文章编号:2095-0411(2017)01-0086-07

基于遗传算法对有机物热导率的预测研究

时静洁,袁雄军,邵 辉,王凯全,陈海群
(常州大学 环境与安全工程学院,江苏 常州 213164)

摘要:根据定量构效关系(Quantitative Structure-Property Relationship,QSPR)原理,研究热导率与其分子结构间的内在定量关系。以 178 种有机化合物作为样本集,随机选择 142 种作为训练集,36 种作为测试集,采用遗传算法(Genetic Algorithm,GA)进行变量选择,得到 5 个特征描述符作为模型的输入变量,结合多元线性回归(Multiple Linear Regression,MLR)方法建立了遗传-多元线性回归(GA-MLR)预测模型。研究结果表明:GA-MLR 模型的训练集和测试集的复相关系数分别为0.808 0和0.742 2,其均方根误差分别为0.109 8和0.129 3,预测效果令人满意。随后采用残差分析图对样本集进行了残差分析,进一步验证模型在建立过程中未产生系统误差。采用"Y-随机性检验"方法对模型进行了研究,发现预测模型不存在"偶然相关"现象,具备较强的稳定性。该研究提供了一种有效预测有机化合物热导率的方法。

关键词:热导率;遗传算法;多元线性回归;预测;定量构效关系
中图分类号:X 937 **文献标志码**:A **doi**:10.3969/j.issn.2095-0411.2017.01.015

Prediction of the Thermal Conductivity of Organic Compounds Based on the Genetic Algorithm

SHI Jingjie, YUAN Xiongjun, SHAO Hui, WANG Kaiquan, CHEN Haiqun
(School of Environmental & Safety Engineering, Changzhou University, Changzhou 213164, China)

Abstract: The quantitative relationship between the thermal conductivity and the molecular structure of the organic compound was investigated based on the QSPR principle. The datasets of 178 kinds of organic compounds were randomly divided into the train set (142) and the test one (36). GeneticAlgorithm (GA) was well adapted to the variable selection. As a result, five descriptors were screened out from a large pool of calculated descriptors as input parameters for the model. Coupled with these descriptors, Multiple Linear Regression (MLR) method was employed to build the GA-MLR model. In the model, the square correlation coefficient (R^2) of the train set and the test one were 0.808 0 and 0.742 2, and the Root Mean Square Error(RMSE) were 0.109 8 and 0.129 3, respectively. The above results showed that the built modelis robust are very satisfying. Then, the sample was analyzed with the residual analysis as to further validate no systematic error in the process of the model. In addition, Y-randomization was applied to determine that there is no chance correlation in the model. It was seen that the established model had strong stability. This research provides a new and effective method for predicting the thermal conductivity of organic compounds.

收稿日期:2016-05-10。
基金项目:常州大学科研启动基金(ZMF15020112);常州市科技支撑计划项目(社会发展)(CE20155025);建筑消防工程技术公安部重点实验室开放课题(KFKT2014MS02)。
作者简介:时静洁(1987—),女,江苏常州人,博士,讲师。通讯联系人:陈海群(1970—),E-mail:shijingjie@cczu.edu.cn

Key words: thermal conductivity; genetic algorithm; multiplelinearregression; prediction; quantitativestructure-property relationship

