定量结构-性质关系(QSPR)中的计算方法研究进展

张钰¹, 魏世丞^{1*}, 董超芳², 王博¹, 梁义¹, 王玉江¹, 陈茜¹

1. 陆军装甲兵学院装备再制造技术国防科技重点实验室, 北京 100072;

2. 北京科技大学腐蚀与防护中心, 北京 100083

* 联系人, E-mail: wsc33333@163.com

2020-10-22 收稿, 2020-12-07 修回, 2020-12-28 接受, 2021-01-12 网络版发表

国家重点研发计划(2019YFB1311100)、国家自然科学基金(51675533, 51701238和51905543)、国防科技卓越青年科学基金(2017-JCJQ-ZQ-001)、“十三五”装备预研共用技术项目(404010205)和中国博士后科学基金(2018M643857)资助

摘要 定量结构-性质关系(QSPR)方法是一种将材料的微观定量结构与其某些性能构建关系的方法. 通过构建的关系模型可以对具有其他结构的材料性能进行预测, QSPR已成为材料基因组计划的主要实现途径之一. 计算方法是决定QSPR模型准确程度、构建速度的主要因素, 选择最佳的计算方法对改善QSPR模型预测精度、计算速度等指标至关重要. 本文首先对QSPR模型进行了概述, 简单介绍了QSPR方法的计算步骤; 重点说明了QSPR中分子描述符计算方法、模型优化算法, 并对各个方法进行了优缺点分析; 最后总结了近些年QSPR中各类计算方法的应用趋势.

关键词 定量结构-性质关系, 材料基因组, 模型优化算法, 分子描述符, 机器学习

新材料的技术突破引领产业革命, 然而传统的“爱迪生”试错法开发新材料耗时久、投入高, 无法满足当今快速发展的工业需求, 采用新技术开发新材料迫在眉睫^[1]. 2011年6月, 美国白宫提出材料基因组计划(Materials Genome Initiative, MGI), 旨在利用数据驱动发现新材料, 缩短科研周期, 从传统的经验试错式实验转为理论预测指导实验^[2]. 2012年年底, 我国《材料科学系统工程发展战略研究——中国版材料基因组计划》重大项目启动, 其中一个重要内容就是高通量计算与材料预测^[3]. 这种绿色环保型低消耗、更快速、高通量的“计算-实验”方法已成为新的材料研究趋势, 体现了材料基因组计划较强的指导意义和实际应用价值^[4-8]. 材料基因组计划主要有两个实现途径, 即发展量子化学计算和统计力学、建立定量结构-性质关系(quantitative structure-property relationship, QSPR), 而后者凭借优秀的预测特性成为近年来的研究热点^[9].

QSPR方法自1962年被提出以来, 广泛应用在材

料、生物、医药、环境等领域^[10]. QSPR方法旨在找出材料的宏观性质与分子水平微观结构之间的数学关系, 从而预测材料的性质, 以寻找具有期待性质的新材料^[11]. 当今, 材料基因组计划极大地促进了QSPR方法的发展, 尤其体现在其计算方法的改变. 在QSPR中合理使用计算方法, 能够大幅提高QSPR模型的形成速度、预测准确程度, 这对预测新材料十分关键.

基于材料基因组计划, 对QSPR方法进行深入研究, 对我国新材料的研发工程具有重要意义. 结合大数据和人工智能的技术背景, 了解与掌握QSPR方法中的计算方法是使用QSPR方法的核心要求. 因此, 本文首先对QSPR方法进行简要介绍, 其次重点总结归纳了目前QSPR方法中几种主要的分子描述符计算方法, 最后评述了QSPR方法中近年常用的优化算法及其相关应用.

1 QSPR方法概述

QSPR方法作为“爱迪生”试错方法的替代方法, 最

引用格式: 张钰, 魏世丞, 董超芳, 等. 定量结构-性质关系(QSPR)中的计算方法研究进展. 科学通报, 2021, 66: 2832-2844

Zhang Y, Wei S C, Dong C F, et al. Progress in QSPR modelling methods (in Chinese). Chin Sci Bull, 2021, 66: 2832-2844, doi: [10.1360/TB-2020-1362](https://doi.org/10.1360/TB-2020-1362)

早以多元线性原理为理论基础,通过对材料的结构和指定性能建立数学关系,形成QSPR模型,对具有其他假设结构的材料性能进行预测,以寻找具有预期性能的材料结构^[11]。QSPR方法广泛应用于材料化学、生物医药、环境等领域,是目前预测材料性能的最重要手段之一^[12]。

1995年, Katritzky等人^[11]提出的QSPR研究流程图作为传统QSPR方法的代表,如图1(a)所示,主要包括以下五部分: (1) 准备输入数据; (2) 三维(3D)结构优化; (3) 分子描述符计算; (4) 统计学分析; (5) QSPR预测。在材料基因组的背景下,随着大数据理念以及机器学习算法的引入, QSPR方法也发生了较大的变化。图1(b)所示为目前常用的QSPR方法实现流程。

QSPR模型构建步骤主要包括:

(1) 准备输入数据集。QSPR方法是一种有监督训练,需要收集待研究材料已有的实验数据,包括所研究材料的结构和相应的性质。传统的QSPR方法不需要庞大的数据量,但是材料基因组使其对数据量需求较大,因此需在开始时构建一个输入数据集。QSPR输入数据集的构建一般有3种方法,即从实验或前期工作中获取、从已发表论文收集以及从数据库收集。此外,为了测试模型的准确度,通常还需要将输入数据集划分成训练数据集和测试数据集两部分。划分方法主要有Kennard Stone算法^[13]、随机选择法^[14]等。

(2) 分子描述符的计算。分子描述符是指分子在某一或某些方面的度量,包括分子的物理化学性质和基于分子结构计算出的数据指标,如键长、最高占据分

图1 (网络版彩色)QSPR研究方法对比图

Figure 1 (Color online) QSPR study method comparison chart

子轨道(highest occupied molecular, HOMO)、最低未占分子轨道(lowest unoccupied molecular, LUMO)等,是QSPR方法的重要组成部分^[15]。根据不同方法,分子描述符有多种分类方式。按照定量定性分类,可以将其分成定量描述符和定性描述符。定量描述符包括分子场描述符和分子形状描述符;定性描述符通常被称为“分子指纹”,即使用某种编码表示分子的结构、性质或片段等。QSPR方法采用一定的计算方式将材料的分子结构转化成可以用算法计算的分子描述符形式。如图2所示, Ogunyemi等人^[16]采用HOMO和LUMO作为分子描述符以研究嘌呤衍生物的缓蚀作用^[16]。

