逐步回归法与 PLS-Bootstrap 法在肼衍生物与 Pu(Ⅳ)反应定量构效关系中的应用

朱晓乐,肖松涛,欧阳应根*

中国原子能科学研究院 放射化学研究所,北京 102413

摘要:钚是一种重要的核材料,在洗锝槽还原试剂的开发过程中,还原试剂与 Pu(Ⅳ)的化学反应是关键的影响因素之一,通常要求其与钚几乎不发生反应。为了建立良好的 Pu(Ⅳ)与肼衍生物反应的定量构效关系,找到该反应的特征参数,为后续工作提供支撑,采用 PLS-Bootstrap 法与逐步回归法分别对该反应进行定量构效关系研究,并使用交叉检验和外部检验来对模型进行验证。在疏水性参数作为该反应特征参数的基础上,得到了最高占据轨道能是该反应的又一个特征参数的结果,且疏水性参数与最高占据轨道能的值越大,该反应进行得越慢。

关键词:定量构效;肼;偏最小二乘法;钚

中图分类号:TL241.1 **文献标志码**:A **文章编号**:0253-9950(2020)02-0102-09

doi:10.7538/hhx.2020.42.02.0102

Application of Stepwise Regression Method and PLS-Bootstrap Method in Quantitative Structure-Activity Relationship of Hydrazine Derivatives and Pu(Ⅳ) Reaction

ZHU Xiao-le, XIAO Song-tao, OUYANG Ying-gen*

China Institute of Atomic Energy, P. O. Box 275(26), Beijing 102413, China

Abstract: Plutonium is an important nuclear material. The chemical reaction of reducing reagents with Pu(Ⅳ) is one of the key influencing factors in the development of reducing reagents in technetium scrubbing stage. It is usually required to react little with hydrazine. In order to establish a good quantitative structure-activity relationship between the reaction of Pu(Ⅳ) and hydrazine derivatives, the characteristic parameters of the reaction were found to provide support for the subsequent work. In this paper, the PLS-Bootstrap method and the stepwise regression method were used to study the quantitative structure-activity relationship of the reaction, and the cross-test and external test were used to verify the model. Based on the hydrophobic parameter as the characteristic parameter of the reaction, it is obtained that the highest occupied orbital energy is another characteristic parameter of the

收稿日期:2019-01-09;**修订日期**:2019-02-21

基金项目:国家自然科学基金资助项目(21790371)

作者简介:朱晓乐(1994—),男,湖北荆州人,硕士研究生,分析化学专业,E-mail: sogalas@qq.com

*通信联系人:欧阳应根(1964—),男,安徽太湖人,研究员,从事核燃料循环与材料研究,E-mail: 13911262687@139.com

reaction, and the larger the value of the hydrophobic parameter and the highest occupied orbital energy are, the slower the reaction proceeds.

Key words: quantitative structure-activity relationship; hydrazine; partial least square; plutonium

钚是一种重要的战略资源,是制造核武器和核燃料的关键材料。Pu 在水溶液中的化学行为比较复杂,以多价态共存的方式存在。其中 Pu(Ⅳ)易被 TBP 萃取到有机相,而 Pu(Ⅲ)不易被萃取到有机相,在 PUREX 流程中可利用这个特点将 Pu(Ⅳ)还原为 Pu(Ⅲ)实现铀钚的分离和钚的纯化。

王婷[1]、邵开元[2]、石灵娟[3]等分别对有机污染物的混合毒性、催眠类药物的毒性、多环芳烃的荧光强度进行了定量构效关系研究,均得到了理想的结果。其中邵开元[2]为了改善定量构效关系(quantitative structure-activity relationship, QSAR)模型,提出了化学势变化率这一量化参数,最终使 QSAR 相关性得到提高,降低了预测误差。

Norinder[4]、邓景景[5]和张永红[6]将偏最小二乘法(partial least square, PLS)应用于 QSAR 研究,在处理多重相关性的问题上表现良好。邓景景还指出,使用 PLS 方法建模,较多元线性回归(multiple linear regression, MLR)具有更好的拟合能力和稳定性。

PLS-Bootstrap 法用于定量构效关系研究的情况较少。金志超[7]提出,使用 PLS-Bootstrap 法对变量进行筛选,能解决受实验条件限制导致的样本数较少的问题,并可应用于化学药物的活性与亲油性等参数之间相关关系的预测。Bras[8]则将 PLS-Bootstrap 法进行了拓展,用于筛选波数间隔。虽尚未发现有文献报道将 PLS-Bootstrap 法用于 QSAR 研究工作中,但由于该方法可实现变量的筛选,因此本课题组将使用 PLS-Bootstrap 法进行 QSAR 研究。本工作将在之前的研究基础上[9],在剔除错误样本点后重新建立定量构效关系。将分别使用逐步回归法和 PLS-Bootstrap 法来建立定量构效关系模型并进行交叉检验和外部检验。