热导率又称导热系数[1]，是物质的一种物理性质，反映物质的热传导能力，对于任何热过程的工程设计来说都是非常重要的[2]。热导率是化工生产过程、石油工业和能源工程等有关传热设计中必不可少的重要基础数据之一[3-5]。因此，从 18 世纪中叶人们就对其测量方法进行了大量的探索[6-7]。热导率主要取决于物质的分子结构和所处温度，其数值主要依赖实验测定，如热通量法[8]、闪蒸法[9]、瞬态板热源法[10]、瞬态热线法[11]等。但实验测定非常耗时，且费用昂贵，更重要的是所测得的热导率值存在一定误差，其主要原因是在测试过程中无法控制对流和辐射所产生的热量流失，确定准确的热导率值是非常困难的。因此，探索热导率的理论预测方法具有重要的理论意义和实用价值。关于热导率的理论估算成为近代物理和物理化学中一个非常活跃的课题，由此建立了相当多的理论估算模型。目前，国内外学者采用的方法主要为基团贡献法和定量构效关系(Quantitative Structure-Property Relationship，QSPR)法，但由于基团贡献法存在无法区分同分异构体、无法考虑基团间的交互作用对性质的影响等局限性。因此，QSPR 法成为目前预测热导率的主要理论方法[12-13]。2001 年，Kauffman 等人[14]选用 213 种有机化合物溶剂作为研究样本，采用目标特征选择法和主观特征选择法筛选出了 9 个分子描述符作为模型的输入参数，建立了 MLR 模型和人工神经网络模型。2007 年，Toropov 等人[15]对 58 种纳米材料的热导率进行了 QSPR 研究。2012 年，蒋海燕等人[16]建立基于反向传播(Back Propagation，BP)神经网络的热导率预测模型以获取无岩心深度段的岩石热导率。2013 年，Gharagheizi 等人[17]将 1 635 种有机物在不同温度下的 19 000 个数据样本点分为训练集、测试集和验证集，提出了预测热导率的 QSPR 模型。

GA 通过全面模拟自然选择和遗传机制，是一种具有“生成＋检测”迭代过程的搜索算法。目前，GA 在很多方面得到了广泛应用[18]。在最优化技术问题上，GA 应用于寻找复杂搜索空间中的全局最优解[19]；在定量构效关系模型研究中，GA 被作为一种灵活高效常用的变量筛选方法[20]；此外，GA 同样应用于曲线拟合[21]等相关问题。在 GA 的应用过程中，人们往往结合问题的特征和领域知识对 GA 进行各种改变，形成了各种各样的具体遗传算法，使得 GA 具备求解不同类型优化问题的能力和强大的全局搜索能力。因此，本文选用遗传算法变量筛选方法，与多元线性回归方法相结合，建立遗传-多元线性回归模型，预测有机化合物的热导率值，并探讨影响有机化合物热导率的结构因素。

1　材料和方法

1.1　试验样本

QSPR 研究是从大量已知样本数据中提取相关的结构-性质关系的信息，发现规律，运用数学统计方法建立模型，用以预测化合物的性质[22]。因此，必须有足够量和可靠的样本数据作为基础。由于液、固体的热导率一般与压力关系不大，但受温度的影响很大，因此本文选取 20℃的有机物液体热导率值，并将所有的热导率取自然对数(用 lnλ 表示)作为模型的因变量。实验条件的不同必然导致实验结果的不一致，样本数据的准确性会直接影响所建立模型的预测效果，因此选择一个可靠的数据库是十分重要的。为了消除不同数据库中数据的差异可能给实验预测结果造成的影响，故本文所选用的 178 种有机物液体的热导率值均来源于权威数据库《有机化合物实验物性数据手册》[23]。随后随机选择 142 种化合物作为训练集用于建立模型，剩下的 36 种作为测试集，用于评估模型的预测能力。

1.2　分子描述符的计算

描述符是将化学物质分子结构的特征量化。这些描述符量化给定化学结构，是由化学组成，拓扑，几何，波函数，势表面或者是由以上描述符的结合而获得的。这些特定的描述符的值是由实验者或通过一些软件计算而获得的。每一个描述符必须和所给定的结构相对应。目前已开发的分子描述符种类繁多、数目庞大，对于如此多的分子描述符，目前已有多种化学软件可进行计算。本文采用了应用非常广泛的分子描述符计

算软件——Dragon 2.1 软件进行分子描述符的计算[24-26]。Dragon 软件[27-28]是一款高效计算分子描述符的模拟软件,主要用于评估定量构效关系、相似性分析和高通量分子数据库的筛选等。Dragon 2.1 软件可以计算如组成描述符、拓扑描述符、RDF 描述符、官能团描述符等 18 类共 1 481 种描述符,有效地表征了有机物的分子结构。

为解决由于分子描述符过多而造成数据分析和模型建立的困难,有必要首先对众多描述符进行预筛选。利用 Dragon 2.1 软件的筛选功能进行初步筛选,筛除常数或者近似常数及描述符间相关系数达到 0.95 以上的描述符。经过预筛选,分子描述符减少至 592 个,但仍无法满足 QSPR 建模的需要。GA 作为智能优化算法,已在组合优化等问题中被证明是高效的变量选择方法,分子描述符的选择问题正属于组合优化问题。因此,本文将 GA 用于分子描述符的筛选,以期取得良好的效果。