(3) 分子描述符的筛选。分子描述符在MGI中担任材料“基因”的核心角色,对其合理筛选对性能预测的准确度至关重要。当前计算性能的大幅提升使计算得

图2 (网络版彩色)嘌呤衍生物的HOMO(a)和LUMO(b)轨道^[16]

Figure 2 (Color online) HOMO (a) and LUMO (b) orbitals for the purine derivatives^[16]

到的分子描述符数量达到几百甚至上千,因此采用合适的方法选择影响比重较大的若干分子描述符,可以有效地排除噪音数据干扰,从而提升模型的构建速度和质量。

(4) QSPR模型优化. 传统的QSPR模型优化方法通常选择统计学方法,基于多元线性原理在分子描述符与所研究性质之间建立数学关系,直观看出各个分子描述符在模型中的比重. 随着机器学习方法的引入,机器学习算法也常被用来构建QSPR模型,比如人工神经网络、支持向量机、决策树、随机森林等. 此外,还有很多研究人员使用混合式方法构建QSPR模型,综合多种方法的优点改善模型。

(5) QSPR预测. 利用训练数据集和优化算法构建QSPR模型后,使用测试数据集对模型进行测试. 由于QSPR方法的目的是预测出某一性能的指标,因此QSPR中采用的算法模型多为回归模型. 回归预测过程主要采用5个参数对模型进行检验,即均方误差(mean squared error, MSE)、均方根误差(root mean square error, RMSE)、平均绝对误差(mean absolute error, MAE)以及决定系数 R^2 . 前3个参数越小说明预测结果越准确, R^2 越接近100%说明模型越拟合。

2 QSPR模型中分子描述符的计算

QSPR在材料基因组中的应用思路是将分子描述符看作材料的一个个“基因”,计算这些微观的“基因”以预测出宏观的材料性质. 选择分子描述符的方法主要有以下三类: (1) 经验选择法; (2) 算法描述法; (3) 软件计算法,其对比如表1所示。

(1) 经验选择法. 经验选择法适用于对所研究的材料非常熟悉的情况,且直接选择的分子描述符对所研究性能具有直接影响的关系. Ser等人^[17]在预测缓蚀剂性能时,选择了前沿轨道、激发能、吸附能等20个与缓蚀相关的化学参数作为QSPR模型中的分子描述符。

Sharifian等人^[18]对有机化合物进行研究时,选择了物质的分子量、原子极化率的总和及每种物质中的原子数作为分子描述符,利用QSPR模型预测了这些有机化合物的标准摩尔化学能。

(2) 算法描述法. 直接计算尺度不同的分子描述符对QSPR模型的准确度会有较大影响,因此研究人员采用算法描述法. Wang等人^[19]对污染物毒性进行研究时,基于化学图论提出了范数指数描述符、定义位置矩阵描述符以及原子性质矩阵描述符,并通过矩阵公式计算得到了QSPR模型的矩阵形式分子描述符. Jia等人^[20]对有机化合物正辛醇/水分配系数预测研究中,在范数指数描述符的基础上,改善了矩阵公式,提高了分子描述符的信息权重,进而提升了预测能力。

(3) 软件计算法. 前两种方法由于选择的分子描述符个数有限,且有时由于经验限制而导致某些重要分子描述符被忽略,使得计算准确度有限. 因此,避免上述两个问题的软件计算法是当前最为常用的方法. 最早由Milano Chemometrics和QSAR研究小组^[49]开发的Dragon专门用于WHIM描述符的计算. 目前Dragon已经更新到版本7.0,分子描述符数量已从1600多个增长到5270个,因此被广泛使用(https://chm.kode-solutions.net/products_dragon.php)^[50]. 为了得出爆速与高能共晶体分子结构之间的最佳关系,Zohari和Mohammadkhani^[23]使用Dragon计算分子描述符,并选择了多达488个不同类型的分子描述符. Joudaki和Shafiei^[40]使用Dragon计算了环烷烃化合物数据集中每种化合物的1502个理论描述符. Amiri等人^[41]对43种有机磷杀虫剂研究时,使用Dragon软件计算产生的理论描述符作为自变量来开发具有4个描述符的QSPR模型。

由Yap课题组^[51]开发的PaDEL-Descriptor软件是一款基于Java语言的开源软件,主要利用The Chemistry Development Kit(CDK)程序包计算分子描述符,并配有图形用户界面. Ghosh等人^[42]为了构建有害合成有机

表1 分子描述符选择方法

Table 1 Molecular descriptor selection method

方法	优点	缺点	文献
经验选择法	简单、快速	要求对所研究材料非常熟悉; 容易忽略某些对性质有影响的分子描述符	[17,18]
算法描述法	统一了分子描述符尺度,提高了准确度	容易忽略某些对性质有影响的分子描述符; 增加了计算量,实现较复杂,用时长	[19-21]
软件计算法	快速计算出大量分子描述符	计算得到的分子描述符数量过大,后续需要进行预处理	[22-48]

化学物质(SOCs)在单壁碳纳米管(SWCNTs)的吸附QSPR模型,使用PaDEL-Descriptor软件计算扩展的拓扑化学原子(extended topochemical atom, ETA)指数。Zhu等人^[25]为了得到预测低密度聚乙烯(LDPE)中疏水性有机污染物(HOCs)扩散系数的QSPR模型,使用PaDEL-Descriptor软件计算得到1305个分子描述符。

由Katritzky等人^[52]指导开发的CODESSA软件是用于开发定量构效/性质关系(QSAR/QSPR)的综合程序,将QSPR方法中的数学和化学工具集成,并根据化合物2D/3D结构以及量子化学波函数计算出各类分子描述符。Banchero和Manna^[45]对有机化合物的临界温度、临界压力和偏心因子研究时,使用CODESSA软件计算出400个分子描述符。Yang等人^[44]为了研究55种双(亚氨基)吡啶铁和钴配合物催化性能的QSPR模型,使用CODESSA软件计算分子描述符。CORAL是一款由Toropova等人^[30-34]开发以建立QSPR模型的免费软件,通过分子描述符构建端点模型来预测材料的性能, CORAL也可以计算并优化分子描述符。另外,有些研究人员自行开发程序以计算分子描述符,例如, Wu等人^[43]开发了iqspr程序包,计算聚合物的分子描述符来构建聚合物的玻璃转移温度与熔点温度的QSPR模型。

一些具有综合集成功能的分子模拟软件中也包含分子描述符计算模块,如Gaussian、VASP、Materials Studio等。Liu等人^[46]采用Gaussian 09计算有机化合物的理论分子描述符; Wang等人^[47]使用VASP对金属氧化物纳米颗粒的理论分子描述符进行计算; Ma等人^[48]