1　数据与方法

Pu(Ⅳ)与肼衍生物反应的半反应时间 $t_{1/2}$ 及它们的量化参数列入表 1 和表 2[9],肼衍生物结构示于图 1[9]。考虑到肼衍生物与 Pu(Ⅳ)的半反应时间存在着数量级差异,因此将它们转化成自然对数后再进行定量构效关系研究。肼衍生物与 Pu(Ⅳ)反应的半反应时间可用于表示该反应进行的快慢程度。半反应时间越小,该反应进行得越快。量化参数有最高占据轨道能(E_{HOMO})、最低非占轨道能(E_{LUMO})、前线轨道能量差($\Delta E=E_{LUMO}-E_{HOMO}$)、分子总能量($E$)、分子偶极矩($\mu$)、疏水性参数(lg P)、分子折射率(R)、分子摩尔体积(V)、分子表面积(A)、网格化分子表面积(G)、相对分子质量(M_r)、分子极化率(P)以及水合能(E_H)等。

表 1　肼衍生物与 Pu(Ⅳ)反应的半反应时间[9]

Table 1　Half-reaction time of hydrazine derivatives with Pu(Ⅳ)[9]

还原剂	$t_{1/2}$/s	ln[$t_{1/2}$/s]
肼	328 000	12.7
甲基肼	29 000	10.28
偏二甲基肼	520 000	13.16
对二甲基肼	499 000	13.12
乙基肼	2 550 000	14.75
烯丙基肼	3 780 000	15.14
羟乙基肼	2 110	7.65
氯乙基肼	1 270	7.15
氨基胍	9 600	9.17
乙基脂醋酸肼	3 560	7.85

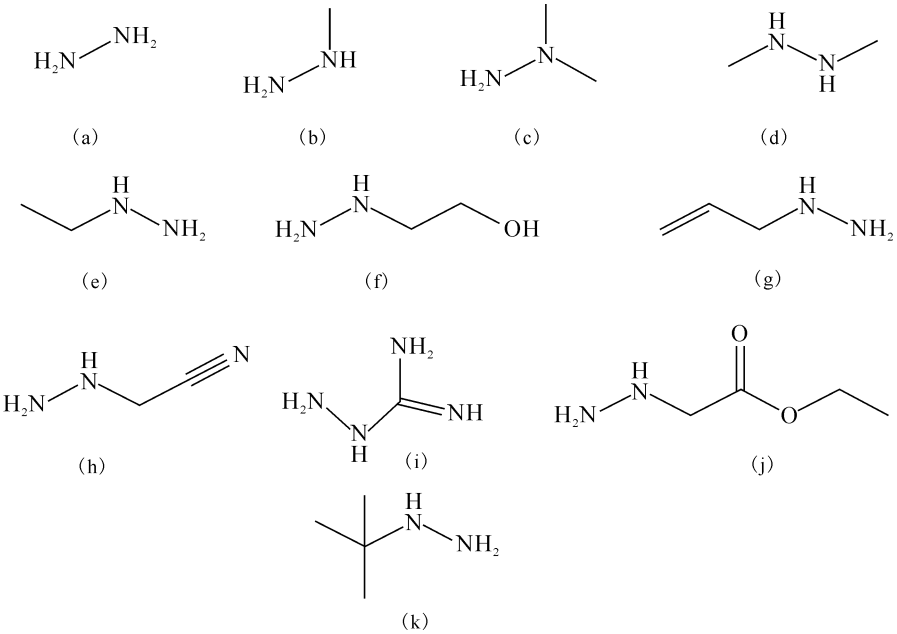
为了考察半反应时间与量化参数的相关性、量化参数之间的相关性,求解其 Pearson 矩阵,列于表 3。由于 Pearson 矩阵为对称矩阵,且主对角元的相关性系数均为 1,概率 p 值均为 0,因此仅求出其上半部分的相关性系数及对应的 p 值。由表 3 可知,半反应时间($t_{1/2}$)与最高占据轨道能(E_{HOMO})、前线轨道能量差(ΔE)、分子偶极矩(μ)、疏水性参数(lg P)有较强的线性相关性,相关系数均达到 0.6 以上,且不相关的可能性 p 均小于 0.05,即半反应时间与这些参数均存在相关性,但线性相关性并不强。此外,各参数之间也存在相

