大 学 化 学
Univ. Chem. **2025,** *40* (8), 345

•自学之友•　　doi: 10.12461/PKU.DXHX202410107　　www.dxhx.pku.edu.cn

定量构效关系方法学习探索：以钴卟啉活化氧气为例

张学鹏*，龙宇驰，潘禹澍，王继顶，白宝钰，丁瑞
陕西师范大学化学化工学院，西安 710119

摘要：金属卟啉配合物由于在氧气还原反应中具有优异的催化活性以及良好的反应选择性，受到了人们的广泛关注。但是，常规的实验合成表征或高精度量化计算较难大批量研究其取代基效应。为此，本文基于定量构效关系(QSAR)方法，利用Chem3D、Gaussian等软件计算了不同取代基钴卟啉配合物的拓扑参数、量化参数，并结合现有物化参数用SPSS进行相关性分析，可以快速筛选出影响钴卟啉活化氧气分子的特征描述符。并且，本文进一步采用逐步回归分析得到了多元回归方程，其具有较好的拟合优度以及泛化能力。本文详细阐述了QSAR中代表性描述符的计算与采集，并展示了常用的相关性分析与回归分析过程，旨在帮助学生自学Chem3D、Gaussian、SPSS等软件，且可以有效提高学生在分子建模、计算化学、统计学软件等方面的实际操作能力以及数据分析和处理能力。

关键词：钴卟啉；氧气结合能；定量构效关系；相关性分析；逐步回归分析
中图分类号：G64；O6

Exploring Quantitative Structure-Activity Relationship Methods: A Case Study on Oxygen Activation by Cobalt Porphyrins

Xue-Peng Zhang *, Yuchi Long, Yushu Pan, Jiding Wang, Baoyu Bai, Rui Ding
School of Chemistry and Chemical Engineering, Shaanxi Normal University, Xi'an 710119, China.

Abstract: Metalloporphyrin complexes have attracted much attention because of their excellent catalytic performance and good selectivity in oxygen reduction reactions (ORR). However, massive investigations on substituent effects with conventional experimental synthetic approaches as well as high accuracy quantum chemistry computations would be difficult. Herein, quantitative structure-activity relationship (QSAR) methods were utilized. Based on collected physicochemical parameters, calculated topological parameters and quantum-chemical parameters of cobalt porphyrins with various substituents by Chem3D and Gaussian, the correlation analysis was performed by SPSS program, which enables quick screening of feature descriptors on dioxygen activation mediated by cobalt porphyrins. Subsequently, a stepwise regression analysis was performed, which outputs a multi-variant regression equation with acceptable goodness-of-fit and generalization ability. Our work systematically demonstrated the calculation and collection of representative descriptors in QSAR, and the detailed process of correlation analysis and regression analysis were shown. This work can assist in self-learning the relevant functions of Chem3D, Gaussian, and SPSS software, enhancing practical abilities in molecular modeling, computational chemistry and statistical software, as well as data analysis and processing skills.

Key Words: Cobalt porphyrin; Oxygen binding energy; QSAR; Correlation analysis; Stepwise regression analysis

收稿：2024-10-30；录用：2024-12-23；网络发表：2025-03-21
*通讯作者，Email: zhangxp@snnu.edu.cn
基金资助：国家自然科学基金(22003036)

1 前言

能源存储与转换技术的持续发展是目前解决能源短缺问题的重要策略，其中氢氧燃料电池、金属-空气电池以及电催化水分解制氢等因其高能效和环保特性而备受青睐。氧气还原反应(oxygen reduction reaction，ORR)作为上述过程的关键化学反应，研究其反应机理与催化剂设计具有重要意义[1]。在自然界中，细胞色素C氧化酶可以高效催化氧气还原反应，而其催化活性中心被认为是具有氧化还原能力的铁卟啉基团[2]。受自然界启发，大量过渡金属卟啉配合物被发现具有ORR催化活性[3]。相比于昂贵的贵金属基催化剂，较为廉价的过渡金属卟啉配合物同时还具有：1) 刚性的配位环境，能够稳定高价的金属离子；2) 丰富的取代位点，卟啉*meso*位、*β*位以及中心金属轴向配体均可修饰进而调控其催化活性；3) 特征的光谱信号，便于进行催化反应的原位检测和表征。基于以上优势，过渡金属卟啉配合物已被广泛用于氧气还原反应的催化机理研究以及潜在工业应用[4–9]。

一般认为，氧气分子与催化剂活性中心的低价态金属中心成键并得到电子发生活化，这一过程的氧气结合能是评估ORR催化活性的关键因素。研究表明，钴卟啉配合物催化ORR的反应机理通常为低价态的Co^{2+}金属中心直接与O_2键合并向其提供1个电子，生成超氧物种$Co^{3+}-O_2^{\bullet-}$中间体[10,11]，接着进一步发生单电子还原生成过氧物种$Co^{2+}-O_2^{\bullet-}$中间体，该中间体可以发生远端氧质子化继而O－O断键生成4电子还原产物H_2O，或者近端氧被质子化得到2电子还原产物H_2O_2 [12]。钴卟啉配合物的氧气活化能力与卟啉配体*meso*位和*β*位取代基种类息息相关，但是目前常用的实验合成表征手段以及高精度量化计算在繁杂多变的卟啉配体取代下，均存在工作量巨大、实验成本高昂的情况，难以系统地研究其取代基效应。

定量构效关系(Quantitative Structure-activity Relationship，QSAR)方法，又叫定量结构-性质关系(Quantitative Structure-property Relationship，QSPR)方法，是一类用于定量描述一批已知分子的结构与其生理/环境/物理/化学等方面的目标效应之间的数学关系，并预测另一批已知结构但未知其目标效应的研究方法，可在加速化合物开发、降低科研成本以及提升预测精确度等方面呈现明显的优势[13–15]。20世纪60年代，构效关系开始由定性转向定量并应用于药物化学[16]。随着方法学和计算科学的飞速发展，QSAR方法已广泛应用于生命科学、环境科学、材料科学等诸多领域[17,18]。对于分子体系，QSAR研究的关键是构建数学函数：$y = f(x_1, x_2, \cdots, x_m)$。其中，$y$表示所研究分子的相关性质函数(因变量)，$x_1, x_2, \cdots, x_m$为所研究分子的结构参数和特性量(自变量)。所以，QSAR可以通过测定化合物的相关参数，探索其与分子结构的相互作用规律，实现利用得到的关系式凭借基本的化学结构即可合理预测化合物的性质[19]。例如：清华大学罗三中课题组利用不同溶剂中的实测酸解离常数pK_a数据，以结构-物理化学描述符作为自变量，建立了适用于多种溶剂的pK_a预测模型[20]。中国科学院孙文华团队选取了一批用于乙烯聚合的双(亚胺)吡啶铁/钴催化剂，用一种兼具变量筛选与回归拟合功能的算法分别得出了适于拟合催化活性、产物分子量及熔点的模型[21]。圣彼得堡国立大学Buglak等人通过密度泛函理论(Density Functional Theory，DFT)计算了一系列卟啉类光敏剂，并建立了线性及非线性回归模型以推测选定自变量与单线态氧的量子产率的数学关系[22]。