对64种二元共沸物计算了336个分子描述符。表2汇总了本文介绍的分子描述符计算软件。

3 QSPR模型中的主要算法

在QSPR模型中,算法主要应用在分子描述符的预处理以及模型的优化两方面:预处理利用算法对分子描述符进行筛选,模型优化则用算法拟合构建出一个最准确的QSPR预测模型。

3.1 分子描述符预处理方法

采用软件算法得到的分子描述符有时候多达几千个,使后续计算量大且速度慢;某些分子描述符具有密切相关性,使用这些冗余描述符可能会导致QSPR模型出现过拟合的情况。因此,为了后续的模型优化,通常对计算出的分子描述符进行预处理,去除冗余的特征描述符。目前常用的分子描述符预处理方法有以下几种,如表3所示。

主成分分析法(principle component analysis, PCA)是一种常用的降维方法,通过正交变换将高维数据映射到低维,将一组相关变量转换成一组更小的不相关新变量,达到既保留描述符绝大部分特征信息又减少描述符数量的目的。Belhassan等人^[60]用PCA方法确定了51种气味分子(15种醇、11种醛、9种酮和16种酯)分子描述符的非线性,并用其筛选出与气味分子保留/释放特性高度相关的描述符。Ermondi和Caron^[61]对77种化合物的色谱保留系数QSPR预测中,使用PCA方法对

表2 分子描述符计算软件

Table 2 Molecular descriptor calculation software

软件	类型	优点	缺点	应用平台	文献
Dragon	付费	可以计算5270个分子描述符;图形用户界面;批处理	不能计算扩展拓扑化学原子(ETA)指数	Windows/Linux	[49]
PaDEL-Descriptor	免费	可移植性强;擅长计算ETA指数	最大计算分子描述符数量为1875	Mac/Windows/Linux	[51]
CODESSA	付费	集成了有关QSPR的必要的数学和计算工具	数年未更新	Windows/AXI/Solari	[52]
CORAL	免费	可以建立与纳米材料有关的终点模型	只接受SMILES格式	Windows	[34]
iqspr	免费	包括机器学习算法和蒙特卡洛方法	只接受SMILES格式;仅提供应用程序接口	R语言环境平台	[43]
Gaussian	付费	全面的量子化学软件包	Windows平台计算较慢	Mac/Windows/Unix/Linux	[53]
VASP	付费	可以快速准确计算周期性结构	没有图形用户界面	Linux	[54]
Materials Studio	付费	可以用其他模块优化材料结构;能够方便建立整个QSPR模型	并行计算效率低	Windows/Unix/Linux	[55]

表3 分子描述符预处理方法

Table 3 Molecular descriptor preprocessing method

方法	类型	优点	缺点	文献
PCA	统计学	结果容易理解; 无参数限制	降维过程无法人为干预; 数据集需满足高斯分布	[56]
PLS	统计学	适合处理高维小样本数据; 有监督方法	冗余特征数据影响降维结果	[57]
mRMR	信息学	可处理高维数据	计算时间长	[58]
GA	生物学	可利用经验选择设计; 计算时间少, 鲁棒性高; 可处理大样本数据	实现相对复杂	[59]

化合物分子描述符构建的数据库进行初步分析和离群值检测, 大幅提高了计算效率。

偏最小二乘法(partial least squares, PLS)也是一种与主成分有关的统计方法, 与PCA方法不同, PLS将主成分分析、多元线性回归、典型相关三者结合, 使筛选后的特征变量既能很好概括原始特征, 还能对输出变量有较强的解释能力。Yang等人^[44]对催化剂进行研究时, 采用了PLS方法对计算得到的分子描述符进行筛选, 防止出现过拟合以及偶然相关的问题。Noda等人^[62]使用QSPR方法对钙钛矿型氧化物的介电常数预测时, 对计算得到的分子描述符进行了PLS回归分析, 以提取合适的描述符。

最大相关最小冗余(maximum-relevance minimum-redundancy, mRMR)是一种特征选择方法, 其核心是找到一个特征子集, 使特征与变量之间的相关度最大化, 特征之间的相关度最小化。Mirshahvalad等人^[63]为了预测来自不同化学族的纯有机化合物的闪点, 使用mRMR方法从1926个计算出的描述符中筛选出20个最有效的描述符。

遗传算法(genetic algorithm, GA)是一种强大的随机全局搜索和优化方法, 基于生物学中生物进化的规律提出, 选择具有最大适应度的个体即最优解输出。Zhang等人^[64]研究钙钛矿居里温度时, 设计了在线材料数据挖掘计算平台OCPMDM, 并采用了基于GA的特征选择方法, 为预测QSPR模型提供了紧凑的描述符特征集。Wang等人^[47]构建金属氧化物纳米颗粒QSPR模型时, 利用R程序实现GA, 用于筛选QSPR模型的分子描述符。Joudaki和Shafiei^[40]预测环烷烃的量子性质时, 使用GA筛选出具有最高预测能力的且最少量的分子描述符。Yao等人^[65]从分子结构预测二元生物柴油混合物的闪点, 使用GA获得了与闪点显著相关的最佳描述符子集。

传统的统计学方法虽然计算相对简单, 但是对成百上千的分子描述符进行筛选时存在计算速度较慢、降维不准确的问题。随着计算机科学的迅速发展, GA的实现愈发简便。因此, 近3年分子描述符预处理方法使用率较高的是GA以及PLS, 如图3所示, 然而具体方法的选择仍需要结合分子描述符计算的结果以及研究的体系。

3.2 模型优化方法

以分子描述符与预测性能之间的关系来划分, 模型构建的优化方法可分为三大类, 即线性优化方法、非线性优化方法以及混合优化方法, 如图4所示。QSPR方法常用的线性方法主要是多元线性回归和偏最小二乘回归; 常用的非线性方法主要为人工神经网络、支持向量机、K最近邻算法、决策树、随机森林等机器学习算法; 结合实际计算的需求, 很多研究综合了多种优化方法, 也取得了不错的预测效果。

3.2.1 线性优化方法

多元线性回归(multiple linear regression, MLR)是最常用的回归方法之一, 通常用来研究多个变量与因变量之间的线性关系。在QSPR模型优化中, MLR旨在研究众多分子描述符与某一特性之间的线性关系, 构建材料性能的预测模型。Zohari和Mohammadkhani^[23]采用MLR方法构建了通过分子结构(sp, OB, DU, n(AT))预测相应高能共晶体密度的模型, 为预测新型高能共晶体提供了理论指导。