表 2 肼衍生物的量化参数[9]

Table 2 Quantitative parameters of hydrazine derivatives[9]

还原剂	E_{HOMO}/a. u.	E_{LUMO}/a. u.	ΔE/a. u.	E/a. u.	μ/D	lg P	R
肼	−0.218	0.011	0.229	−111.912	0	−0.68	10.57
甲基肼	−0.243	−0.002	0.241	−151.237	1.74	−0.46	14.28
偏二甲基肼	−0.203	−0.003	0.201	−190.554	0.564	−0.55	19.18
对二甲基肼	−0.228	−0.005	0.223	−190.555	1.736	−0.25	17.99
乙基肼	−0.242	−0.001	0.241	−190.566	1.797	−0.12	19.03
烯丙基肼	−0.213	−0.011	0.202	−228.648	0.359	0.28	23.44
羟乙基肼	−0.247	−0.005	0.241	−265.806	1.183	−0.91	20.57
氰乙基肼	−0.277	−0.018	0.260	−243.210	3.255	−0.34	19.49
氨基胍	−0.233	−0.011	0.222	−260.793	3.299	−0.64	19.81
乙基脂醋酸肼	−0.249	−0.012	0.238	−418.518	3.732	−0.55	29.88

还原剂	V	A	G	M_r	P	E_H/(kcal·mol^{-1})
肼	183.87	135.89	158.850	32.050	3.480	−21.860
甲基肼	239.45	143.59	190.220	46.070	5.310	−13.590
偏二甲基肼	291.36	134.66	217.930	60.100	7.150	−0.640
对二甲基肼	300.95	149.83	225.610	60.100	7.150	−7.400
乙基肼	295.76	143.61	223.980	60.100	7.150	−10.780
烯丙基肼	333.73	258.62	247.910	72.110	8.790	−12.800
羟乙基肼	319.84	246.11	239.980	76.100	7.780	−17.450
氰乙基肼	298.39	239.98	227.090	71.080	7.160	−15.210
氨基胍	294.44	221.08	223.120	74.090	7.600	−25.320
乙基脂醋酸肼	437.67	269.48	308.280	118.140	11.540	47.610



(a)——肼,(b)——甲基肼,(c)——偏二甲基肼,(d)——对二甲基肼,(e)——乙基肼,(f)——羟乙基肼,(g)——烯丙基肼,(h)——氰乙基肼,(i)——氨基胍,(j)——乙基脂醋酸肼,(k)——叔丁基肼

图 1 肼衍生物结构式[9]

Fig. 1 Hydrazine derivatives structure[9]

表 3　Pearson 相关性矩阵

Table 3　Pearson correlation matrix

参数	E_{HOMO}/ a. u.	E_{LUMO}/ a. u.	ΔE/ a. u.	E/ a. u.	μ/D	lg P	R	V	A	G	M_r	P	E_H/ (kcal·mol^{-1})
$t_{1/2}$/s	0.733	0.475	−0.643	0.563	−0.677	0.639	−0.269	−0.299	−0.541	−0.313	−0.484	−0.265	−0.207
	0.016	0.166	0.045	0.090	0.032	0.047	0.452	0.401	0.106	0.379	0.156	0.460	0.566
E_{HOMO}/a. u.		0.514	−0.933	0.392	−0.720	0.158	−0.203	−0.250	−0.419	−0.271	−0.351	−0.200	−0.110
		0.129	0.000	0.263	0.019	0.662	0.573	0.486	0.228	0.450	0.320	0.580	0.763
E_{LUMO}/a. u.			−0.173	0.681	−0.690	−0.266	−0.711	−0.687	−0.752	−0.693	−0.709	−0.713	−0.268
			0.633	0.030	0.027	0.458	0.021	0.028	0.012	0.026	0.022	0.021	0.455
ΔE/a. u.				−0.167	0.545	−0.291	−0.062	0.002	0.159	0.022	0.109	−0.068	0.030
				0.644	0.104	0.414	0.866	0.996	0.660	0.951	0.764	0.853	0.935
E/a. u.					−0.691	0.113	−0.928	−0.933	−0.802	−0.939	−0.993	−0.925	−0.736
					0.027	0.756	0.000	0.000	0.005	0.000	0.000	0.000	0.015
μ/D						−0.141	0.518	0.549	0.465	0.552	0.662	0.539	0.412
						0.698	0.125	0.100	0.175	0.098	0.037	0.108	0.237
lg P							0.195	0.152	0.057	0.154	−0.017	0.191	−0.009
							0.590	0.674	0.875	0.672	0.962	0.596	0.980
R								0.988	0.760	0.989	0.965	0.996	0.732
								0.000	0.011	0.000	0.000	0.000	0.016
V									0.717	0.999	0.969	0.993	0.773
									0.020	0.000	0.000	0.000	0.009
A										0.740	0.791	0.735	0.316
										0.014	0.006	0.015	0.374
G											0.972	0.992	0.763
											0.000	0.000	0.010
M_r												0.964	0.749
												0.000	0.013
P													0.730
													0.017