本文基于QSAR方法，以不同取代基钴卟啉配合物为研究对象，用常见的化学软件获取分子描述符，首先获取各描述符之间的相关性，之后建立筛选自变量与因变量之间物理意义明确的数学方程，进而定量研究取代基对钴卟啉配合物活化氧气分子的影响。通过本论文钴卟啉活化氧气的实例学习探索，学生将能够使用Chem3D计算拓扑参数，用Gaussian进行分子体系的量化计算来获取量化参数，并使用定量构效关系对参数进行相关性分析和逐步回归分析，提高数据分析和处理问题的能力，进而实现对定量构效关系方法的学习和探索。

2 分子描述符以及QSAR分析方法

2.1 分子描述符

分子描述符含编码为符号的分子信息，其可转化为有用的数字或一些标准化实验的结果，用于

分析分子的性质和特征[14,15,23,24]。本文使用如图1所示的三类分子描述符：拓扑描述符、量化描述符和物化描述符。过去极性表面积(Polar Surface Area，PSA)须通过量化计算的方式来获取，后来Ertl等提出了一种基于分子拓扑结构(Topological Structure)估算的方法，在Chem3D中即可获取[25]，因此下文也将其归为拓扑描述符内。

图1　定量结构性质主要研究步骤

2.1.1　拓扑描述符

拓扑描述符根据分子拓扑结构建立邻接矩阵、距离矩阵后所计算得出，用以反映分子的大小、形状、分支、杂原子、不饱和键等结构特征，从而实现分子结构信息的数值化。近年来，在研究饱和烷烃的理化性质[26]和无机含氧酸的酸性[27]等方面，分子拓扑学作为一个有前景的工具得到了广泛应用[28]。其优势在于形式简明、计算简单。拓扑指数的计算一般包括三个步骤：(1) 用化学拓扑图(分子结构图)表示分子即隐氢图，下面以对氯苯酚为例简要介绍，如图2所示，重原子用数字编号表示，原子之间键长用小写英文字母编号表示；(2) 用矩阵表示分子的化学拓扑图；(3) 分子的化学拓扑图的数值化，即对分子的某种矩阵进行数学运算，从而得到一个或一组参数。常用到邻接矩阵和距离矩阵。

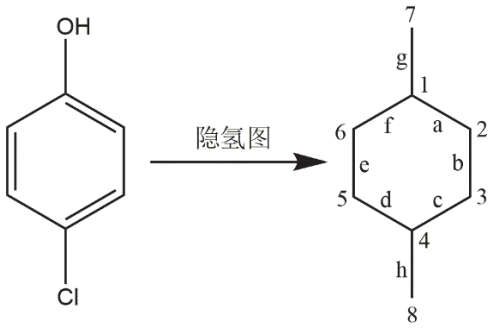

图2　对氯苯酚的分子结构图表示

① 邻接矩阵

1936年，König将邻接矩阵$\boldsymbol{A}$引入化学分子图，对于任一有n个顶点(非氢原子，本节下同)的分子图，可构成一个$n \times n$阶矩阵。矩阵中第i行第j列元素——矩阵元a_{ij}定义为[29]：

$$a_{ij}=\begin{cases}1 & \text{顶点邻接}\\ 0 & \text{其他(包括不邻接和重合)}\end{cases}$$

其邻接矩阵$\boldsymbol{A}$用于表示图中顶点之间相邻关系的一种矩阵，其中的矩阵元表示图中节点之间的连接关系。若顶点i与顶点j之间有边相连，则$a_{ij}=1$；反之，则$a_{ij}=0$。

② 距离矩阵

1947年，Wiener将距离矩阵$\boldsymbol{D}$引入，对于任一有n个顶点的分子图，可构成一个$n\times n$阶矩阵。矩阵元d_{ij}定义为[30]：

$$d_{ij}=\begin{cases}d & \text{连接顶点最小边数}\\ \infty & \text{当顶点时，两点不连通}\end{cases}$$

其距离矩阵$\boldsymbol{D}$用于表示图中各节点之间距离的一种矩阵形式，其中的矩阵元表示节点之间的最短路径长度或距离。若顶点i与顶点j之间有最短距离相连，则$d_{ij}=d$；反之，则d_{ij}设置为∞。

本文主要介绍如下拓扑描述符，均基于Chem3D获取，如表1所示。

表1　拓扑描述符以及其相关含义*

符号	表达式	含义
$W(G)$	$W(G)=\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}d_{ij}$	结构G的Wiener指数，等于距离矩阵$\boldsymbol{D}$所有矩阵元之和的一半，即所有非氢原子之间的距离之和，可表征同分异构烷烃分子的支链性及沸点、表面张力等性质[30]
J	$J=\frac{q}{(\mu+1)\sum_{1\le i<j\le n,\ a_{ij}=1}\sqrt{s_i s_j}}$	Balaban指数，q是非氢成键原子对数，环数$\mu=q-n+1$，s_i是距离矩阵$\boldsymbol{D}$第i行矩阵元之和，在Wiener指数的基础上还能反映分子内的成环性[31]
TC	$\mathrm{TC}=\frac{1}{\sqrt{\prod_{i=1}^{n}\left(\sum_{j=1}^{n}a_{ij}\right)}}$	考虑所有非氢原子点价的分子总连接性(total connectivity，TC)指数，每个原子的点价指与其共价键合的非氢原子数，即顶点度数，作用与Wiener指数类似，但区分度更高[32]
TVC	$\mathrm{TVC}=\frac{1}{\sqrt{\prod_{i=1}^{n}\delta_i^{v}}}$	在TC基础上修正点价的分子总价连接性(total valence connectivity，TVC)指数，修正后的点价称为价点价δ^{v}，可区分不同类型的杂原子，并考虑了不饱和键影响[32]
MTI	$\mathrm{MTI}=\sum_{i=1}^{n}\boldsymbol{E_i}$	Schultz分子拓扑指数(Molecular Topological Index，MTI)，其中每个非氢原子的$\boldsymbol{E}_i$综合考虑了邻接矩阵、距离矩阵及顶点度数对其的影响，对烷烃分子沸点也有较好预测性[33]
PSA	–	氧、氮原子以及与之结合的氢原子组合的在不同化学环境下的片段表面积，其中TPSA无需考虑片段构象，有较好的脂溶性预测效果[25]

*其中价点价$\delta^{v}=\frac{Z-h}{Z-Z^{v}-1}$，$Z$为核外总电子数，价电子数$Z^{v}=\sigma+\pi+\mathrm{LP}$，即$\sigma$电子、$\pi$电子及孤对电子数三者之和，$h$为相连H原子数

2.1.2　量化描述符

量化描述符是由量子化学理论计算出来的物理化学参数，以下量化描述符是基于Gaussian获取。主要如表2所示。

2.1.3　物化描述符

物化描述符是指用来描述分子物理和化学性质的数学参数，广义上与量化描述符存在一定的交集，故本文的物化描述符特指不经过量化计算获取的。其中，脂水分配系数通过Chem3D计算获取，而典型取代基的Hammett参数可根据查表得到。主要如表3所示。