偏最小二乘回归(partial least squares regression, PLSR)是基于偏最小二乘法的回归方法, 是评估两组自变量和因变量之间关系的多元线性方法。在QSPR方法中, PLSR建立分子描述符与材料性能的线性关系, 从而可以预测其他结构的相应性能。Luo等人^[67]使用PLSR构建手性化合物的分离因子的QSPR模型, 对50

图 3 (网络版彩色)通过任意混合方法选择的前10个重要分子描述符的条形图^[66]
 Figure 3 (Color online) Bar chart of top 10 important descriptors selected by any blending^[66]

图 4 (网络版彩色)QSPR优化方法
 Figure 4 (Color online) QSPR optimization methods

种手性化合物的分离因子进行预测, 预测结果显示准确度达到96.9%。

3.2.2 非线性优化方法

人工神经网络(artificial neural networks, ANN)是基于模拟大脑神经网络处理模式的一种非线性算法, 凭借其能够识别各种高度非线性系统中复杂关系的优势, 成

为近年来的研究热点。在QSPR中, ANN能够利用其固有的能力对纯物质和混合物的物理性质进行预测, 并且避免因缺乏信息而带来的模型局限性。Safa等人^[68]在对有机物溶解焓的预测模型中, 使用具有3-6-1拓扑型的ANN优化QSPR预测模型, 预测结果显示 $R^2=0.9796$, 接近最佳数值1, 且高于相同条件下MLR模型。

支持向量机(support vector machine, SVM)是一种有监督的分类器, 基于材料结构-性能的标准数据集可以预测具有其他结构的材料性能。Wang等人^[29]利用SVM建立了燃料混合物的过热极限温度的QSPR非线性模型, 实现了对过热极限温度的快速预测。Jin等人^[69]基于SVM建立了醇类、酮类、苯类等二元液体混合物自燃温度的QSPR模型, 结果显示预测模型具有较好的拟合性、鲁棒性和外部可预测性。

K最近邻算法(k-nearest neighbor, kNN)是一种通过测量不同特征值间距离来进行分类的算法, 即用 k 个和待测样本距离最近值的类别代表该样本的类别。kNN是最简单的机器学习方法之一, 因此在QSPR方法中也广泛应用。Mettou等人^[70]为了改善化合物的水溶性, 构建了化合物溶解度的QSPR模型并使用kNN方法($k=5$)优化, 对55种化合物的溶解度进行测试后验证了模型具有较高准确度。

决策树(decision tree, DT)是一种具有树形结构的分类模型,可对数据进行分类.模型中每个节点代表对分子描述符性质的测试,分支代表测试的结果,叶片代表分类后的性质标记. Tangestani等人^[71]使用DT预测457种不同水污染物的羟基自由基速率常数,与LSSVM和ANFIS两种方法(混合优化方法,见3.2.3节)相比,DT具有更高的准确度. Chaudhari等人^[72]使用DT开发了65种可燃粉尘的最小燃烧能的QSPR预测模型,正确预测测试集中的最小燃烧能类别的概率为95%.

随机森林(random forest, RF)是一种基于集成学习的分类和回归方法,将若干决策树集合在一起各自对数据进行训练学习和预测,再把预测结果结合成单预测,即得到优于单分类器的预测模型. Rakhimbekova等人^[73]对4种不同分子化学反应数据集构建QSPR预测模型,并分别采用RF方法和高斯过程回归方法进行优化,结果显示RF方法在该研究中优于高斯过程回归方法.

3.2.3 混合优化方法

遗传函数算法(genetic function approximation, GFA)是由遗传算法和多元自适应回归样条结合的一种优化方法.与其他方法不同,GFA可以在模型的构造中使用更广泛的基函数,例如样条函数、高斯函数或更高阶多项式,最早由Rogers和Hopfinger^[74]提出时便应用于QSPR模型建立中,现在可以在Materials Studio中直接使用. Ma等人^[48]在Materials Studio中运用GFA针对64种二元共沸物构建了二元共沸物组成-共沸温度的QSPR模型,测试结果表明该模型具有较好的预测性能,提高了分离含低级醇的二元共沸物的有效性.

自适应神经模糊系统(adaptive network-based fuzzy inference system, ANFIS)是基于ANN和模糊逻辑(fuzzy logic, FL)提出的一种模糊推理系统结构,与ANN相比,可以更好地表达近似人脑的推理. Sharifian等人^[18]基于ANFIS建立了QSPR模型以实现有机化合物的标准摩尔化学能值的预测,并使用粒子群算法(particle swarm optimization, PSO)对ANFIS的参数进行了优化,实验结果显示预测表现出色, R^2 达到0.9999. Dashti等人^[75]使用ANFIS设计预测有机化合物的自燃温度的QSPR模型,并分别使用遗传算法、粒子群算法、差异进化算法(differential evolution, DE)对ANFIS进行优化.实验结果显示,对有机化合物,该QSPR模型仅使用9个分子描述符即可对自燃温度进行准确预测.

最小二乘法支持向量机(least squares support vector machine, LSSVM)是对SVM计算过程中的损失函数

计算使用最小二乘法优化,使原方法的不等式约束转换成线性方程求解,进而提高运算速率. Aghaie和Zendeheboudi^[76]对离子液体中二氧化碳溶解度预测时,选用LSSVM对QSPR预测模型进行了优化,结果显示LSSVM的预测准确程度要优于MLR; Rahmani等人^[38]使用LSSVM和GA-MLR优化化合物对乳腺癌细胞HCT15的最大半数抑制活性的预测模型,结果显示LSSVM预测准确度明显优于GA-MLR.

由于遗传算法(GA)、粒子群算法(PSO)、差分进化算法(DE)等进化算法具有优秀的全局优化能力,很多研究人员利用GA、PSO、DE等对已有的方法进行优化,以达到快速生成最优QSPR模型的目的,比如GA-MLR^[38,40,46,65,77-80]、GA-PLS^[17,81]、GA-SVM^[82]、GA-ANN^[83]、GA-ANFIS^[84]、BPSO-MLR^[85]、PSO-MLR^[86]、PSO-SVR^[87].

表4为QSPR优化方法的优缺点分析,可以看出机器学习算法已渐趋成熟,然而单一算法的劣势在现实需求中逐渐突出,因此综合利用多种优化方法对原有算法进行一定的改进,对于提升模型整体的质量具有重要作用.此外,对于MLR、ANN、DT、RF等需要多次迭代寻找最佳参数的算法,提升模型构建速度还可以通过引入优化算法得以实现.

4 小结与展望

伴随着材料基因组的大力发展,QSPR方法有了新的发展方向,并且近年来在各个领域广泛应用.作为QSPR模型构建速度、准确程度的关键,QSPR中的计算方法在近年来也有了多样化发展趋势.本文主要对近3年广泛应用于QSPR中的计算方法研究进展进行了介绍,并对各类方法进行优劣势分析.

