930	0.1384	25.7553	25.3005	9.6808	25.2641	4.7238	0.9357
	4	4-OCH ₃	2.00	183.76	0.0887	0.16750	0.8752	0.17998	0.1397	24.9823	22.7676	8.9715	25.9855	4.7228	0.9281
	5	3-Cl	2.10	172.82	0.0915	0.16523	0.8454	0.18027	0.1424	29.1827	28.6390	9.5244	25.4616	4.6711	1.0609
	6	4-Cl	2.60	175.60	0.0914	0.16705	0.8746	0.18090	0.1434	29.2392	28.6498	9.7706	25.3517	4.7271	1.0563
	7	3-Br	4.00	180.74	0.0897	0.16448	0.8453	0.18098	0.1423	29.1375	30.4762	9.8077	25.2835	4.6713	1.0535
	8	4-Br	3.53	179.31	0.0918	0.16555	0.8744	0.18112	0.1442	29.2476	30.5043	10.0559	25.3873	4.7272	1.0578
	9	3-CN	4.99	174.67	0.0945	0.16254	0.8452	0.18228	0.1431	27.9573	28.5668	9.4532	25.6225	4.6725	1.0249
	10	4-CN	4.84	172.43	0.0983	0.16387	0.8725	0.18310	0.1472	28.1442	30.6156	9.9127	25.3683	4.7295	1.0147
	11	3-NO ₂	4.90	174.31	0.0930	0.15903	0.8299	0.18850	0.1550	36.8571	35.8119	8.6642	32.4149	4.6762	1.2467
	12	4-NO ₂	5.10	173.76	0.0956	0.15778	0.8688	0.18732	0.1524	38.1068	32.4798	4.7737	34.4467	4.7311	1.3249
C ₂ H ₅	13	H	3.20	193.12	0.0842	0.16700	0.8487	0.17851	0.1389	23.3204	22.4214	8.4431	27.1830	4.6652	0.9061
	14	4-CH ₃	3.00	211.22	0.0844	0.16897	0.8804	0.17913	0.1379	21.0580	20.3574	8.5672	26.9849	4.726	0.8177
	15	3-Cl	3.80	208.61	0.0865	0.16563	0.8478	0.17993	0.1433	23.3693	22.5815	8.5556	27.2851	4.6675	0.9095
	16	4-Cl	3.72	211.16	0.0859	0.16730	0.8477	0.17993	0.1434	23.3329	22.5707	8.5776	27.2270	4.6672	0.9076
	17	3-Br	4.11	216.78	0.0851	0.16491	0.8480	0.18065	0.1432	23.3337	23.7017	8.6473	27.1110	4.6676	0.9037
	18	4-Br	4.06	214.65	0.0864	0.16583	0.8475	0.18018	0.1441	23.3400	23.7101	8.7010	27.2632	4.6674	0.9088
	19	3-CN	5.00	206.29	0.0909	0.16299	0.8477	0.18192	0.1440	22.5835	22.5336	8.3905	27.4499	4.6701	0.8855
	20	4-CN	5.10	208.51	0.0914	0.16417	0.8464	0.18225	0.1477	22.6599	23.8059	8.5708	27.2390	4.6713	0.8787
	21	3-NO ₂	5.10	208.44	0.0887	0.15963	0.8309	0.18817	0.1573	29.9536	28.7278	8.3255	34.3080	4.6729	1.0721
	22	4-NO ₂	5.20	209.21	0.0907	0.15826	0.8692	0.18698	0.1530	30.9226	26.7101	6.6199	36.1900	4.7314	1.1309
	23	2,4-Cl	4.30	223.46	0.0879	0.16722	0.8448	0.18095	0.1446	23.3335	22.6851	8.5043	27.4328	4.6672	0.9144
	24	2,5-Cl	4.10	226.04	0.0864	0.16679	0.8426	0.18078	0.1554	23.3668	22.6912	8.4353	27.4891	4.6628	0.9163
C ₄ H ₉	25	H	2.50	259.73	0.0798	0.16698	0.8759	0.17960	0.1394	16.6625	16.0409	7.3210	28.3085	4.7277	0.6740
	26	3-CH ₃	2.00	278.05	0.0789	0.16780	0.8761	0.17937	0.1405	15.5324	14.9914	7.1642	28.6357	4.7284	0.6363
	27	4-CH ₃	2.10	274.90	0.0804	0.16891	0.8761	0.17944	0.1389	15.4721	14.9750	7.1658	28.2630	4.7259	0.6281
	28	4-OCH ₃	2.10	289.58	0.0793	0.16793	0.8575	0.17950	0.1237	15.1701	14.0373	7.0515	28.9961	4.6898	0.6303
	29	3-Cl	2.80	280.25	0.0797	0.16556	0.8453	0.18028	0.1432	16.6456	16.1330	7.2506	28.4681	4.6731	0.6778
	30	4-Cl	2.50	274.77	0.0824	0.16731	0.8748	0.18098	0.1436	16.6690	16.1294	7.3406	28.3583	4.7293	0.6752
	31	4-Br	2.95	278.03	0.0829	0.16583	0.8747	0.18119	0.1445	16.6758	16.7098	7.3299	28.3948	4.7297	0.6761
	32	3-CN	4.00	273.43	0.0851	0.16302	0.8735	0.18252	0.1446	16.2972	16.1269	7.1376	28.6070	4.7306	0.6653
	33	4-CN	4.01	274.49	0.0859	0.16422	0.8724	0.18304	0.1478	16.3453	16.7827	7.1693	28.3847	4.7315	0.6601
	34	3-NO ₂	4.21	283.26	0.0805	0.15948	0.8290	0.18826	0.1568	21.6944	20.9116	7.7969	35.4914	4.6786	0.8066
	35	4-NO ₂	4.38	281.87	0.0824	0.15817	0.8692	0.18713	0.1528	22.4157	19.8805	7.5371	37.4883	4.7332	0.8520