表2　量化描述符及其相关含义

量化描述符	单位	符号及表达式	含义
垂直电离势	eV	$\mathrm{VIP}=-\varepsilon_{\mathrm{HOMO}}$	(第一)垂直电离势(Vertical Ionization Potential，VIP)，是最高占据轨道能量的相反数，反映失电子难易程度
垂直电子亲和能	eV	$\mathrm{VEA}=-\varepsilon_{\mathrm{LUMO}}$	(第一)垂直电子结合能(Vertical Electron Affinity，VEA)，是最低未占据轨道能量的相反数，反映得电子难易程度
Mulliken电负性	eV	$\chi=-\frac{(\varepsilon_{\mathrm{HOMO}}+\varepsilon_{\mathrm{LUMO}})}{2}$	与Mulliken电负性相关，可表示体系被电子还原的难易性
化学硬度	eV	$\eta=\varepsilon_{\mathrm{LUMO}}-\varepsilon_{\mathrm{HOMO}}$	与化学硬度相关，可以根据软硬酸碱理论预测物质的溶解度、化学反应的方向等
NPA电荷[34–36]	|e|	q	按照体系的波函数以自然键轨道算法划分电子云空间分布得到原子自然布居分析(Natural Population Analysis，NPA)电荷，常用于分析体系中各原子的局部电荷
偶极矩	Debye	μ	正、负电荷中心间的距离和电荷中心所带电量的乘积，为矢量(本文取其模)，可用于衡量体系极性大小及极性区域，继而预测在极性不同的溶剂中的溶解性
极化率	a.u.	$\alpha=\frac{\mu}{E}$	分子的平均偶极矩μ与电场强度E的比值，为标量，可衡量电子在体系中流动的难易

表3　物化描述符以及其相关含义

符号	表达式	含义
Log$P$	$\mathrm{Log}P=\frac{\Delta_{\mathrm{hyr}}G-\Delta_{\mathrm{sol}}G}{2.303RT}$	脂水分配系数，指物质在1-辛醇和水中分配系数的比值的对数值，是描述物质亲水性的参数。$\Delta_{\mathrm{hyr}}G$为水中的溶剂化自由能(hydration free energy)，$\Delta_{\mathrm{sol}}G$为1-辛醇的溶剂化自由能(solvation free energy)，R和T分别是理想气体常数$\left(R=8.314\ \mathrm{J\cdot mol^{-1}\cdot K^{-1}}\right)$和温度(K)
σ	$\sigma=\lg K_{\mathrm{X}}-\lg K_{\mathrm{H}}$	用于描述反应速率与平衡常数和反应物取代基类型(吸电子/给电子)之间线性自由能关系的方程。若σ为正值，说明取代基是吸电性的，反之则具有供电性；对于多取代苯甲酸，若取代基之间不存在明显的位阻、氢键等相互作用，Hammett参数可近似认为具有加和性[25]*

* K_{H}为苯甲酸在常温下于水溶液中的酸解离平衡常数，K_{X}为相同条件下在对位p/间位m含有取代基X的单取代苯甲酸的酸解离常数

2.2　QSAR分析方法

2.2.1　皮尔逊(Pearson)相关

皮尔逊相关系数$r\in[-1,1]$，主要研究一对变量之间的线性相关性。基于上述三类分子描述符共21种，研究候选自变量对金属卟啉活化氧气时吉布斯自由能变ΔG的影响，其取值范围与相关程度如表4所示，相关系数的符号只影响相关的正负性，不影响相关程度。其分析主要分为两步：

(1) 相关系数计算(相关程度分析)

$$r=\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2\sum_{i=1}^{n}(y_i-\overline{y})^2}} \tag{1}$$

表4 |*r*|的取值与相关程度

|*r*|的取值	0	0–0.4	0.4–0.7	0.7–1	1
|*r*|的意义	完全不相关	弱相关	中等相关	强相关	完全相关

其中n是样本数，x_i和y_i分别是两个变量的第i项观测值，$\overline{x}$和$\overline{y}$分别是两个变量的平均观测值。

(2) 显著性检验

常用t检验来检验皮尔逊相关系数的显著性，根据t值可得到假设两个变量间完全线性不相关(而非根据公式求得的大于0的|*r*|值)成立的概率：

$$t = \frac{r}{\sqrt{(1-r^2)(n-2)}} \tag{2}$$

其中n为样本个数，r为相关系数，t为检验值。t值对应的显著性水平P与样本数有关，P值越接近0越显著，其检验阈值一般取0.05 (显著相关)或0.01 (极显著相关) [37]，也可根据实际情况适度放宽。

2.2.2 回归分析

回归分析方法是通过建立统计模型来确定因变量与自变量间相关关系的密切程度、结构状态及模型预测的一种有效工具。一元线性回归主要用于研究一个自变量与一个因变量之间的线性关系，而多元线性回归则在此基础上扩展到多个自变量与一个因变量之间的线性关系。其主要目的都是通过建立数学模型来描述这种关系，并进行预测。回归分析效果可用拟合优度R^2来衡量：

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2} \tag{3}$$

其中$\hat{y}_i$是拟合方程下的第i项因变量预测值，其他符号含义与公式(1)中的相同。

对于线性回归，$R^2 \in [0, 1]$，拟合优度越接近1，效果越好。在各种衡量模型拟合效果及其泛化能力的统计学指标中，拟合优度是最常用的。一个模型经训练后如果拟合优度不佳，说明存在欠拟合，可能还需要继续训练，如引入更多的自变量；若训练集上的拟合优度良好，但测试集上的拟合优度明显较差，说明发生了过拟合，泛化能力不佳，此时可采取增加训练集样本或者剔除次要自变量重新拟合的措施。

逐步回归则是多元回归中用于变量选择和模型优化的一种有用方法，通过回归过程中反复对方程中各自变量系数进行t检验(公式较为复杂，此处不列出)，可既不重复也不遗漏地引入关键自变量，以确定最优的回归模型[38]。其不仅可以揭示自变量对因变量的影响大小，还可以根据回归方程进行预测和控制。逐步回归建模应满足的条件有：① 自变量与因变量之间存在相关关系；② 自变量之间无高相关性；③ 变量数小于样本数。

3 实验部分

3.1 研究对象

如图3所示，本文选取四苯基钴卟啉(CoTPP)及其苯环取代衍生物共27种组成训练集，从中获取了21种分子描述符作为候选自变量，其对氧气的结合能ΔG作为因变量；另选取8种CoTPP衍生物组成测试集以检验模型的泛化能力。所选取分子均使用Gaussian 16 Rev A.03程序[39]基于DFT [40–42]完成计算。结构优化、振动分析均使用经D3(BJ)色散校正[43]的B3LYP明星泛函[44]搭配def2-SVP基组[45]；为获得更精确的吉布斯自由能数据，采用隐式溶剂模型中可极化连续模型(Polarizable Continuum Model，PCM) [46]，以水作溶剂(介电常数ε = 78.36)，使用D3色散校正[47]的M06泛函[48]搭配def2-TZVP基组[45]进行高精度单点能计算。