1.3 遗传算法(GA)

遗传算法是 20 世纪 70 年代初期由美国密歇根大学的 Holland 教授发展起来的,是一种模拟生物进化过程的计算模型[29]。遗传算法的基本处理流程主要有以下几个步骤:对参数集进行编码、初始化群体、适应度值评价、遗传操作、终止循环条件。本文所选用的遗传算法(GA)是由 Rogers 和 Hopfinger[30]提出的以拟合缺失分数(Lack-Of-Fit,LOF)作为其适应度函数的遗传函数算法(Genetic Function Approximation,GFA),应用于变量选择。将 GA 运用到函数逼近问题上[30],给出大量影响函数值的潜在因素,找到最好的关联函数值的变量子集。GA 在进化搜索中基本不用外部信息,仅用目标函数即适应度函数作为依据。由于适应度值是群体中个体生存机会选择的唯一确定性指标,所以适应度函数的形式直接决定着群体的进化行为。在进化过程中,很多统计指标可以用来评价模型是否合适。本文选用的 LOF 适应度函数定义如下

$$f_{\mathrm{LOF}}=\frac{\varepsilon_{\mathrm{SSE}}}{\left[1-\lambda\left(\frac{c+dp}{M}\right)\right]^2} \tag{1}$$

式中:$\varepsilon_{\mathrm{SSE}}$ 为误差平方和;c 为基础方程的数目;d 为平滑系数($d=\{\alpha(M-c_{\max})/c_{\max}\}$,$c_{\max}$为模型中所取最大变量数,$\alpha$ 为平滑参数);p 为模型中所用变量的数目;M 为训练集样本数目;λ 为安全系数,一般设为 0.99,这个系数可以确保表达式的分母不为 0。

由 f_{LOF} 的表达式可知,f_{LOF} 与 $\varepsilon_{\mathrm{SSE}}$ 不同,当增加回归模型变量的同时,f_{LOF} 值并不会一直减小。增加新的变量一方面会减小 $\varepsilon_{\mathrm{SSE}}$ 的值,同时 c 和 p 的值反而会增加,进而 f_{LOF} 的值就会增加。因此,增加变量会减小 $\varepsilon_{\mathrm{SSE}}$ 值,却增加了 f_{LOF} 的值。因此,f_{LOF} 与传统的 $\varepsilon_{\mathrm{SSE}}$ 相比,可以有效防止通过简单增加变量而产生的"过拟合",可以有效估测到最适宜的变量个数,且可以使用户控制拟合的平滑度等。

1.4 模型的验证

模型验证是 QSPR 研究中非常重要的部分。仅仅对模型的"拟合能力、稳定性和预测能力"中的一种或两种进行评价,缺乏对模型全面有效的验证。本文根据"OECD Principles"[31],从模型拟合能力、稳定性和预测能力 3 个方面,对所建 QSPR 模型进行全面的评价和验证。本文主要以复相关系数 R^2 和均方根误差 RMSE 作为模型拟合能力评价指标。复相关系数 R^2 是测量一个变量与其他多个变量之间线性相关程度的指标,均方根误差 RMSE 表示随机误差的分散程度。R^2 越大,RMSE 越小,说明所建模型拟合能力越强,但是并不能保证模型具有更好的预测精度。交互检验的 Q^2 是目前使用较为广泛的一种内部检验方法,其中"留一法"(Q^2_{loo})是最常用的交叉验证方法。其定义如(2)式所示

$$Q^2_{\mathrm{loo}}=1-\frac{\sum_{i=1}^{\text{训练集}}(y_i-y_i)^2}{\sum_{i=1}^{\text{训练集}}(y_i-\bar{y}_{tr})^2} \tag{2}$$