(1)目前的分子描述符计算软件种类繁多,而且大部分软件计算速度比较快,计算的分子描述符数量大,可以达到较好的计算结果.但是各个软件仍然存在自己的局限,研究人员可以根据自己的需要选择某个或者某些软件进行计算.

(2)预处理算法方面主要为筛选算法,包括统计学方法、信息学方法以及生物学方法.其中,由生物学方法发展出来的进化类算法,凭借其优秀的全局筛选能力,愈来愈多被应用.随着计算得到的分子描述符数量大幅提升,对分子描述符全局性意义的逐渐重视,近几年来进化类算法在分子描述符筛选方面逐渐成为热门选择.

表 4 QSPR优化方法

Table 4 QSPR optimization method

优化方法	类型	优点	缺点	文献
MLR	线性	简单方便; 可以直观表现每个成分在模型中的影响比重; 在模型和输入数据相同的情况可得到唯一的预测结果	仅适于自变量与因变量具有线性关系的情况	[88]
PLSR	线性	可用于曲线拟合和小样本数据	不适用于自变量数量少的情况	[89]
ANN	非线性	避免了缺少局部信息对模型的影响	容易过拟合	[90]
SVM	非线性	稳定, 不受数据微小改变影响	计算内存需求高; 计算大量数据时用时长	[91]
kNN	非线性	擅长处理大样本数据	计算量较大	[92]
DT	非线性	可以识别异常值; 算法直观	需要大样本训练; 处理缺失值能力有限	[93]
RF	非线性	不容易过拟合; 擅长处理高维数据、遗漏特征数据	需要大样本训练	[94]
GFA	混合	可以建立多模型体系	与MLR相比预测准确度较低	[95]
ANFIS	混合	可以使用低维数据训练	迭代次数多计算较慢	[96]
LSSVM	混合	与SVM相比计算速度更快	与SVM相比预测准确度较低	[76]

(3) 优化方法主要为机器学习算法, 具体分类为线性方法、非线性方法以及混合方法, 前两种方法在各个成分重要性表达以及模型准确度方面各有优势。为

了提高算法参数生成的效率, 近些年越来越多地引入全局优化算法作为参数的优化方法, 形成混合优化方法。

参考文献

- Xiang Y, Yan Z K, Zhu Y L, et al. Progress on materials genome technology (in Chinese). J Univ Electron Sci Technol China, 2016, 45: 634–649 [向勇, 闫宗楷, 朱焱麟, 等. 材料基因组技术前沿进展. 电子科技大学学报, 2016, 45: 634–649]
- Andrea W. Materials Genome Initiative. Chem Eng News, 2013, 31: 25–27
- Wan Y, Feng R H, Wang G F. Comparison analysis and implications of U.S. and European materials genetic engineering policy (in Chinese). Sci Technol Manage Res, 2014, 34: 69–72 [万勇, 冯瑞华, 王桂芳. 美欧材料基因工程技术政策对比及其启示. 科技管理研究, 2014, 34: 69–72]
- Fan X L. Materials Genome Initiative and first-principles high-throughput computation (in Chinese). Mater China, 2015, 34: 689–695 [范晓丽. 材料基因组计划与第一性原理高通量计算. 中国材料进展, 2015, 34: 689–695]
- Zhao J C. A perspective on the Materials Genome Initiative (in Chinese). Nat J, 2014, 36: 89–104 [赵继成. 材料基因组计划简介. 自然杂志, 2014, 36: 89–104]
- Liu L M. Material genome engineering: Materials design and simulation (in Chinese). J New Ind, 2015, 5: 71–88 [刘利民. 材料基因工程: 材料设计与模拟. 新型工业化, 2015, 5: 71–88]
- Liu Z K. Perspective on Materials Genome® (in Chinese). Chin Sci Bull (Chin Ver), 2013, 58: 3618–3622 [刘梓葵. 关于材料基因组的基本观点及展望. 科学通报, 2013, 58: 3618–3622]
- Yin H Q, Jiang X, Zhang R J, et al. Role of materials data in materials innovation development and thoughts on the existing problems (in Chinese). Mater China, 2017, 36: 401–405 [尹海清, 姜雪, 张瑞杰, 等. 材料数据在材料创新发展中的作用与存在问题的思考. 中国材料进展, 2017, 36: 401–405]
- Li Y Q, Liu L Y, Chen W D, et al. Materials genome: Research progress, challenges and outlook (in Chinese). Sci Sin-Chim, 2018, 48: 243–255 [李云琦, 刘伦洋, 陈文多, 等. 材料基因组学的发展现状、研究思路与建议. 中国科学: 化学, 2018, 48: 243–255]
- Hansch C, Maloney P P, Fujita T, et al. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. Nature, 1962, 194: 178–180
- Katritzky A R, Lobanov V S, Karelson M. QSPR: The correlation and quantitative prediction of chemical and physical properties from structure. Chem Soc Rev, 1995, 24: 279–287
- Toropov A A, Toropova A P. QSPR/QSAR: State-of-Art, weirdness, the future. Molecules, 2020, 25: 1292
- Rahal S, Hadidi N, Hamadache M. In silico prediction of Critical Micelle Concentration (CMC) of classic and extended anionic surfactants from