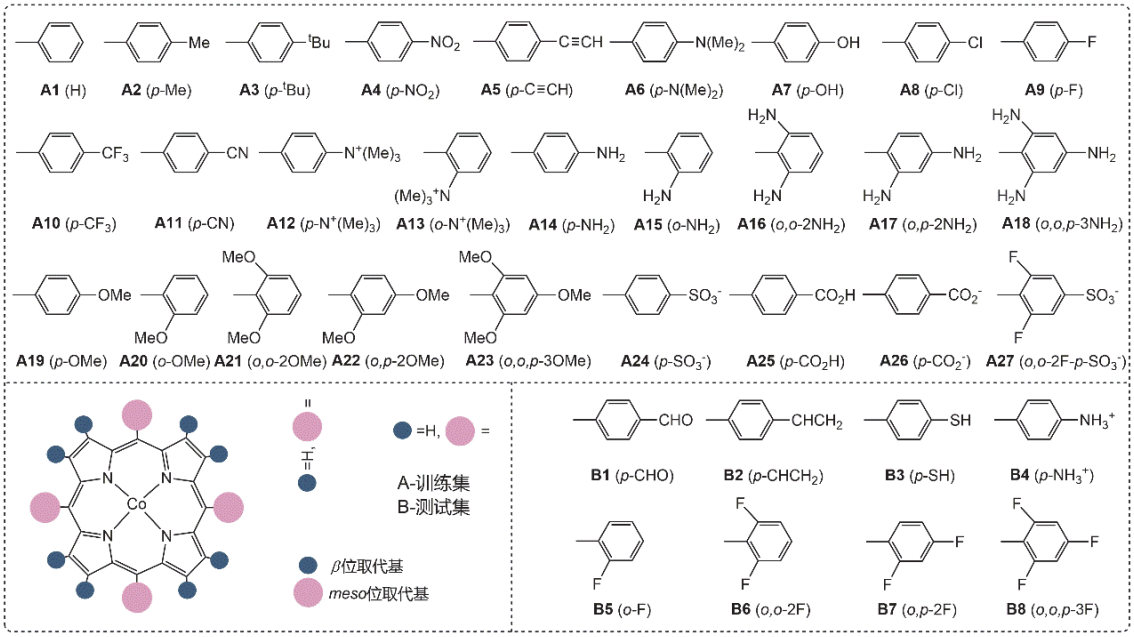

图3　CoTPP及其衍生物组成的训练集和测试集

3.2　实验设备及软件

台式机或者笔记本电脑(本实验基于Windows 11操作系统进行)，Gaussian 16量子化学计算程序及与其搭配的GaussView 6图形界面[49]，ChemOffice 19.0 (含ChemDraw、Chem3D等组件) [50]，SPSS (Statistical Package for the Social Sciences) Statistics 27软件[51]。

3.3　分子描述符的获取

3.3.1　拓扑描述符、物化描述符的获取

1) 在ChemDraw绘制分子的结构。

2) 右击分子选择Molecule→Edit→ Copy As→SMILES。这里的SMILES是指“简化分子线性输入规范”(Simplified Molecular Input Line Entry System)。它是一种文本格式，用于表示分子结构，通过简单的字符序列来描述分子中原子间成键关系及键级。

3) 粘贴到Chem3D里，选择Calculations→Compute Properties；在ChemPropStd中勾选Mol Weight (M_r)，在Molecular Networks中勾选LogP (物化描述符)，在Molecular Topology中勾选Balaban Index (J)、Molecular Topological Index (MTI)、Polar Surface Area (TPSA)、Total Connectivity (TC)、Total Valence Connectivity (TVC)和Wiener Index (W)，操作页面如图4所示；所有参数选择完成后，点击“OK”即可。对于Wiener指数和Balaban指数，选取外围取代基时分别记作Sub$W$和Sub$J$，由于TC、TVC原始值非常小，故对TC乘10^{10}进行放大，对TVC乘10^{20}进行放大。

4) 典型取代基的σ值(Hammett参数)可根据查表得到，具体可以参考补充材料。为简化计算，本文只选取4个*meso*位的其中一组进行计算，且由于取代位点离中心金属较远，仅取σ_p值(对位Hammett参数)。

3.3.2　量化描述符的获取

1) 使用GaussView打开反应物分子的输出文件。

2) 右击选择Tools→MOs打开，有Alpha Mos和Beta Mos两列，分别表示α电子和β电子，HOMO轨道记录两列中已占据电子能级最高轨道所对应的数值，HOMO-1及HOMO-2轨道记录两列中已占据电子能级次高和第三高轨道所对应的数值，即$\varepsilon_{\mathrm{HOMO}(-i)} = max\{\varepsilon_{\alpha-\mathrm{HOMO}(-i)}, \varepsilon_{\beta-\mathrm{HOMO}(-i)}\}$，其中$i \in \{0,1,2\}$；同理，LUMO轨道记录两列中未占据电子能级最低轨道所对应的数值，$\varepsilon_{\mathrm{LUMO}} =$

$\min\{\varepsilon_{\alpha-\mathrm{LUMO}}, \varepsilon_{\beta-\mathrm{LUMO}}\}$。以上数据的单位为原子单位a.u.，在进行数据分析时，通常需要将a.u.转化为eV (1 a.u. = 27.2113838 eV)。此外，还需通过公式$\eta = \varepsilon_{\mathrm{LUMO}} - \varepsilon_{\mathrm{HOMO}}$和$\chi = -(\varepsilon_{\mathrm{HOMO}} + \varepsilon_{\mathrm{LUMO}})/2$分别计算并记录化学硬度($\eta$)和Mulliken电负性($\chi$)。分子轨道能级数据获取操作见图5。