式中:y_i 和 y_i 分别表示训练集的实验值和预测值;$\bar{y}_{tr}$ 表示训练集样本实验值的平均值。

Q^2_{loo} 的结果可以说明 QSPR 模型的稳健性和内部预测性能,但并不能保证模型的真实预测能力也较强。

Tropsha 等人[32-33]指出，对模型预测能力的评价必须通过对未参与训练的物质进行预测。这里采用测试集样本预测值与实验值之间的交互验证系数 Q^2_{ext} 来衡量模型的外部预测能力。

$$Q^2_{ext}=1-\frac{\sum_{i=1}^{测试集}(y_i-y_i)^2}{\sum_{i=1}^{测试集}(y_i-\bar{y}_{tr})^2} \tag{3}$$

式中：y_i 和 y_i 分别表示测试集的实验值和预测值；$\bar{y}_{tr}$ 表示训练集样本实验值的平均值。

2　结果与讨论

2.1　GA-MLR 热导率预测模型

本文运用 GA 对 592 个分子描述符进行进一步筛选，其运行过程均在 Materials Studio 6.0 软件中实现。相应的参数均采用软件默认设置，初始方程式长度为 5，最大方程式长度为 10，种群数为 50，最大代数为 500，变异概率为 0.1，比例控制 LOF 的平滑参数 α 为 0.5。

针对 142 个训练集样本，以 LOF 函数作为 GA 中的适应度函数，对分子描述符进行筛选，确定了 5 个特征描述符，其类型与定义列于表 1。随后，将 5 个特征描述符作为模型的输入变量，热导率自然对数值作为输出变量，运用 SPSS17.0 统计软件，在 95%的置信区间内，建立了 MLR 模型，如(4)式所示

$$\ln\lambda=-1.727-0.199\times x_{nF}-0.057\times x_{SEigp}+0.094\times x_{nHDon}-0.075\times x_{GGI1}-0.526\times x_{E1v} \tag{4}$$

$$n=142,\quad R^2=0.808,\quad D_s=0.112\,14,\quad F=114.134,\quad p<0.001$$

式中：lnλ 为热导率自然对数值；n 为训练集样本数；R^2 为决定系数；D_s 为模型标准误差；F 为 F 检验值。$F_{实际}=114.134>F_{理论}(5,136,0.05)=2.29$，实际 F 值大于理论 F 值，则认为回归关系假设的因果关系是显著的。模型的显著性概率 p 远小于 0.05，认为该回归方程及所筛选变量的影响均是显著的。

表 1　GA-MLR 模型中的特征描述符及其统计学参数

描述符	类型	定义	回归系数	标准误差	标准系数	t-值
常数项	—	—	−1.727	0.047	—	−37.010
nF	组成描述符	氟原子的数目	−0.199	0.021	−0.782	−9.442
SEigp	拓扑描述符	根据极化率加权，距离矩阵的特征值之和	−0.057	0.008	−0.515	−6.826
nHDon	官能团描述符	与 N 原子和 O 原子相连的 H 原子数	0.094	0.020	0.202	4.645
GGI1	Galvez 拓扑电性指数	第 1 拓扑电性指数	−0.075	0.012	−0.311	−6.451
E1v	WHIM 描述符	1st 成分可达性定向 WHIM 指数/根据原子范德华体积加权	−0.526	0.079	−0.280	−6.693

在这 5 个描述符中，涉及 5 种不同的描述符，分别是组成描述符、拓扑描述符、官能团描述符、Galvez 拓扑电性指数和 WHIM 描述符。它们从不同方面表征了分子结构特征。nF 属于组成描述符，表示分子中氟原子的数目。SEigp 属于拓扑描述符，由分子图论获得，主要表征分子的极化率。nHDon 属于官能团描述符，表示与 N 原子和 O 原子相连的 H 原子数目，氢键的形成可能与该官能团密切相关。GGI1 属于 Galvez 拓扑电性指数，能够表征原子对间以及整个分子中的电荷转换。E1v 主要表征原子范德华体积的大小，属于 WHIM 描述符。通过分子几何坐标获得加权协方差矩阵，然后由加权协方差矩阵计算得到主成分坐标，最后计算主成分坐标即可获得 WHIM 描述符。为消除自变量量纲和数量级的不同所造成的影响，采用标准系数来衡量描述符的贡献度大小。方程中，若标准系数为正，表明该描述符与热导率值成正相关，反之，则为负相关。标准系数的绝对值与相关程度成正比。比较表 1 中各描述符的标准系数，发现 nF 前的标准系数的绝对值最大，其次为 SEigp，因此，可以说明分子中氟原子的数目及分子的极化率大小是影响有机化合物热导率的主要因素。