- their molecular structural descriptors. *Arab J Sci Eng*, 2020, 45: 7445–7454
- 14 Mathieu D. QSPR versus fragment-based methods to predict octanol-air partition coefficients: Revisiting a recent comparison of both approaches. *Chemosphere*, 2020, 245: 125584
- 15 Ren W, Kong D X. Correlation of molecular descriptors in QSAR study (in Chinese). *Comput Appl Chem*, 2009, 26: 1455–1458 [任伟, 孔德信. 定量构效关系研究中分子描述符的相关性. *计算机与应用化学*, 2009, 26: 1455–1458]
- 16 Ogunyemi B T, Latona D F, Adejoro I A. Molecular modeling and quantitative structure-property relationships (QSPRs) of purine derivatives as corrosion inhibitor in acid medium. *Sci Afr*, 2020, 8: e00336
- 17 Ser C T, Žuvela P, Wong M W. Prediction of corrosion inhibition efficiency of pyridines and quinolines on an iron surface using machine learning-powered quantitative structure-property relationships. *Appl Surf Sci*, 2020, 512: 145612
- 18 Sharifian S, Madadkhani M, Rahimi M, et al. QSPR based ANFIS model for predicting standard molar chemical exergy of organic materials. *Pet Sci Technol*, 2019, 37: 2174–2181
- 19 Wang Q, Jia Q, Yan L, et al. Quantitative structure-toxicity relationship of the aquatic toxicity for various narcotic pollutants using the norm indexes. *Chemosphere*, 2014, 108: 383–387
- 20 Jia Q, Shi Q, Yan F, et al. Norm index-based QSPR model for describing the n-octanol/water partition coefficients of organics. *Environ Sci Pollut Res*, 2020, 27: 15454–15462
- 21 Safa F. Atomic-level topological indices for prediction of the infinite dilution activity coefficients of oxo compounds in water. *J Solut Chem*, 2020, 49: 222–238
- 22 Zohari N, Ghiasvand Mohammadkhani F. Prediction of the density of energetic co-crystals: A way to design high performance energetic materials. *Cent Eur J Energ Mater*, 2020, 17: 31–48
- 23 Zohari N, Mohammadkhani F G. Detonation velocity assessment of energetic cocrystals using QSPR approach. *Z Anorg Allg Chem*, 2020, 646: 30–35
- 24 Zohari N, Fareghi-Alamdari R, Sheibani N. Model development and design criteria of hypergolic imidazolium ionic liquids from ignition delay time and viscosity viewpoints. *New J Chem*, 2020, 44: 7436–7449
- 25 Zhu T, Jiang Y, Cheng H, et al. Development of pp-LFER and QSPR models for predicting the diffusion coefficients of hydrophobic organic compounds in LDPE. *Ecotox Environ Saf*, 2020, 190: 110179
- 26 Zhu T, Chen W, Wu J, et al. Predicting low density polyethylene-water partition coefficients based on pp-LFER and QSPR models using molecular descriptors. *Fluid Phase Equilib*, 2020, 506: 112374
- 27 Zhou L, Wang B, Jiang J, et al. A mathematical method for predicting flammability limits of gas mixtures. *Process Saf Environ Protect*, 2020, 136: 280–287
- 28 Wu T, Li W, Chen M, et al. Estimation of ionic liquids toxicity against Leukemia rat cell line IPC-81 based on the empirical-like models using intuitive and explainable fingerprint descriptors. *Mol Inf*, 2020, 39: 2000102
- 29 Wang B, Zhou L, Liu X, et al. Prediction of superheat limit temperatures for fuel mixtures using quantitative structure-property relationship model. *J Loss Prev Process Ind*, 2020, 64: 104087
- 30 Toropova A P, Toropov A A, Leszczynska D, et al. The index of ideality of correlation: Models of the flash points of ternary mixtures. *New J Chem*, 2020, 44: 4858–4868
- 31 Toropova A P, Toropov A A, Carnesecchi E, et al. The index of ideality of correlation: Models for flammability of binary liquid mixtures. *Chem Pap*, 2020, 74: 601–609
- 32 Toropova A P, Toropov A A, Carnesecchi E, et al. The using of the Index of Ideality of Correlation (IIC) to improve predictive potential of models of water solubility for pesticides. *Environ Sci Pollut Res*, 2020, 27: 13339–13347
- 33 Toropov A A, Toropova A P, Marzo M, et al. Pesticides, cosmetics, drugs: Identical and opposite influences of various molecular features as measures of endpoints similarity and dissimilarity. *Mol Divers*, 2020, doi: 10.1007/s11030-020-10085-3
- 34 Toropov A A, Raška Jr I, Toropova A P, et al. The study of the index of ideality of correlation as a new criterion of predictive potential of QSPR/QSAR-models. *Sci Total Environ*, 2019, 659: 1387–1394
- 35 Talevi A, Morales J F, Hather G, et al. Machine learning in drug discovery and development Part 1: A primer. *CPT Pharmacometrics Syst Pharmacol*, 2020, 9: 129–142
- 36 Rybińska-Fryca A, Sosnowska A, Puzyn T. Representation of the structure—A key point of building QSAR/QSPR models for ionic liquids. *Materials*, 2020, 13: 2500
- 37 Rezaei B, Riahi S, Gorji A E. Molecular investigation of amine performance in the carbon capture process: Least squares support vector machine approach. *Korean J Chem Eng*, 2020, 37: 72–79
- 38 Rahmani N, Abbasi-Radmoghaddam Z, Riahi S, et al. Predictive QSAR models for the anti-cancer activity of topoisomerase II α catalytic inhibitors against breast cancer cell line HCT15: GA-MLR and LS-SVM modeling. *Struct Chem*, 2020, 31: 2129–2145