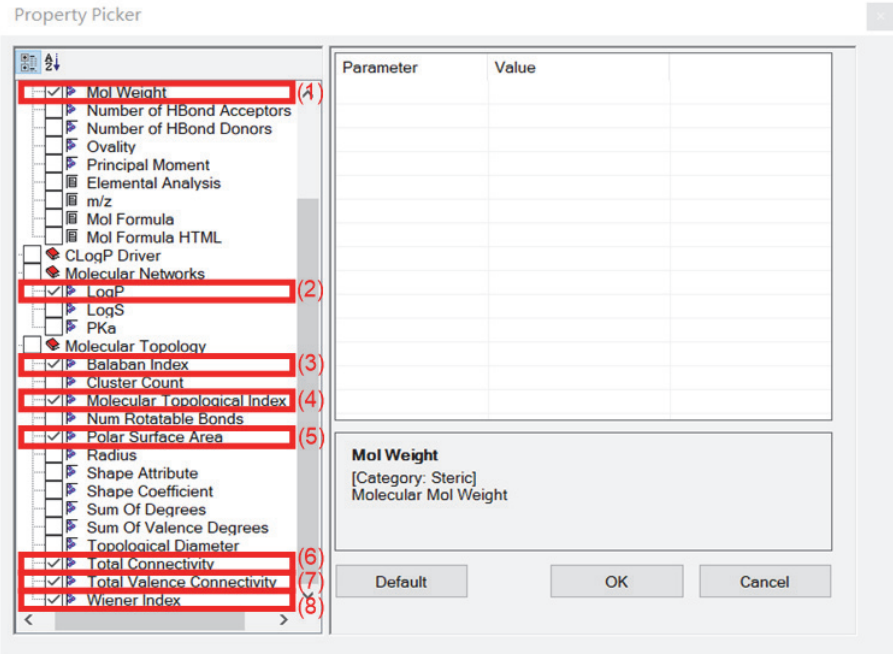

图4　获取分子的结构参数

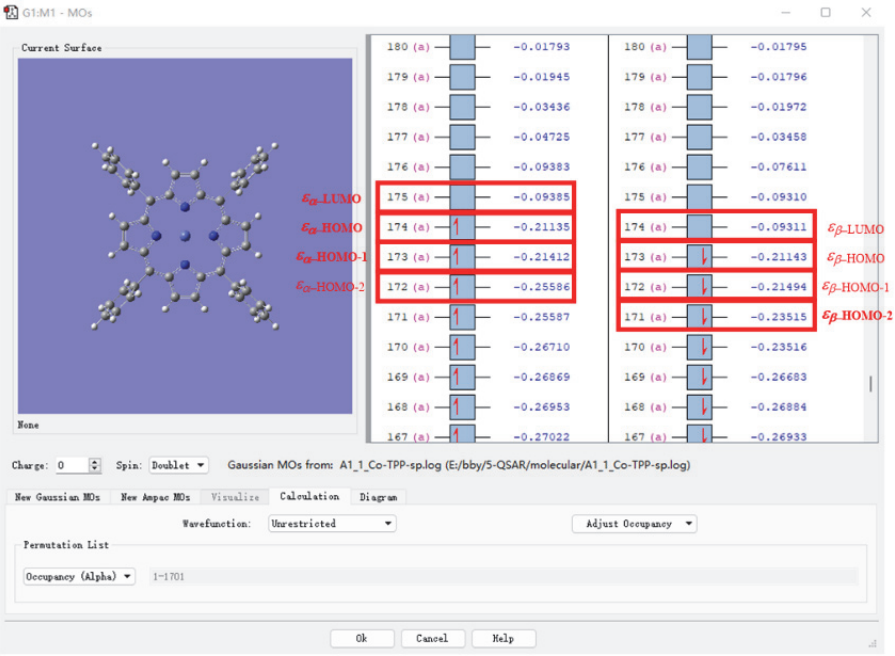

图5　获取HOMO(−*i*)、LUMO轨道能量值

加粗者为实际选取的轨道

3) 右击选择Results→Summary→Overview，分别记录Dipole Moment (μ)和Polarizability (α)，即偶极矩和极化率，操作页面如图6(a)所示。

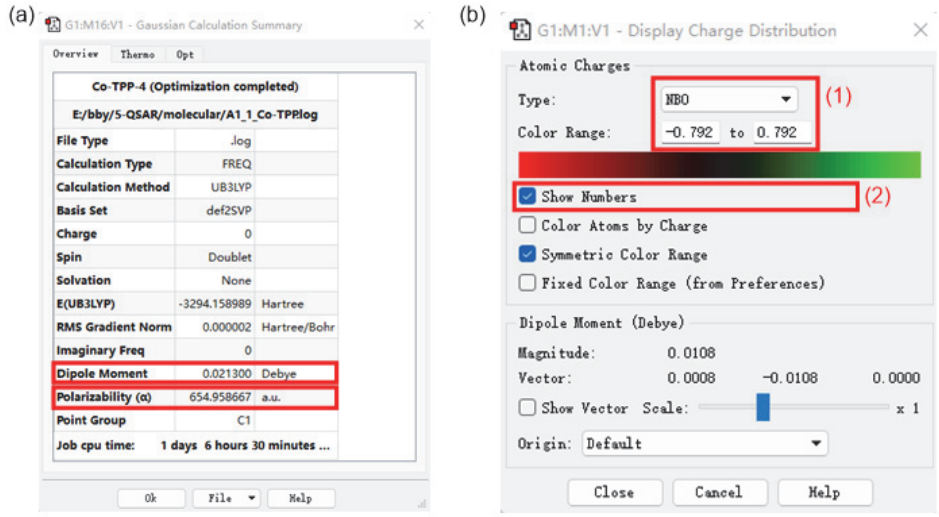

图6 (a) 获取偶极矩和极化率的值；(b) 通过NBO获取NPA电荷的值

4) 右击选择Results→Charge Distribution，在Type中选择NBO并记录，操作页面如图6(b)所示。本文统计了中心金属和卟啉环4个N的NPA电荷之和以及卟啉环4个*meso*位C原子NPA电荷均值，分别记作$q(CoN_4)$和$\bar{q}$(*meso*-C)。

5) 在GaussView中，右击选择Results→Summary→Thermo查看各分子的热力学数据。其中Thermal Correction to Free Energy是分子的吉布斯自由能的校正因子，EE + Thermal Free Energy Correction是分子的吉布斯自由能的大小。本文选用的高精度吉布斯自由能通过高精度单点能与结构优化的吉布斯自由能校正因子相加得到[52]。若催化剂、氧气分子、结合氧气后产物的吉布斯自由能分别记为G_1，G_2，G_3，则所选催化剂分子与氧气的结合能计算为$\Delta G = G_3 - G_1 - G_2$。这里需要对能量单位进行转换，Gaussian计算的能量单位为原子单位Hartree，这里转换为kJ·mol^{-1} (1 Hartree = 2625.5 kJ·mol^{-1})。热力学数据获取操作页面如图7所示。