关关系，如最高占据轨道能（E_{HOMO}）与前线轨道能量差（ΔE）存在强相关性，与偶极矩存在较强的相关性；最低非占轨道能（E_{LUMO}）则与前线轨道能量差（ΔE）、疏水性参数（lg P）、水合能（E_H）之外的其他量化参数均存在较强的相关性。即量化参数之间存在着较强的相关性，无法直接进行回归建模，因此将分别使用逐步回归法和 PLS-Bootstrap 法对量化参数进行筛选，再通过回归分析的方法建立定量构效关系模型，且半反应时间与量化参数之间存在相关性但线性相关性并不强，故还需要考虑非线性定量构效关系模型。

逐步回归法是结合了变量筛选与回归的一种方法，在每一步中，都要分别考查所有的参数与因变量的相关性，将未引入方程中但相关性最显著的参数引入方程，即将 p 值最小的参数引入方程，而将方程中相关性不够强的参数剔除掉。为防止被剔除的参数能再次被引入而形成死循环，一般设置剔除参数的 p 临界值（p_{remove}）要大于进入方程的 p 临界值（p_{enter}）。一般默认设置为 $p_{enter}=0.05$，$p_{remove}=0.10$。逐步回归法的结果是否有意义需要由 F 检验确定，各参数与因变量的相关性是否显著则应当由 t 检验确定。

PLS 被认为是一种能较好处理自变量共线性和样本数较少的回归方法，在对每一个成分进行回归时都要完成提取主成分、相关性分析和回归三个步骤。首先对自变量与因变量各提取一个主成分，并使自变量的主成分与因变量的主成分相关性最大，其次使用自变量的主成分对因变量进行回归，然后分别求得经过回归之后自变量与因变量的残差，并使用残差分别代替自变量与因变量重复前述步骤进行回归。一般情况下不需要选择所有的成分来建立回归方程，仅需要选用前几个成分即可得到预测能力较好的回归方程。需要注意的是，使用 PLS 进行回归时所有的参数都会被引入方程，其中可能会包括部分相关性不显著的参数，考虑到 Bootstrap 作为一种检验方法，能较好地处理样本数较少的情况，通过蒙特卡罗随机抽样的方法建立 Bootstrap 样本，然后对 Bootstrap 样本进行回归，将回归系数按大小顺序排列并选定一定的分位数，若该数大于原样本回归方程中的回归系数，则该参数应当从方程中剔除。将 PLS 与 Bootstrap 结合起来进行回归分析时，可以将相关性不显著的参数从方程中剔除掉，使方程的意义更明确。由于 Bootstrap 本身已包含了对各参数的检验，故无需对各参数进行 t 检验。

本工作使用了 Matlab 2014b 软件来进行回归处理，在逐步回归法中调用了程序自带的 stepwise 和 stepwisefit 函数来进行逐步回归，在 PLS-Bootstrap 法中调用了程序自带的 plsregress 函数实现 PLS 的部分，使用 rand 函数结合 for 循环语句，按照文献[7]的步骤来实现 Bootstrap 的部分。

2 结果与讨论

2.1 逐步回归法对参数的筛选及模型的建立

利用逐步回归法进行回归分析，在完成回归分析任务的同时，也完成了变量的筛选。在默认的设置（$p_{enter}\leqslant 0.05$，$p_{remove}>0.10$）下，对其进行逐步回归，得到的结果列入表 4。由表 4知，第 1、2、6 个变量，即最高占据轨道能（E_{HOMO}）、最低非占轨道能（E_{LUMO}）和疏水性参数（lg P）保留在方程中，方程为 $t_{1/2}=54.939\ E_{HOMO}+176.841\ E_{LUMO}+6.363\ \lg P+27.718$。其方程的分子自由度为 3，分母自由度为 6，$F$ 统计量值为 37.719，即 $F(3,6)=37.719>4.76=F_{0.05}(3,6)$，即在 95% 的置信度上该回归方程有意义，对应的调整相关系数即调整 $r^2=0.924$。标准化回归方程为 $y=0.384x_1+0.463x_2+0.701x_6$。标准化回归系数的绝对值反映了该自变量对因变量的影响大小，故疏水性参数是该反应的主要特征参数，且疏水性参数越大，半反应时间越大，该反应进行得越慢；最高占据轨道能（E_{HOMO}）和最低非占轨道能（E_{LUMO}）是该反应的次要特征参数，且它们的值越大，半反应时间越大，该反应进行得越慢。