- 39 Maouz H, Laidi M, Hamadache M, et al. QSPR studije karbonilnih, hidroksilnih, polienskih indeksa i prosječne molekulske težine polimera pod fotostabilizacijom pristupom ANN i MLR. *Kem Ind*, 2020, 69: 1–16
- 40 Joudaki D, Shafiei F. QSPR models to predict thermodynamic properties of cycloalkanes using molecular descriptors and GA-MLR method. *Curr Comput-Aided Drug Des*, 2020, 16: 6–16
- 41 Amiri R, Messadi D, Bouakkadia A, et al. QSPR models for the prediction of octanol/water partition coefficient of organophosphorous insecticides. *Egypt J Chem*, 2019, 62: 1563–1574
- 42 Ghosh S, Ojha P K, Roy K. Exploring QSPR modeling for adsorption of hazardous synthetic organic chemicals (SOCs) by SWCNTs. *Chemosphere*, 2019, 228: 545–555
- 43 Wu S, Kondo Y, Kakimoto M, et al. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. *npj Comput Mater*, 2019, 5: 66
- 44 Yang W, Ahmed S, Fidelis T T, et al. Effect of cycloalkyl-fused ring on the catalytic performance of bis(imino)pyridine metal complexes by QSPR modeling. *Catal Commun*, 2019, 132: 105820
- 45 Banchero M, Manna L. Comparison between multi-linear- and radial-basis-function-neural-network-based QSPR models for the prediction of the critical temperature, critical pressure and acentric factor of organic compounds. *Molecules*, 2018, 23: 1379
- 46 Liu S, Jin L, Yu H, et al. Understanding and predicting the diffusivity of organic chemicals for diffusive gradients in thin-films using a QSPR model. *Sci Total Environ*, 2019, 706: 135691
- 47 Wang J, Wang Y, Huang Y, et al. Development of a nano-QSPR model to predict band gaps of spherical metal oxide nanoparticles. *RSC Adv*, 2019, 9: 8426–8434
- 48 Ma Y, Ma K, Wang H, et al. QSPR modeling of azeotropic temperatures and compositions for binary azeotropes containing lower alcohols using a genetic function approximation. *Chin J Chem Eng*, 2019, 27: 835–844
- 49 Mauri A, Consonni V, Pavan M, et al. Dragon software: An easy approach to molecular descriptor calculations. *Match-Commun Math Chem*, 2006, 56: 237–248
- 50 Yu X. Prediction of chemical toxicity to *Tetrahymena pyriformis* with four-descriptor models. *Ecotox Environ Saf*, 2020, 190: 110146
- 51 Yap C W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem*, 2011, 32: 1466–1474
- 52 Katritzky A R, Perumal S, Petrukhin R, et al. CODESSA-based theoretical QSPR model for hydantoin HPLC-RT lipophilicities. *J Chem Inf Comput Sci*, 2001, 41: 569–574
- 53 Lu C, Ni S, Sun Y, et al. Prediction of thermodynamic properties of actinide and lanthanide compounds. *Nucl Sci Techn*, 2008, 19: 152–158
- 54 Isayev O, Fourches D, Muratov E N, et al. Materials cartography: Representing and mining material space using structural and electronic fingerprints. *Am Chem Soc Div Fuel Chem*, 2015, 60: 656
- 55 Sharma S, Kumar P, Chandra R. Molecular Dynamics Simulation of Nanocomposites Using BIOVIA Materials Studio, Lammmps and Gromacs. *Holland: Elsevier*, 2019. 329–341
- 56 Puzyn T, Falandysz J. Computational estimation of logarithm of n-octanol/air partition coefficient and subcooled vapor pressures of 75 chloronaphthalene congeners. *Atmos Environ*, 2005, 39: 1439–1446
- 57 Cardoso S P, Hollauer E, Borges L E P, et al. QSPR prediction analysis of corrosion inhibitors in hydrochloric acid on 22%-Cr stainless steel. *J Braz Chem Soc*, 2006, 17: 1241–1249
- 58 Peng H C, Long F H, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*, 2005, 27: 1226–1238
- 59 Andzelm J W, Alvarado-Swaigood A E, Axe F U, et al. Heterogeneous catalysis: Looking forward with molecular simulation. *Catal Today*, 1999, 50: 451–477
- 60 Belhassan A, Chtita S, Lakhlifi T, et al. QSPR study of the retention/release property of odorant molecules in pectin gels using statistical methods. *J Taibah Univ Sci*, 2017, 11: 1030–1046
- 61 Ermondi G, Caron G. MLR, PLSR-BR analysis and MBPLSR to interpret multivariate QSPR Models. The case of a micellar liquid chromatography descriptor ($\log K_w^{SDS}$). *Mol Inf*, 2019, 38: 1800144
- 62 Noda Y, Otake M, Nakayama M. Descriptors for dielectric constants of perovskite-type oxides by materials informatics with first-principles density functional theory. *Sci Technol Adv Mater*, 2020, 21: 92–99
- 63 Mirshahvalad H, Ghasemiasl R, Raoufi N, et al. A neural network QSPR model for accurate prediction of flash point of pure hydrocarbons. *Mol Inf*, 2019, 38: 1800094
- 64 Zhang Q, Chang D, Zhai X, et al. OCPMDM: Online computation platform for materials data mining. *Chemom Intell Lab Syst*, 2018, 177: 26–34
- 65 Yao J, Qi R, Pan Y, et al. Prediction of the flash points of binary biodiesel mixtures from molecular structures. *J Loss Prev Process Ind*, 2020, 65: 104137
- 66 Chen C H, Tanaka K, Kotera M, et al. Comparison and improvement of the predictability and interpretability with ensemble learning models in

- QSPR applications. *J Cheminf*, 2020, 12: 19
- 67 Luo C, Hu G, Huang M, et al. Prediction on separation factor of chiral arylhydantoin compounds and recognition mechanism between chiral stationary phase and the enantiomers. *J Mol Graphics Model*, 2020, 94: 107479
- 68 Safa F, Dakhel A A, Shariati S. Predictive artificial neural network model for solvation enthalpy of organic compounds in *N,N*-dimethylformamide. *Russ J Phys Chem*, 2019, 93: 2661–2668
- 69 Jin Y, Jiang J, Pan Y, et al. Prediction of the auto-ignition temperature of binary liquid mixtures based on the quantitative structure-property relationship approach. *J Therm Anal Calorim*, 2020, 140: 397–409
- 70 Mettou A, Papaneophytou C, Melagraki G, et al. Aqueous solubility enhancement for bioassays of insoluble inhibitors and QSPR analysis: A TNF- α study. *Slas Discov*, 2018, 23: 84–93
- 71 Tangestani E, Keivanimehr F, Ghadiri M. Insights into the estimation of hydroxyl radical rate constant of water contaminants in AOP using new smart QSPR models. *Water Environ J*, 2020, 35: 312–321
- 72 Chaudhari P, Ade N, Pérez L M, et al. Quantitative structure-property relationship (QSPR) models for minimum ignition energy (MIE) prediction of combustible dusts using machine learning. *Powder Technol*, 2020, 372: 227–234
- 73 Rakhimbekova A, Madzhidov T I, Nugmanov R I, et al. Comprehensive analysis of applicability domains of QSPR models for chemical reactions. *Int J Mol Sci*, 2020, 21: 5542
- 74 Rogers D, Hopfinger A J. Application of genetic function approximation to quantitative structure-activity relationships and quantitative structure-property relationships. *J Chem Inf Model*, 1994, 34: 854–866
- 75 Dashti A, Jokar M, Amirkhani F, et al. Quantitative structure property relationship schemes for estimation of autoignition temperatures of organic compounds. *J Mol Liq*, 2020, 300: 111797
- 76 Aghaie M, Zendejboudi S. Estimation of CO₂ solubility in ionic liquids using connectionist tools based on thermodynamic and structural characteristics. *Fuel*, 2020, 279: 117984
- 77 Shen S, Ji X, Pan Y, et al. A new method for predicting the upper flammability limits of fuel mixtures. *J Loss Prev Process Ind*, 2020, 64: 104074
- 78 Rajabi M, Shafiei F. Structure-property relationships of aliphatic esters using topological descriptors and backward multiple linear regression method. *J Chin Chem Soc*, 2020, 67: 1338–1347
- 79 Pan Y, Qi R, He P, et al. Thermal hazard assessment and ranking for organic peroxides using quantitative structure-property relationship approaches. *J Therm Anal Calorim*, 2020, 140: 2575–2583
- 80 Borhani T N G, Afzali A, Bagheri M. QSPR estimation of the auto-ignition temperature for pure hydrocarbons. *Process Saf Environ Protect*, 2016, 103: 115–125
- 81 Shiri F, Salahinejad M, Momeni-Moogaei N, et al. Predicting stability constants of transition metals Y³⁺, La³⁺, and UO₂²⁺ with organic ligands using the 3D-QSPR methodology. *J Recept Signal Transduct*, 2021, 49: 59–66
- 82 Qi J, Wei J, Sun C, et al. A comparative QSPR study on aqueous solubility of polycyclic aromatic hydrocarbons by GA-SVM, GA-RBFNN and GA-PLS. *Front Earth Sci*, 2011, 5: 245–251
- 83 Wilczyńska-Piliszek A J, Piliszek S, Falandysz J. QSAR and ANN for the estimation of water solubility of 209 polychlorinated *trans*-azobenzenes. *J Environ Sci Health Part A*, 2012, 47: 155–166
- 84 Jalili-Jahani N, Zeraatkar E, Fatehi A, et al. Pixels of chemical structures correlate to chromatographic detector responses using genetic algorithm-adaptive neuro-fuzzy inference system as a novel nonlinear feature selection method. *Chemom Intell Lab Syst*, 2020, 202: 104032
- 85 Borhani T N G, Saniedanesh M, Bagheri M, et al. QSPR prediction of the hydroxyl radical rate constant of water contaminants. *Water Res*, 2016, 98: 344–353
- 86 Bagheri M, Borhani T N G, Zahedi G. Estimation of flash point and autoignition temperature of organic sulfur chemicals. *Energy Convers Manage*, 2012, 58: 185–196
- 87 Pei J F, Cai C Z, Zhu Y M, et al. Modeling and predicting the glass transition temperature of polymethacrylates based on quantum chemical descriptors by using hybrid PSO-SVR. *Macromol Theor Simul*, 2013, 22: 52–60
- 88 Chen F F, Breedon M, White P, et al. Correlation between molecular features and electrochemical properties using an artificial neural network. *Mater Des*, 2016, 112: 410–418
- 89 Begam B F, Kumar J S, Chae G S. Optimized peer to peer QSPR prediction of enthalpy of formation using outlier detection and subset selection. *Peer-to-Peer Netw Appl*, 2018, 11: 1156–1165
- 90 Hook A L, Alexander M R, Winkler D A. Materiomics: A toolkit for developing new biomaterials. In: Van Blitterswijk C A, De Boer J. *Tissue Engineering*. 2nd ed. Oxford: Academic Press, 2014. 253–281
- 91 Frenkel M. A never-ending search for the truth: Thermodynamics in the uncertain era of the internet. *J Chem Thermodyn*, 2015, 84: 18–40
- 92 Melagraki G, Afantitis A. Enalos KNIME nodes: Exploring corrosion inhibition of steel in acidic medium. *Chemom Intell Lab Syst*, 2013, 123: 9–14