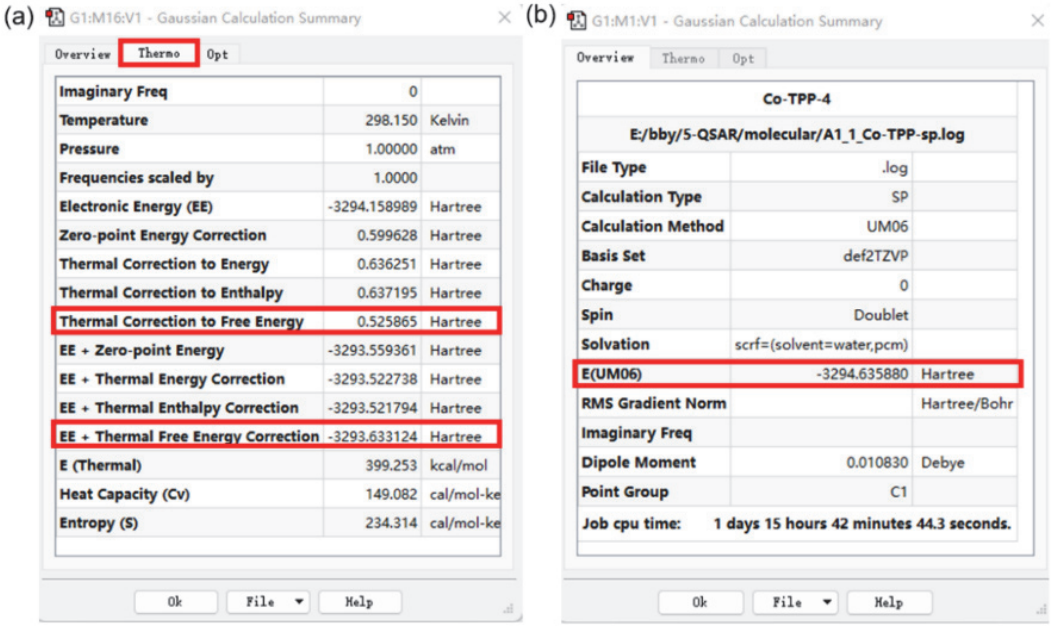

图7 (a) 获取分子的吉布斯自由能大小和校正因子；(b) 获取高精度的单点能

3.3.3 SPSS计算过程

1) 将上述所有分子描述符数据汇总到一个Excel文件中，如图8所示。

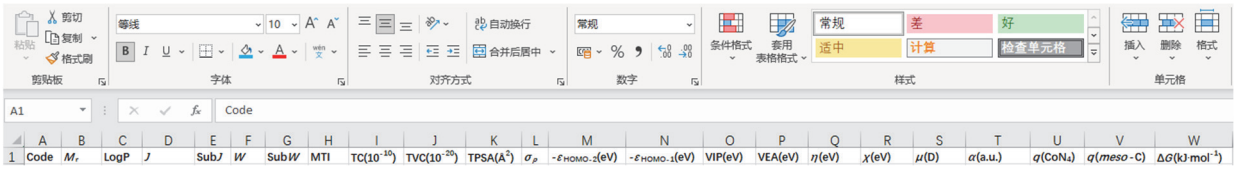

图8　用于导入SPSS的Excel文件示意图

2) 打开SPSS Statistics 27，点击菜单栏“文件”→“导入数据”→导入数据汇总的“Excel”。

3) 相关性分析：分析→相关→双变量，通过中间的箭头将变量导入右侧，同时选择图9中“皮尔逊”相关系数。勾选“仅显示下三角形”表示隐去对角线重复部分。

图9　相关性分析

4) 逐步回归分析：分析→回归→线性，打开对话框，将变量选入，选择方法“步进”，选择图10中的“统计”，默认选项“估算值”；“模型拟合”；另选择“描述”和“共线性诊断”；再打开“选项”，根据实际情况修改引入变量的显著性阈值和剔除变量的显著性阈值，注意“除去”的设值必须大于“进入”的设值。

图10　逐步回归分析

4 结果与讨论

4.1 相关性分析

变量间的皮尔逊相关性可利用Origin绘制热图来直观表示(图11)，具体操作步骤并非本文教学重点，此处不做赘述。相关程度中等以上的变量如表5所示。

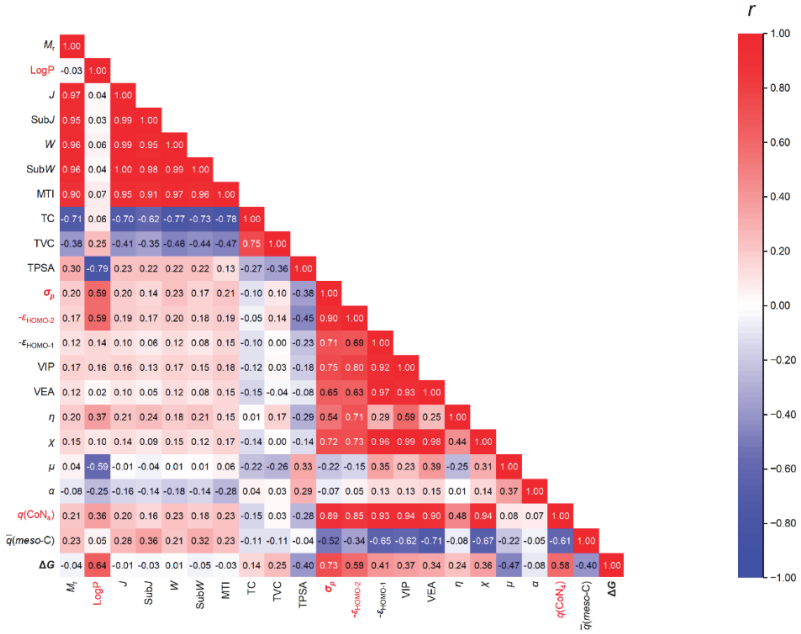

图11 变量之间皮尔逊相关性热图

表5 相关性分析结果

因变量	相关系数|*r*|	相关程度	显著性(*P*)
Log*P*	0.64	中	< 0.001
σ_p	0.73	强	< 0.001
$-\varepsilon_{HOMO-2}$	0.59	中	0.001
μ	0.47	中	0.014
$q(CoN_4)$	0.58	中	0.001
$\bar{q}$(*meso*-C)	0.4	中	0.038

表5展示了与因变量相关性较高的候选自变量及相关系数。对于其中与因变量极显著相关(P < 0.01)的候选自变量，我们认为Log*P*反映了钴卟啉在脂溶性各异的基团调节下自身在低极性溶剂中的稳定性：在低极性溶剂中越稳定，能量越低，使得氧气结合后自由能上升越高。而σ_p、$-\varepsilon_{HOMO-2}$、$q(CoN_4)$则均能反映内核向氧气分子提供电子的难易程度。其中，取代基对位Hammett参数σ_p与取代基的吸电子能力有关，吸电子能力越强，中心金属上的电子密度越低，越难向氧气分子提供电子，使得氧气结合能越高。而$-\varepsilon_{HOMO-2}$与中心金属的关系密切，如图12所示(需打开与输出文件同时生成的波函数文件查看)，轨道能量越低，电子越难向氧气分子的未占据轨道跃迁，故氧气结合能越高；此外，样本中的$q(CoN_4)$均为负值，即CoN_4内核的局部电荷越高(越接近0)，内核向氧气分子提供电子越困难，导致氧气结合能越高。

4.2 逐步回归分析

利用SPSS进行逐步回归分析，本文将引入变量的显著性阈值(P_{in})设置为0.200，剔除变量的显著

性阈值(P_{out})设置为0.300。从无自变量的常数方程$y = \bar{y}$开始，在当前的回归方程的基础上，建立所有可能的仅新引入1个自变量的线性回归方程，当这些回归方程中新自变量得到的最小的P值(回归方程自变量系数的t检验对应的显著性水平)小于引入变量的显著性阈值时，则作为引入变量。引入后，对方程中的所有自变量系数进行t检验，若存在自变量系数的P值超出剔除阈值，则将其移出方程；移除后继续进行t检验，直至无可剔除变量，此刻返回引入变量步骤。反复进行上述引入-剔除操作，直至既无可引入的变量，也无可剔除的变量。