为评价模型的拟合能力和外部测试能力,分别针对训练集中的 142 种化合物和测试集中的 36 种化合物进行分析验证,模型的主要性能参数见表 2。GA-MLR 模型所得热导率预测值和实验值见附录 1,预测值和实验值的比较见图 1。

由图 1 可见,整体基本位于对角线两边附近,但无论是训练集还是测试集化合物,均有部分物质偏离对角线较远。这些物质的预测值出现了比较大的偏差,被称为"异常值"。产生"异常值"主要有 2 个可能原因造成:①是热导率的实验数据本身存在问题;②是这些分子的某些结构特征并未被所筛选出的分子描述符很好表征。

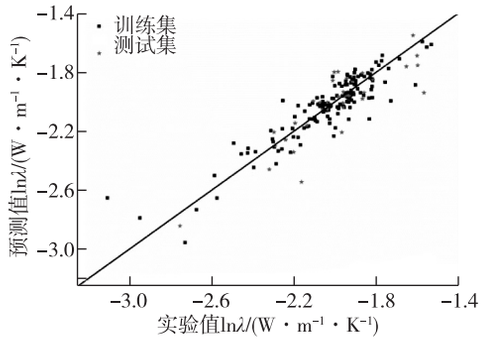

图 1 GA-MLR 模型所得热导率预测值与实验值的比较

表 2 GA-MLR 模型的主要性能参数

性能参数	R^2	Q^2_{loo}	Q^2_{ext}	RMSE
训练集	0.808 0	0.807 6	—	0.109 8
测试集	0.742 2	—	0.780 8	0.129 3

从图 1 中可以看出,GA-MLR 热导率预测模型对训练集中 142 个样本和测试集中 36 个样本的预测值均与实验值有较好的一致性,预测精度令人满意。比较表 2 中各热导率预测模型的主要性能参数发现,模型中训练集和测试集的 R^2 均比较高,预测误差较低,而且比较接近。一般认为,若 Q^2_{loo} 和 Q^2_{ext} 均大于 0.6[34],则说明所建立的模型不但比较稳定,而且具备较强的预测能力和泛化推广性能。

随后,对 GA-MLR 模型的样本集进行了残差分析,讨论在建立过程中是否存在系统误差,其残差图如图 2 所示。

由图 2 可以看出,模型的计算残差均随机分布于基准线的两侧,不存在明显的规律性。由此可以推断,预测模型在建立过程中未产生系统误差。

为检验所建模型是否存在机会相关,用"Y-随机性检验"方法对模型运行 50 次,所得最大 R^2 为0.509 0,其结果不如原始模型。由此可见,本文所建立的热导率预测模型不存在"偶然相关"现象,具备较强的稳定性。