3 Bootstrap

Bootstrap是建立在样本可以表征该样本所在的总体的假设之上. 对于多元线性回归模型中参数的估计有残差重抽样和观测值重抽样^[4-6]. 由于观测值重抽样不要求模型的残差必须方差齐性^[4-6], 因此, 本文采用观测值重抽样来估计 TLSER模型各个参数的稳健性. 假定观测值的数目为 n , 自变量为 $x = (x_1, x_2, \dots, x_n)$, 应变量为 $y = (y_1, y_2, \dots, y_n)$, 则 Bootstrap统计模拟为: (1) 从 $(1, 2, \dots, n)$ 中采用可放回方式随机抽取 n 个数, 记作 $i = (i_1, i_2, \dots, i_n)$; (2) 从 x 和 y 中以 i 中的记录的数值作为下标抽取观测值, 组成 Bootstrap样本 x^* 和 y^* ; (3) 对 x^* 和 y^* 进行线性回归, 求得模型的回归系数与标准偏差的 Bootstrap估计; (4) 重复步骤 (1) 至步骤 (3) B 次; (5) 根据得到的模型参数的 Bootstrap估计分布, 求取模型参数的偏置, 方差及置信区间. 计算过程参考文献 [4]. 根据 Davison和 Hinkley的建议^[4], 这里取 $B = 1999$. 所有的统计分析均采用统计计算的语言与环境 R^[14]完成.

4 TLSER 模型

根据 Famin等的 TLSER理论^[10-11], 对 35种 OP化合物的 $\lg(1/LC_{50})$ 与 12种 TLSER描述符之间进行回归分析, 得到如下模型:

$$\lg(1/LC_{50}) = 94.078 + 72.808\pi_1 - 211.915\varepsilon_b - 13.157\mu_{max} \quad (1)$$

$$n = 34, R^2 = 0.830, F = 78.720, SE = 0.475, Q^2 = 0.785, SE_P = 0.501, p < 0.001$$

Gandhe等的研究表明^[7], 取代基 R的不同, 会导致其生物活性的差异. 将 35种 OP化合物分成三类, 即双甲基取代, 双乙基取代及双丁基取代. 它们的 TLSER模型分别如下:

双甲基取代 OP化合物 (化合物 1—12):

$$\lg(1/LC_{50}) = 8.893 + 284.261\pi_1 - 36.350q^- \quad (2)$$

$$n = 11, R^2 = 0.929, F = 52.440, SE = 0.387, Q^2 = 0.856, SE_P = 0.471, p < 0.001$$

双乙基取代 OP化合物 (化合物 13—24):

$$\lg(1/LC_{50}) = -17.938 + 221.125\pi_1 - 0.121q\sigma^+ \quad (3)$$

$$n = 12, R^2 = 0.951, F = 86.810, SE = 0.183, Q^2 = 0.917, SE_P = 0.206, p < 0.001$$

双丁基取代 OP化合物 (化合物 25—35):

$$\lg(1/LC_{50}) = 25.860 + 142.660\pi_1 - 208.76\varepsilon_b \quad (4)$$

$$n = 11, R^2 = 0.977, F = 172.900, SE = 0.155, Q^2 = 0.942, SE_P = 0.213, p < 0.001$$

其中, TLSER模型 (1) 与 (2) 均不包括化合物 5. 因为在回归模型的诊断中, 化合物 5与其它化合物之间存在较大差异. 由于该化合物与同组其它化合物的结构无显著差异, 所以差异可能是由于生物活性测定上的误差; 该化合物与生物体的作用机理与其它化合物不同^[15].

TLSER模型 (1) — (4)在模型拟合及预测能力上都具有显著的统计意义. π 是唯一进入所有四个模型的参数. 根据 Famin等的研究^[10-11], π 表征是单位范德华体积上的极化率, 在非特异性反应中, 随着 π 值的增大, 化合物倾向于溶解于水, 在细胞膜上的分配可能会少一些. 然而, 本文研究的 OP类化合物与生物体乙酰胆碱酯酶 (AChE) 之间存在特异反应^[8-9, 15, 16], 模型中生物毒性随 π 的增大而增大, 表明 OP对家蝇的急性毒性中, OP与 AChE之间的相互作用可能是致毒的最重要机理.

模型 (1) 与 (4) 中的 ε_b 及模型 (2) 中的 q^- 表示化合物共价键与静电性对氢键受体的碱性的贡献, 其值愈大, OP的生物活性就愈低. 这与 OP化合物的特异性反应机理一致^[14, 15]. 也与此前对 OP化合物氢键受体的比较分子相似性指数分析 (CoMSIA) 的结果一致^[15]. 相似的, μ_{max} 与 $q\sigma^+$ 均与 OP化合物的负电荷区域相关, 因而, 当其增大时, 不利于 OP的生物活性.

5 TLSER 模型稳健性分析

通过对小样本进行大量可放回重抽样得到的 Bootstrap估计的分布, 可以较好地反映样本总体的内在分布^[4-5], 因此, 可以用 Bootstrap估计来判断模型稳健性. 表 2为 TLSER模型中 π 系数的偏置、标准误差及置信区间与 Bootstrap估计的比较. 从表 2可以看出, Bootstrap估计与模型参数非常接近.

对其它参数的模拟也取得了相似的结果，表明 OP对家蝇急性毒性的 TLSEr模型是稳健的。

表 2 TLSEr模型 π_1 系数的偏置、标准误差及置信区间的理论值与 Bootstrap估计

Table 2 Comparison of Bootstrap bias, standard error and confidence interval of the π_1 coefficient of TLSEr model with theoretical ones					
统计量		TLSEr (1)	TLSEr (2)	TLSEr (3)	TLSEr (4)
理论值	回归系数	72.808	284.261	221.125	142.655
	偏置	0	0	0	0
	标准误差	17.927	37.079	25.042	22.330
	置信区间	(36.197, 109.419)	(198.756, 369.765)	(164.475, 277.776)	(91.168, 194.143)
Bootstrap估计	偏置	-1.839	5.159	-4.062	-8.701
	标准误差	17.778	49.982	32.009	30.054
	置信区间	(44.260, 113.130)	(159.500, 376.000)	(175.400, 302.300)	(105.700, 226.300)