下面考虑非线性关系，由于参数较多，考虑指数形式，假设存在指数关系，则有

$$y=Ae^{(Bx_1+Cx_2+Dx_3)}=e^{(Bx_1+Cx_2+Dx_3+A')}$$

其中 $A'=\ln A$。两边同取对数得：

$$\ln y=A'+Bx_1+Cx_2+Dx_3$$

若将半反应时间取自然对数，则其对数值与变量依旧呈线性关系。按此方法处理后进行回归分析，得非线性回归方程 $t_{1/2}=e^{(5.543E_{HOMO}+17.201E_{LUMO}+0.560\lg P+4.010)}$，方程的调整 $r^2=0.928$，F 统计量为 39.964$>$4.76$=F_{0.05}(3,6)$。标准化回归方程 $t_{1/2}=e^{(0.413x_1+0.479x_2+0.656x_3)}$。疏水性参数（lg P）仍为该反应的主要特征参数，最高占据轨道能（E_{HOMO}）和最低非占轨道能（E_{LUMO}）

表 4　逐步回归结果

Table 4　Stepwise regression results

变量	相关系数	标准差	状态	$p_{变量}$	分母自由度	分子自由度	总平方和	残余平方和	F 统计值	$p_{方程}$	均方根误差	常数项
E_{HOMO}	54.939	16.324	In	0.015	6	3	82.993	4.179	37.719	0.000	0.835	27.718
E_{LUMO}	176.841	44.650	In	0.007								
ΔE	−39.146	521.525	Out	0.943								
E	0.002	0.006	Out	0.800								
μ	0.142	0.419	Out	0.750								
lg P	6.363	0.923	In	0.000								
R	0.003	0.089	Out	0.977								
V	0.001	0.007	Out	0.910								
A	−0.009	0.007	Out	0.255								
G	0.001	0.011	Out	0.951								
M_r	−0.003	0.020	Out	0.908								
P	0.028	0.218	Out	0.904								
E_H	−0.006	0.015	Out	0.732								

注:“In”为变量在方程中出现,“Out”为变量在方程中未出现

仍为该反应的次要特征参数,且它们的值越大,半反应时间越大,该反应进行得越慢。

2.2　PLS-Bootstrap 法对参数的筛选及模型的建立

在使用 PLS 进行回归分析时,需要确定提取的成分数。一般情况下,当所选取成分包含的信息量达到 85%时,便认为信息已经提取完全,因此,把 X 信息量与 Y 信息量均达到 85%时的最小成分数作为 PLS 回归分析时提取成分数,对所有样本进行处理,结果列入表 5。

表 5　PLS 回归分析的成分数与信息量关系

Table 5　Relationship between number of components and amount of information in PLS regression analysis

成分数	第一次回归		第二次回归	
	X 信息量	Y 信息量	X 信息量	Y 信息量
1	0.555	0.574	0.515	0.786
2	0.801	0.872	0.813	0.889
3	0.877	0.946	0.999	0.890
4	0.945	0.963	1.000	0.897
5	0.962	0.989	-	-
6	0.990	0.990	-	-
7	0.998	0.992	-	-
8	1.000	0.994	-	-
9	1.000	1.000	-	-

注:1) “-”,在第二次回归中,当成分数为 4 时,X 信息量已达到 1,继续增加成分数也不会提升信息量,故无需对更多成分数的情况进行计算

在第一次筛选中,成分数为 2 时,包含的 X 信息量仅为 0.801,而 Y 信息量达到了 0.872;当成分数为 3 时,包含的 X 信息量为 0.877,Y 信息量为 0.946,因此,选定 PLS 提取的成分数为 3,进行 Bootstrap 检验,结果列入表 6。在经过第一次筛选后,仅有最高占据轨道能、分子总能量、疏水性参数和相对分子质量被保留。然后进行第二次的 PLS-Bootstrap 法筛选,首先进行 PLS 回归分析(表 5),确定提取的成分数为 3,然后再进行 Bootstrap 检验,结果列入表 6。由表 6 可知,最终仅有最高占据轨道能和疏水性参数被保留下来。下面将使用经过两次筛选后仍被保留的参数进行回归分析。使用最高占据轨道能和疏水性参数对半反应时间进行回归分析,得方程 $t_{1/2}=92.673E_{HOMO}+4.866\ \lg P+34.956$。调整 $r^2=0.766$,统计量 $F=15.730>4.74=F_{0.05}(2,7)$,即在 95%的置信度上该回归方程有意义,模型示于图 2。所有的 PLS-Bootstrap 法的分位数均选为 0.95。