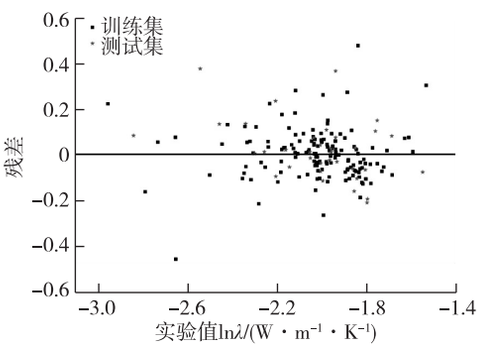

图 2 GA-MLR 模型的实验值与残差关系图

2.2 模型的比较

本文将所建 GA-MLR 模型与文献[17]中的热导率预测模型进行比较。Gharagheizi 等人收集了来自 DIPPR801 数据库的 1 635 种有机化合物在不同温度和不同压力下的 19 000 个数据点,其中 15 417 个数据点作为训练集建立模型,1 926 个数据点作为测试集,文献[17]采用顺序查找算法作为分子描述符筛选方法,建立了包含有 20 个变量的多元线性模型。所建线性模型预测结果比较理想,其平均相对误差为 7.4%,均方根误差为 0.01。文献选取的样本集范围比较广,化合物也比较齐全,但所选化合物的热导率是在不同温度下的数据。而热导率受温度影响较大,因此选取同一温度下的化合物热导率数据将会提高所建模型质量。本文正是基于这方面考虑,选取的样本数据统一来自 20℃时有机物液体的热导率。文献[17]采用的顺序查找算法作为分子描述符筛选方法,其方法简单直观,但不能遍历所有变量组合,不能保证得到的是全局最优解。而本章选用的筛选方法 GA 是一种高效的全局优化概率搜索方法,目前应用十分广泛。综合以上比较,作者认为本文所建立的 GA-MLR 模型在样本集的选取、筛选方法这两方面均优于文献[17]。

3 结　论

搜集了权威数据库《有机化合物实验物性数据手册》中的 178 种有机化合物的热导率，采用 QSPR 方法的基本思想，研究了有机物分子结构与热导率之间的关系。采用遗传算法作为变量筛选方法，获得 5 个特征描述符作为模型的输入参数，结合多元线性回归方法，建立了 GA-MLR 模型。本文所建立的 GA-MLR 热导率预测模型，其性能优越性明显，训练集和测试集的 R^2 均达到 0.7 以上。为了进一步检验模型的系统误差和机会相关性，分别采用残差分析法和“Y-随机性检验”方法进行研究，结果表明，预测模型在建立过程中未产生系统误差，且不存在“偶然相关”现象，进一步表明模型的可靠性和稳定性。

参考文献：

[1]张苗，刘幸娜，陈阵，等. 辐射强度对建筑典型外保温材料燃烧性能影响的试验研究[J]. 中国安全科学学报，2013，23(12)：42-47.

[2]漆政昆，张和平，黄冬梅，等. 消防服用织物材料热湿舒适性综合评价[J]. 中国安全科学学报，2012，22(4)：133-138.

[3]娄江峰，张华，王瑞祥. 纳米冷冻机油热导率的实验研究[J]. 化工进展，2015，34(2)：495-499.

[4]杨晨，高思云. 基于热传导反问题的各向异性材料热物性预测方法[J]. 化工学报，2007，58(6)：1378-1384.

[5]袁超，段斌，李岚，等. 热界面材料热导率和接触热阻的测试[J]. 化工学报，2015，66(S1)：349-353.

[6]彭国文，肖方竹，聂长明，等. 液相链烷烃热导率与其结构定量关系[J]. 化工学报，2011，62(3)：604-610.

[7]WEI X H，WANG L Q. Synthesis and thermal conductivity of microfluidic copper nanofluids[J]. Particuology，2010，8(3)：262-271.

[8]RIDES M，MORIKAWA J，HALLDAHL L，et al. Intercomparison of thermal conductivity and thermal diffusivity methods for plastics[J]. Polymer Testing，2009，28(5)：480-489.

[9]COQUARD R，PANEL B. Adaptation of the FLASH method to the measurement of the thermal conductivity of liquids or pasty materials[J]. International Journal of Thermal Sciences，2009，48(4)：747-760.

[10]HUANGL H，LIUL S. Simultaneous determination of thermal conductivity and thermal diffusivity of food and agricultural materials using a transient plane-source method[J]. Journal of Food Engineering，2009，95(1)：179-195.

[11]NAGASAKA Y，NAGASHIMA A. Simultaneous measurement of the thermal conductivity and the thermal diffusivity of liquids by the transient hot-wire method[J]. Review of Science Instruments，1981，52(2)：229-232.

[12]许路加，胡明，杨海波，等.基于微结构参数建模的多孔硅绝热层热导率研究[J].物理学报，2010，59(12)：8794-8800.

[13]时静洁，陈利平，陈网桦，等. 基于启发式方法和支持向量机方法预测有机物的热导率[J]. 物理化学学报，2012，28(12)：2790-2796.

[14]KAUFFMAN G，JURSP C. Prediction of surface tension，viscosity，and thermal conductivity for common organic solvents using quantitative structure-property relationships[J]. Journal of Chemical Information and Computer Science，2001，41(2)：408-418.