为了更清晰地说明 TLSEr 模型的稳健性，以 TLSEr 模型 (1)为例，对拟合标准误差 SE及相 关系数 R^2 进行了 Bootstrap 模拟。图 1给出了 TLSEr 模型 (1) R^2 的 Bootstrap 模拟结果。其中，柱状图给出了 Bootstrap R^2 估计的分布及概率密度。从图 1中可以看出模型相关系数 R^2 (实线) 与 Bootstrap R^2 (虚线) 之间相差很小。从图 1可以看出，Bootstrap R^2 近似呈正态分布。由于 B 足够大，Bootstrap R^2 的分布可以作为模型真实情况的估计。因此，TLSEr 模型 (1) 的线性回归具有良好的稳健性。

图 2是 TLSEr 模型 (1) 的 SE的 Bootstrap 模型结果。与图 1给出的结果相似，Bootstrap SE的估计值与模型的 SE 相近。从正态拟合图中可以看出，Bootstrap 的 SE 呈近似正态分布。这进一步说明了 TLSEr 模型在 OP 化合物对家蝇急性毒性的 QSAR 研究中给出了足够稳定及预测性能良好的模型。

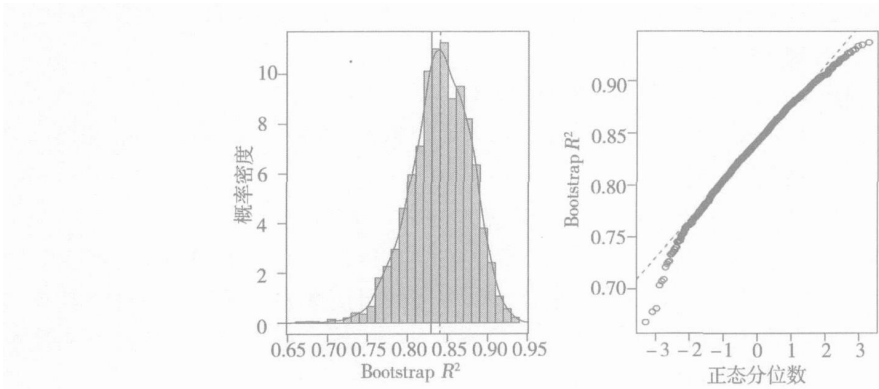


图 1 TLSEr 模型 (1) R^2 的 Bootstrap 模拟
Fig 1 Bootstrap R^2 of TLSEr model (1)

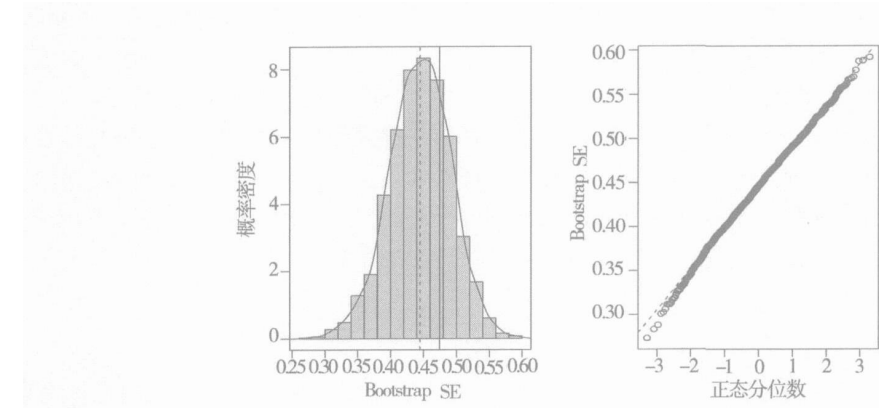


图 2 TLSEr 模型 (1) SE 的 Bootstrap 模拟
Fig 2 Bootstrap SE of TLSEr model (1)

综上所述, 在 OP 化合物对家蝇急性毒性的 QSAR 研究中, TLSEr 模型给出了机理明确的模型, π 在一定程度揭示了 OP 化合物对家蝇的急性毒性可能是由于 OP 化合物与 AChE 之间的特异反应造成的. 另外, 进入模型的其它参数说明了 OP 化合物表面负电荷的积聚在一定程度上阻碍了 OP 化合物与生物体之间的相互作用. 对 TLSEr 模型的系数、标准误差、相关系数的 Bootstrap 模拟结果表明, 模型的相应系数与 Bootstrap 估计之间相差很小, 模型具有良好的稳健性.