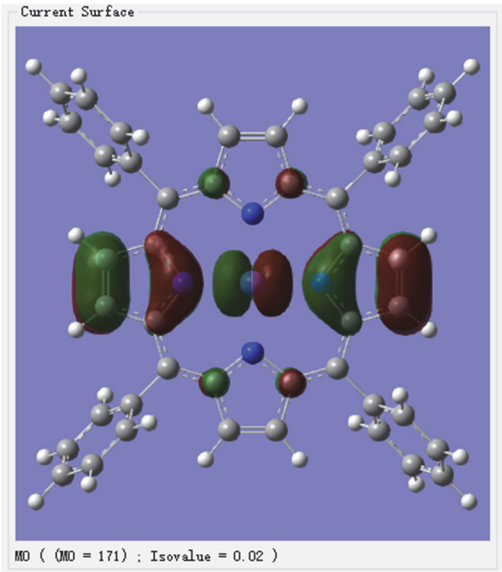

图12　CoTPP的β-HOMO-2形状

逐步回归分析如表6所示依次引入了4个变量，中途无剔除的变量，这里列出最终回归方程：

$$\Delta G/\mathrm{kJ\cdot mol^{-1}} = 111.617 + 5.505\sigma_p - 1.740\ \mu/\mathrm{D} - 17.548\ \eta/\mathrm{eV} + 0.625\ \mathrm{TC}/10^{-10} \quad (4)$$

表6　逐步回归法逐步引入的自变量及相关统计学参数*

步骤	自变量	系数	t值	显著性P	拟合优度R^2
1	C(常数)	55.642	86.283	< 0.001	0.532
	σ_p	4.793	5.331	< 0.001	
2	C	56.752	77.749	< 0.001	0.631
	σ_p	4.328	5.182	< 0.001	
	μ	−1.744	−2.538	0.018	
3	C	110.724	4.236	< 0.001	0.689
	σ_p	5.305	5.796	< 0.001	
	μ	−1.960	−3.001	0.006	
	η	−17.039	−2.065	0.050	
4	C	111.617	4.347	< 0.001	0.713
	σ_p	5.505	6.044	< 0.001	
	μ	−1.740	−2.631	0.015	
	η	−17.548	−2.164	0.042	
	TC	0.625	1.361	0.187	

*试图引入的第五个自变量，其显著性P最小为0.398，不满足引入条件

σ_p、μ、η和TC与因变量氧气结合能呈显著的相关性，该回归模型的拟合优度$R^2=0.713$，即自变量σ_p、μ、η和TC可以解释因变量氧气结合能变化情况的71.3%，说明模型的拟合效果较好，欠拟合现象不显著。值得一提的是部分与因变量相关性也较好的自变量始终未被引入方程，这正好体现了逐步回归法可避免重复引入自变量，如σ_p和Log*P*的相关系数绝对值|*r*|达0.59，引入前者后就不必再重复引入后者(与σ_p相关性更好的自变量自然也不会被引入)；而σ_p和μ的|*r*|仅0.22，令偶极矩得以成为第二个被引入的自变量。新引入的η反映了金属配合物的活性中心作为路易斯酸的硬度，因氧的电负性高，氧气分子作为硬碱需要将其孤对电子配位给中心金属形成σ键并通过d–$p\pi$共轭让金属d电子反馈至氧气分子π^*反键轨道使其活化[53]。钴卟啉的化学硬度越高，ΔG越低，满足软硬酸碱理论“硬亲硬”原则。

4.3 测试集分析

为测试上述所得出的回归方程(4)的泛化能力，本文用前述8种CoTPP衍生物组成的测试集检验模型对其的拟合优度。测试集各样本的标准偏差绝对值均小于3.000，均非离群点。数据集的观测值(量化计算所得氧气结合能)及模型所得预测值如表7所示(样本标准偏差数据参考补充材料)。以观测值ΔG_{obs} (Observations)为x轴，预测值ΔG_{pre} (Predicted values)为y轴进行作图(图13)。如图13所示，利用数学模型验证CoTPP及其衍生物的氧气结合能时，氧气结合能与σ_p、μ、η和TC呈线性关系，根据Consonni等提出的公式[54]，模型对测试集的拟合优度$Q^2=0.547$，泛化能力一般。我们猜测本模型泛

表7 测试集和训练集的观测值和预测值数据(单位：kJ·mol^{-1})

ΔG	ΔG_{obs}	ΔG_{pre}	ΔG	ΔG_{obs}	ΔG_{pre}	ΔG	ΔG_{obs}	ΔG_{pre}	ΔG	ΔG_{obs}	ΔG_{pre}
A1	57.8	58.7	A10	59	57.7	A19	56.7	53.8	B1	53.5	54.2
A2	61.2	56.4	A11	60.7	58.9	A20	53.7	54.3	B2	54.9	56.3
A3	54.6	54.8	A12	58.2	59	A21	47.8	51.8	B3	57	57.1
A4	60.3	61.2	A13	54.5	55.1	A22	52	50.9	B4	59.6	57.5
A5	58.2	57.5	A14	52.9	56.4	A23	50.9	50.8	B5	59.7	55.4
A6	54.7	55.6	A15	53.1	51.4	A24	54.5	57	B6	59	55.9
A7	54.8	55.4	A16	42.4	46.3	A25	55.4	54.8	B7	60.6	55.4
A8	57	58.1	A17	57	49.7	A26	53.3	55.5	B8	60.1	56.3
A9	56.2	57.3	A18	43.4	44.4	A27	60.4	57.8			

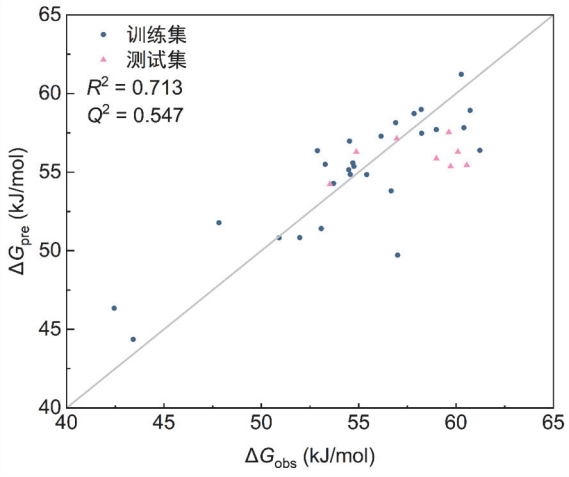

图13 训练集和测试集预测值与实验值拟合效果图

化能力不够好的原因主要为训练集样本数量不足，不同取代基种类所占比例与测试集存在较大偏差，比如亲水性取代基在训练集中所占比例明显高于测试集中的。但是，本论文侧重于QSAR方法的学习，故没有进一步优化模型。

5 结语

本文基于定量构效关系方法，对不同钴卟啉配合物的拓扑参数和量化参数进行计算和统计，通过相关性分析和逐步回归分析建立数学模型，得出其回归方程，拟合优度良好。对额外8种钴卟啉配合物进行测试，但泛化能力一般。本文旨在帮助学生学会用Gaussian获取量化描述符，Chem3D获取拓扑描述符，并利用以上数据凭SPSS软件进行相关性分析及回归分析，得出数据集中影响氧气结合能的主要因素及构效关系，激发学生学习计算化学的兴趣，提高学生在分子建模方面的实际操作能力以及数据分析和处理的能力。学生也可以尝试自行调节逐步回归法的相关阈值进行模型比较，或者采用先引入所有候选变量后依次剔除不满足要求的变量的“后退法”等其他建模方法进行对比。

补充材料：可通过链接 https://www.dxhx.pku.edu.cn 免费下载。

参 考 文 献

[1] 赵苹苹, 胡锴, 蔡苹, 程功臻. 大学化学, **2019,** *34* (5), 33.