[15]TOROPOVA A，LESZCZYNSKA D，LESZCZYNSKI J. Predicting thermal conductivity of nanomaterials by correlation weighting technological attributes codes[J]. Materials Letters，2007，61(26)：4777-4780.

[16]蒋海燕，施小斌，杨小秋，等. 基于 BP 神经网络的测井资料预测岩石热导率[J]. 测井技术，2012，36(3)：304-307.

[17]GHARAGHEIZI F，KASHKOULIP I，SATTARI M，et al. Development of a quantitative structure-liquid thermal conductivity relationship for pure chemical compounds[J]. Fluid Phase Equilibria，2013，355：52-80.

[18]时静洁，陈利平，石宁，等. 基于遗传算法的支持向量机预测有机物自燃点的研究[J]. 中国安全科学学报，2011，21(7)：125-129.

[19]XIAO H，LEE L H. Simulation optimization using genetic algorithms with optimal computing budget allocation[J]. Simulation，2014，90(10)：1146-1157.

[20]ZHANGY X. An improved QSPR method based on support vector machine applying rational sample data selection and genetic algorithm-controlled training parameters optimization[J]. Chemometrics and Intelligent Laboratory Systems，2014，134：34-46.

[21]XIA X H，SUN H W. Curve fitting based on genetic algorithms for quantitative resolution in overlapped fluorescence spec-

tra[J]. Spectroscopy and Spectral Analysis, 2012, 32(8): 2157-2161.

[22]冯琳琳,张兆志,王新颖, 等. 取代芳烃对发光菌急性毒性的 QSAR 研究[J]. 常州大学学报(自然科学版), 2012, 24(4): 8-12.

[23]马沛生. 有机化合物实验物性数据手册——含碳、氢、氧、卤部分[M]. 北京: 化学工业出版社, 2006.

[24]HELGUERA A M, COMBES R D, GONZALEZ M P, et al. Applications of 2D descriptors in drug design: a DRAGON tale[J]. Current Topics in Medicinal Chemistry, 2010, 8(18): 1628-1655.

[25]CUBILLAN N, MARRERO P Y, ARIZA R H, et al. Novel global and local 3D atom-based linear descriptors of the Minkowski distance matrix: theory, diversity-variability analysis and QSPR applications[J]. Journal of Mathematical Chemistry, 2015, 53(9): 2028-2064.

[26]JOUYBAN A, SHAYANFAR A,GHAFOURIAN T, et al. Solubility prediction of pharmaceuticals in dioxane plus water mixtures at various temperatures: Effects of different descriptors and feature selection methods[J]. Journalof Molecular-Liouids, 2014, 195: 125-131.

[27]TEBBY C, MOMBELLI E, PANDARD P, et al. Exploring an ecotoxicity database with the OECD (Q)SAR Toolbox and DRAGON descriptors in order to prioritise testing on algae, daphnids, and fish[J]. Science of the Total Environment, 2011, 409(18): 3334-3343.

[28]BOROTA A, MRACEC M, GRUIA A, et al. A QSAR study using MTD method and dragon descriptors for a series of selective ligands of alpha C-2 adrenoceptor[J]. European Journal of Medicinal Chemistry, 2011, 46(3): 877-884.

[29]MELANIE M. An Introduction to genetic algorithms[M]. London: MIT Press, 1999.

[30]ROGERS D. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships[J]. Journal of Chemical Information and Computer Sciences, 1994, 34(4): 854-866.

[31]STAVROU M, BARDOW A, GROSS J. Estimation of the binary interaction parameter K_{ij} of the PC-SAFT Equation of State based on pure component parameters using a QSPR method[J]. Fluid Phase Equilibria, 2016, 416: 138-149.

[32]TROPSHA A, GRAMATICA P, GOMBAR V K. The importance of being earnest: validation is the absolute essential for successful application and interpretation of QSPR models[J]. QSAR & Combinatorial Science, 2003, 22(1): 69-77.

[33]GRAMATICA P, PILUTTI P, PAPA E. Validated QSAR prediction of OH tropospheric degradation of VOCs: splitting into training-test sets and consensus modeling[J]. Journal of Chemical Information and Computer Sciences, 2004, 44(5): 1794-1802.

[34]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: 2044-2068.

(责任编辑:殷丽莉)