以同样的方法对半反应时间进行处理以考虑其非线性关系,并对其进行回归分析,得非线性回归方程 $t_{1/2}=e^{(0.195\ 8E_{HOMO}+0.138\ 6\lg P+2.371)}$,调整 $r^2=0.758$,模型示于图 3。比较其调整 r^2,显然线性模型较非线性模型略优。

表 6 PLS-Bootstrap 法对参数进行筛选的结果

Table 6 Results of filtering parameters by PLS-Bootstrap method

参数	第一次筛选			第二次筛选		
	临界值	回归绝对值	是否保留	临界值	回归绝对值	是否保留
E_{HOMO}/a. u.	0.148	0.177	保留	0.532	0.541	保留
E_{LUMO}/a. u.	0.225	0.137	排除	-	-	-
ΔE/a. u.	0.188	0.145	排除	-	-	-
E/a. u.	0.098	0.151	保留	0.194	0.185	排除
μ/D	0.201	0.142	排除	-	-	-
lg P	0.363	0.576	保留	0.373	0.530	保留
R	0.101	0.046	排除	-	-	-
V	0.095	0.045	排除	-	-	-
A	0.238	0.232	排除	-	-	-
G	0.099	0.041	排除	-	-	-
M_r	0.075	0.089	保留	0.199	0.104	排除
P	0.088	0.054	排除	-	-	-
E_H/(kcal·mol^{-1})	0.209	0.029	排除	-	-	-

注:1) "-",由于第一次筛选已经将对应的变量排除在方程外,它们将不会参与第二次筛选,故其对应值均不存在

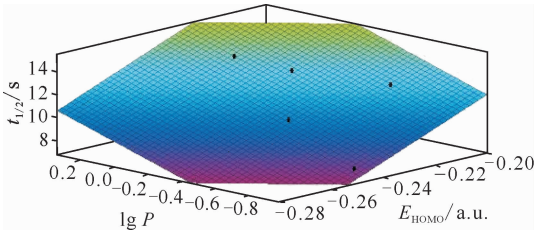


图 2 PLS-Bootstrap 法线性模型

Fig. 2 Linear model of PLS-Bootstrap method

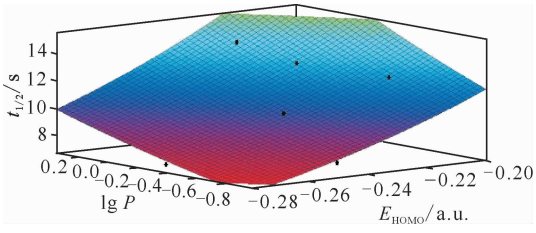


图 3 PLS-Bootstrap 法非线性模型

Fig. 3 Nonlinear model of PLS-Bootstrap method

无论是线性模型还是非线性模型,结果均表明,最高占据轨道能(E_{HOMO})和疏水性参数(lg P)是该反应的特征参数,且它们的值越大,该反应的半反应时间越大,即反应进行得越慢。

2.3 模型的交叉检验与外部检验

为了对模型进行比较,同时也对前人定量构效研究进行补充,将对两种方法建立的模型进行留一法(leave one only, LOO)交叉检验,并对被剔除的样本点叔丁基肼的部分量化参数进行订正后作为外部检验点进行检验[10]。

使用 LOO 交叉检验后得到的结果列入表 7。由表 7 可知:两种方法的交叉检验系数(Q^2)均大于 0.5,其中逐步回归法的 Q^2 达到了 0.820,表明逐步回归法建立的线性模型稳定性较优秀,因此

表 7 逐步回归法及 PLS-Bootstrap 法交叉检验结果

Table 7 Stepwise regression method and PLS-Bootstrap method cross test results

内部检验点	预测误差平方值	
	逐步回归法	PLS-Bootstrap 法
肼	2.732	2.924
甲基肼	0.984	0.008
偏二甲基肼	1.450	0.263
对二甲基肼	0.219	0.349
乙基肼	4.959	12.747
烯丙基肼	5.502	16.731
羟乙基肼	0.074	0.000
氰乙基肼	0.000	1.213
氨基胍	0.159	1.621
乙基脂醋酸肼	0.516	2.571
交叉检验系数(Q^2)	0.820	0.583