- 93 Gupta S, Basant N. Modeling the pH and temperature dependence of aqueousphase hydroxyl radical reaction rate constants of organic micropollutants using QSPR approach. [Environ Sci Pollut Res](#), 2017, 24: 24936–24946
- 94 Chen C H, Tanaka K, Funatsu K. Random forest approach to QSPR study of fluorescence properties combining quantum chemical descriptors and solvent conditions. [J Fluoresc](#), 2018, 28: 695–706
- 95 Das S, Ojha P K, Roy K. Development of a temperature dependent 2D-QSPR model for viscosity of diverse functional ionic liquids. [J Mol Liq](#), 2017, 240: 454–467
- 96 Sepehri B. A review on created QSPR models for predicting ionic liquids properties and their reliability from chemometric point of view. [J Mol Liq](#), 2020, 297: 112013

Summary for “定量结构-性质关系(QSPR)中的计算方法研究进展”

Progress in QSPR modelling methods

Yu Zhang¹, Shicheng Wei^{1*}, Chaofang Dong², Bo Wang¹, Yi Liang¹, Yujiang Wang¹ & Xi Chen¹¹ National Key Laboratory for Remanufacturing, Army Academy of Armored Forces, Beijing 100072, China;² Corrosion and Protection Center, University of Science and Technology Beijing, Beijing 100083, China* Corresponding author, E-mail: wsc33333@163.com

With the introduction of Materials Genome Initiative (MGI) in 2011, high-throughput material computation and material prediction have gradually become a hot topic in materials research. MGI has strong guiding significance and high practical application value as the deepening of the concept of environmental protection, low-consumption, faster, and high-throughput. MGI aims to convert the traditional “Trial Experiment” to “Computation to Experiment”, which has become the trend of materials research. There are two main ways to realize the MGI, the first one is the development of quantum chemical calculations and statistical mechanics, while the other one is the establishment of Quantitative Structure-Property Relationship (QSPR). Among them, the QSPR method has been widely used in recent years due to its excellent predictive properties. Combining data-driven and machine learning algorithms, the QSPR method has undergone major changes today, the most prominent of which is the change in the calculation method used in the QSPR method. Reasonable use of calculation methods in QSPR can greatly increase the speed of QSPR model formation and the accuracy of prediction, which is critical to predicting new materials. The calculation methods are the main factors that determine the accuracy and construction speed of the QSPR model. Choosing the best calculation method is very important to improve the prediction accuracy and calculation speed of the QSPR model. Therefore, this article first summarizes the QSPR model, briefly introduces the calculation steps of the QSPR method, focuses on the calculation methods of molecular descriptors and model optimization algorithms in QSPR, and analyzes the advantages and disadvantages of each method. Finally, this article summarizes the application trend of various calculation methods in QSPR in recent years, and concludes as follow: (1) There are many kinds of molecular descriptor calculation software, and most of the software calculation speed is quite fast. At the same time, the calculated number of molecular descriptors is pretty large, which can achieve better calculation results. However, each software still has its own limitations. Researchers can choose one or some software to calculate molecular descriptors according to their needs. (2) The preprocessing algorithms are mainly screening algorithms, including statistical methods, informatics methods, and biological methods. Evolutionary algorithms, which are developed by biological methods, are increasingly being used because of their excellent global screening capabilities. While the number of molecular descriptors obtained by calculation increase significantly, the global influence of molecular descriptors has been gradually emphasized. Therefore, in recent years, evolutionary algorithms have gradually become a popular choice for molecular descriptor screening. (3) The optimization methods mainly include machine learning algorithms, which can be classified into linear methods, nonlinear methods, and hybrid methods. The first two methods have their own advantages for linear methods can show directly the importance of each molecular descriptor and the nonlinear methods’ models are more accurate. In order to improve the efficiency of algorithm parameter calculation, in recent years, more and more global optimization algorithms have been introduced into machine learning algorithms as parameter optimization methods, which forms the third method—Hybrid optimization method.

quantitative structure-property relationship, Materials Genome Initiative, model optimization algorithm, molecular descriptor, machine learning

doi: [10.1360/TB-2020-1362](https://doi.org/10.1360/TB-2020-1362)