参 考 文 献

- [1] Knaak J B, Dary C G, Power F et al., Physicochemical and Biological data for the Development of Predictive Organophosphorus Pesticide QSARs and PBPK/PD Models for Human Risk Assessment *Crit. Rev. Toxicol.*, 2004, **34** (2): 143—207
- [2] Tropsha A, Gramatica P, Gombar V K, The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSAR Models *QSAR Comb. Sci.*, 2003, **22** (1): 69—77
- [3] Golbraikh A, Tropsha A, Beware of q^2 ! *J. Mol. Graph. Model.*, 2002, **20** (4): 269—276
- [4] Davison A C, Hinkley D V, Bootstrap Methods and Their Application. New York: Cambridge University Press, 1997
- [5] Efron B, Tibshirani R J, An Introduction to the Bootstrap. Boca Raton: Chapman & Hall/CRC, 1994
- [6] Wehrens R, Putter H, Buydens L M C, The Bootstrap: A Tutorial *Chemometr. Intell. Lab.*, 2000, **54** (1): 35—52
- [7] Gandhe B R, Pumanand, Prasad R et al., Use of Gas Chromatographic Retention Indices for Quantitative Structure Activity Relationship Studies of Dialkylphenyl Phosphates *Pestic. Sci.*, 1990, **29** (4): 379—385
- [8] Mileson B E, Chambers J E, Chen W L et al., Common Mechanism of Toxicity: A Case Study of Organophosphorus Pesticides *Toxicol. Sci.*, 1998, **41** (1): 8—20
- [9] 许鹏翔, 袁东星, 陈猛等, 植物酶膜萃取-气相色谱分析水体中痕量有机磷农药. 环境化学, 2002, **21** (3): 296—300
- [10] Farni G R, Wilson L Y, Linear Free Energy Relationships Using Quantum Mechanical Descriptors (in): Kuojiwata K B, Boyd D B, Reviews in Computational Chemistry, Vol. 18. New York: John Wiley & Sons, 2002, 211—255
- [11] Farni G R, Penski C A, Wilson L Y, Using Theoretical Descriptors in Quantitative Structure Activity Relationships: Some Physicochemical Properties *J. Phys. Org. Chem.*, 1992, **5** (7): 395—408
- [12] Stewart J J P, MOPAC: A Semiempirical Molecular Orbital Program. *J. Comput. Aided Mol. Des.*, 1990, **4** (1): 1—105
- [13] Kleinbaum D G, Kupper L L, Muller K E et al., Applied Regression Analysis and Other Multivariable Methods. New York: Duxbury Press, 1998
- [14] R Development Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2003). ISBN 3-900051-00-3. URL <http://www.R-project.org>
- [15] 赵劲松, 王斌, 戴朝霞等, 有机磷酸酯类化合物的三维定量结构-活性相关研究. 科学通报, 2004, **49** (1): 65—70
- [16] Hansch C, Kump A, Garg R et al., Chem-bioinformatics and QSAR: A Review of QSAR Lacking Positive Hydrophobic Terms *Chem. Rev.*, 2001, **101** (3): 619—672

QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS OF ORGANOPHOSPHATE COMPOUNDS AND ROBUST ANALYSIS

ZHAO Jin-song WANG Yu WANG Hai-yan WANG Xiao-dong WANG Lian-sheng

(State Key Laboratory of Pollution Control and Resource Reuse, School of the Environment, Nanjing University, Nanjing 210093)

ABSTRACT

The relationships between the structure and acute toxicity to housefly of 35 organophosphate compounds (OP) were studied using theoretical linear solvation energy relationships (TLSEr) model. A set of specific models was obtained, which could indicate the interaction mechanism. The most important factor that influences the biological activity of OP is the polarizability given by π_1 in TLSEr model. It indicates that the specific reaction might occur. The uncertainty of TLSEr models was analyzed by bootstrap method. The bootstrap estimated R^2 , SE, and coefficients of TLSEr models show that the models are robust.

Keywords organophosphate compounds, QSAR, theoretical linear solvation energy relationships, Bootstrap