就稳定性而言,逐步回归法建立的线性模型更好。由于逐步回归法建立的线性模型与非线性模型相比无明显差别,调整 r^2 分别为 0.924 和 0.928,因此不再单独对其非线性模型进行讨论,即使非线性模型从调整 r^2 上看比线性模型略优。

由于两种方法仅用到了最高占据轨道能(E_{HOMO})、最低非占轨道能(E_{LUMO})和疏水性参数(lg P),使用了参考文献[8]的逐步回归法,在对叔丁基肼的结构进行校正后重新计算了这些参数的值,结果如下:最高占据轨道能 $E_{HOMO}=-0.237$,最低非占轨道能 $E_{LUMO}=-0.006$,疏水性参数 lg P=0.370,半反应时间对数值 ln $t_{1/2}$=16.240。使用两种线性模型分别对 ln $t_{1/2}$ 进行预测,并与观测值进行对比,得到逐步回归模型预测值为 15.991,PLS-Bootstrap 模型预测值为 14.793,结果列入表 8。由表 8 可知,逐步回归法对外部样本点的预测误差更小,误差不足 2%,而 PLS-Bootstrap 模型的预测误差已接近 10%。就叔丁基肼作为外部检验点而言,逐步回归法表现更优秀,但外部检验点样本点数量较少,并不意味着逐步回归法的预测能力一定强于 PLS-Bootstrap 法。

表 8　外部检验结果

Table 8　External inspection results

模型	ln[$t_{1/2,预测}$/s^{-1}]	误差绝对值	误差相对值
逐步回归	15.991	−0.249	−1.533%
PLS-Bootstrap	14.793	−1.447	−8.910%

总体上来看,逐步回归模型中包含了三个参数,且交叉检验、外部检验结果均优于 PLS-Bootstrap 模型,而后者虽然只包含了两个参数,但从 Pearson 相关性矩阵可以看出,这两个参数均与半反应时间有较强的相关性,而差异之处在于相关性相对较弱的最低非占轨道能(E_{LUMO}),这正好表明了通过提取成分进行相关性分析的 PLS 更能有效地保证自变量与因变量之间的相关性。

2.4　两种回归方法的比较

为了进一步探讨最低非占轨道能(E_{LUMO})对模型的影响,以比较两种回归方法的异同,下面将使用相同的变量分别建立回归方程。由于参与回归的变量已经被指定,逐步回归法无需再进行变量筛选,因此可使用最小二乘回归法代替逐步回归法进行比较,而 PLS-Bootstrap 法无需再进行变量筛选,在回归过程中也无须进行 Bootstrap 检验。对比实验一指定的变量有疏水性参数、最高占据轨道能和最低非占轨道能,对比实验二指定的变量有疏水性参数和最高占据轨道能。

对比实验一中,逐步回归法回归方程为 $t_{1/2}$=54.939 E_{HOMO}+176.841 E_{LUMO}+6.363 lg P+27.718,PLS 回归方程为 $t_{1/2}$=70.866 E_{HOMO}+126.893 E_{LUMO}+5.916 lg P+30.992;对比实验二中,逐步回归法回归方程为 $t_{1/2}$=92.673 E_{HOMO}+4.866 lg P+34.956,PLS 回归方程为 $t_{1/2}$=92.673 E_{HOMO}+4.866 lg P+34.956;对比实验一与实验二的各参数列于表 9 。由表 9 可知,当变量数为三个时,PLS 在回归时认为仅需提取两个成分数即可建立效果较好的回归方程,当所有的成分数均提取完全即将三个成分全部提取时,其回归结果与逐步回归法的结果相同,即第三个成分对改善模型的拟合优度,即调整 r^2 的贡献不大。对比可知,第三个成分对调整 r^2 的提高仅为 0.004,较 0.920 提高了 0.43%,表明了 PLS 能选取合适的成分数进行回归的同时,舍弃对回归结果作用不明显的部分。但就对比实验一中对逐步回归法与 PLS 在 LOO 交叉检验的结果进行对比是不合理的,因为逐步回归法在交叉检验时的变量数为 3,而 PLS 由于仅提取了两个成分,在交叉检验时仅有两个成分进行回归,即变量数为 2,故无法简单地将其进行比较。而考虑变量数相同的情况时,在对比实验二中,逐步回归法交叉检验时变量数为 2,而 PLS 在对比实验一中,交叉检验时提取两个成分,变量数也为 2,对比交叉检验系数 Q^2,可得出 PLS 的模型稳定性更好。综上,对比逐步回归法与 PLS,可认为 PLS 的回归结果更好。