[2] Zhang, X.-P.; Chandra, A.; Lee, Y.-M.; Cao, R.; Ray, K.; Nam, W. *Chem. Soc. Rev.* **2021,** *50* (8), 4804.

[3] Zhang, W.; Lai, W.; Cao, R. *Chem. Rev.* **2017,** *117* (4), 3717.

[4] 姜建壮, 吴基培, 谢经雷. 大学化学, **1998,** *13* (4), 6.

[5] 姜建壮, 吴基培, 杜大明, 孙思修. 大学化学, **1999,** *14* (2), 5.

[6] 江国防, 田渊, 郭灿城. 大学化学, **2011,** *26* (2), 1.

[7] 邱晓航, 王竞依, 潘雨辰, 蒙嘉麟, 任红霞. 大学化学, **2015,** *30* (2), 6.

[8] 南志祥, 李珺, 张逢星, 白银娟. 大学化学, **2017,** *32* (4), 46.

[9] 黄宝磊, 孙昊, 李欣烨, 范雨亭, 陶涛. 大学化学, **2021,** *36* (8), 2009060.

[10] Su, B.; Hatay, I.; Trojánek, A.; Samec, Z.; Khoury, T.; Gros, C. P.; Barbe, J.-M.; Daina, A.; Carrupt, P.-A.; Girault, H. H. *J. Am. Chem. Soc.* **2010,** *132* (8), 2655.

[11] Steiger, B.; Anson, F. C. *Inorg. Chem.* **2000,** *39* (20), 4579.

[12] Pegis, M. L.; Wise, C. F.; Martin, D. J.; Mayer, J. M. *Chem. Rev.* **2018,** *118* (5), 2340.

[13] Danishuddin; Khan, A. U. *Drug Discovery Today* **2016,** *21* (8), 1291.

[14] 王鹏. 定量构效关系及研究方法. 哈尔滨: 哈尔滨工业大学出版社, 2011: 22–76.

[15] 焦龙. 拓扑指数的应用: 烃类物质和持久性有机污染物的定量结构性质关系研究. 北京: 中国石化出版社, 2017: 1–100.

[16] Free, S. M.; Wilson, J. W. *J. Med. Chem.* **1964,** *7* (4), 395.

[17] 徐宝峰. 化学教育(中英文), **1996,** *17* (3), 4.

[18] 庞叔薇, 徐晓白. 大学化学, **2002,** *17* (1), 1.

[19] 杨小弟, 崔世海, 毕树平. 大学化学, **2005,** *20* (1), 35.

[20] Yang, Q.; Li, Y.; Yang, J.-D.; Liu, Y.; Zhang, L.; Luo, S.; Cheng, J.-P. *Angew. Chem. Int. Ed.* **2020,** *59* (43), 19282.

[21] Yang, W.; Ma, Z.; Yi, J.; Ahmed, S.; Sun, W.-H. *J. Comput. Chem.* **2019,** *40* (13), 1374.

[22] Buglak, A. A.; Filatov, M. A.; Hussain, M. A.; Sugimoto, M. *J. Photochem. Photobiol. A* **2020,** *403*, 112833.

[23] 戴猷元, 秦炜, 张瑾. 溶剂萃取体系定量结构-性质关系. 北京: 化学工业出版社, 2005: 9–73.

[24] Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*. John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2009; pp. I–XLI.

[25] Ertl, P.; Rohde, B.; Selzer, P. *J. Med. Chem.* **2000,** *43* (20), 3714.

[26] 孟繁宗. 大学化学, **2002,** *17* (6), 45.

[27] 冯长君. 大学化学, **1999,** *14* (4), 47.

[28] 张洪斌. 大学化学, **1998,** *13* (1), 19.

[29] König, D. *Theory of Finite and Infinite Graphs*. Birkhäuser Boston: Boston, MA, USA, 1990; pp. 45–421.

[30] Wiener, H. *J. Am. Chem. Soc.* **1947,** *69* (1), 17.

[31] Balaban, A. T. *J. Chem. Inf. Comput. Sci.* **1992,** *32* (1), 23.

[32] Hall, L. H.; Kier, L. B. *J. Mol. Graphics Modell.* **2001,** *20* (1), 4.

[33] Schultz, H. P. *J. Chem. Inf. Comput. Sci.* **1989,** *29* (3), 227.

[34] Reed, A. E.; Weinstock, R. B.; Weinhold, F. *J. Chem. Phys.* **1985,** *83* (2), 735.

[35] 卢天, 陈飞武. 物理化学学报, **2012,** *28* (1), 1.

[36] Foster, J. P.; Weinhold, F. *J. Am. Chem. Soc.* **1980,** *102* (24), 7211.

[37] 聂长明, 廖力夫. 计算化学. 北京: 北京理工大学出版社, 2010: 8–25.

[38] 何晓群, 刘文卿. 应用回归分析. 第3版. 北京: 中国人民大学出版社, 2011: 16–170.

[39] Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; *et al*. *Gaussian 16*, *Rev. A.03*; Gaussian Inc.: Wallingford, CT, USA, 2016.

[40] Hohenberg, P.; Kohn, W. *Phys. Rev.* **1964,** *136* (3B), B864.

[41] Kohn, W.; Sham, L. J. *Phys. Rev.* **1965,** *140* (4A), A1133.

[42] Calais, J.-L. *Int. J. Quantum Chem.* **1993,** *47* (1), 101.

[43] Grimme, S.; Ehrlich, S.; Goerigk, L. *J. Comput. Chem.* **2011,** *32* (7), 1456.

[44] Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. *J. Chem. Phys.* **1994,** *98* (45), 11623.

[45] Weigend, F.; Ahlrichs, R. *Phys. Chem. Chem. Phys.* **2005,** *7* (18), 3297.

[46] Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. *J. Phys. Chem. B* **2009,** *113* (18), 6378.

[47] Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. *J. Chem. Phys.* **2010,** *132* (15), 154104.

[48] Zhao, Y.; Truhlar, D. G. *Theor. Chem. Acc.* **2008,** *120* (1), 215.

[49] Dennington, R. K. T.; Millam, J. *GaussView Version 6*; Semichem Inc.: Shawnee Mission, KS, USA, 2016.

[50] Irwin, J. J. *J. Chem. Inf. Model.* **2005,** *45* (5), 1468.

[51] IBM. *SPSS Statistics 27*; IBM Corp.: Armonk, NY, USA, 2020.

[52] 王亚妮, 张学鹏. 大学化学, **2023,** *38* (2), 197.

[53] Moltved, K. A.; Kepp, K. P. *ChemPhysChem* **2020,** *21* (19), 2173.

[54] Consonni, V.; Ballabio, D.; Todeschini, R. *J. Chem. Inf. Model.* **2009,** *49* (7), 1669.