表 9　对比实验一与实验二的回归情况

Table 9　Comparing regression of experiment 1 and experiment 2

参数	对比实验一		对比实验二	
	逐步回归	PLS[1)]	逐步回归	PLS
调整 r^2	0.924	0.920	0.766	0.766
F 统计量	37.719	52.776	15.730	15.730
交叉检验系数(Q^2)	0.820	0.676	0.583	0.583

注:1) 回归时提取两个成分,F 统计量为 $F(2,7)$=52.776

将对比实验二的逐步回归法与对比实验一中的 PLS 在交叉检验中的表现进行对比,虽然存在一定的不合理性,但由对比实验二可以看出,PLS

在提取所有的成分数的情况下与逐步回归法有相同的结果。将两个对比实验中 PLS 的表现进行对比，两种情况下 PLS 均提取了两个成分进行回归，差别在于成分中是否包含 E_{LUMO} 这一参数。由调整 r^2 的比较可知，该参数的引入对模型有较大的改善，以调整 r^2 的值来看，拟合优度的改善程度为 20.10%，但以 Q^2 来看，模型稳定性的提高程度却仅为 16.95%。即使这样的比较不准确，却仍能说明 E_{LUMO} 参数的引入对模型拟合效果的改善作用更大，即 E_{LUMO} 参数引入的作用存在着过分追求方程拟合效果的可能性。

综上所述，在肼衍生物与 Pu(Ⅳ)的反应中，PLS-Bootstrap 可使用较少的变量来描述整体的趋势变化，而逐步回归法能获得预测效果更好的模型，两者的结果具有相似性，即两种方法均可得到反映该反应规律的回归方程。就肼衍生物与 Pu(Ⅳ)反应而言，E_{LUMO} 是否能作为该反应的特征参数，还需进一步研究确定。

3 结 论

通过逐步回归法与 PLS-Bootstrap 法分别对 Pu(Ⅳ)与肼衍生物反应的定量构效关系进行了研究，获得了可描述该反应进行快慢的特征参数：最高占据轨道能(E_{HOMO})、疏水性参数(lg P)，其中疏水性参数是该反应的主要特征参数，且它们的值越大，半反应时间越大，该反应进行得越慢。而作为次要特征参数的最低非占轨道能(E_{LUMO})能否作为该反应的特征参数仍需进一步研究确定，因此两种方法得到的结果具有相似性。

对之前的 Pu(Ⅳ)与肼衍生物反应的定量构效关系研究的不足之处进行了补充，对错误的样本点进行了校正并作为外部检验点对模型进行了检验，完成了之前模型未进行的交叉检验等检验工作。

本工作使用 PLS-Bootstrap 法进行定量构效关系研究并进行了交叉检验与外部检验，结果表明，该模型的稳定性较逐步回归模型差，考虑到外部检验点数量少，其预测能力还需进一步研究确定。

参考文献：

[1] 王婷. 有机污染物的混合毒性 QSAR 模型及其机制研究进展[J]. 科学通报，2015，60(19)：1771-1780.

[2] 邵开元. 化学势变化率对催眠药类化合物 QSAR 影响[J]. 化学通报，2017，80(11)：1061-1066.

[3] 石灵娟. 多环芳烃结构和荧光强度的 QSAR 研究[J]. 计算机与应用化学，2004，21(5)：773-776.

[4] Norinder U. A PLS quantitative structure-activity relationship study of some monoanime oxidase inhibitors of the phenyl alkylamine type[J]. Eur J Med Chem, 1994, 29(3): 191-195.

[5] 邓景景. 应用 PLS-QSAR 模型预测纳米金属氧化物的毒性[J]. 癌变·畸变·突变，2015，27(5)：399-403.

[6] 张永红. PLS 变量筛选法用于有机物透聚乙烯膜性能 QSAR 研究[J]. 化学学报，2011，69(10)：1232-1238.

[7] 金志超. 偏最小二乘回归系数的 Bootstrap 假设检验及 SAS 实现[J]. 中国卫生统计，2009，26(4)：340-343.

[8] Bras L P. A Bootstrap-based strategy for spectral interval selection in PLS regression[J]. J Chemometrics, 2008, 22: 695-700.

[9] 肖松涛. 肼衍生物与 Pu(Ⅳ)的氧化还原反应定量构效关系研究[J]. 核技术，2016，39(8)：1-10.

[10] 覃礼堂. QSAR 模型内部和外部验证方法综述[J]. 环境化学，2013，32(7)：1205-1211.