中国科学 B 辑:化学 2008 年 第 38 卷 第 6 期: 461 ~ 474
www.scichina.com chem.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

面向毒害有机物生态风险评价的(Q)SAR 技术: 进展与展望

陈景文*, 李雪花, 于海瀛, 王亚南, 乔显亮

工业生态与环境工程教育部重点实验室, 大连理工大学环境科学与工程系, 大连 116024

* 联系人, E-mail: jwchen@dlut.edu.cn

收稿日期: 2007-08-14; 接受日期: 2007-12-03

国家重点基础研究计划(973)(批准号: 2006CB403302)资助项目

摘要 有机物结构-活性关系(SAR)和定量结构-活性关系(QSAR)技术(合并简写为(Q)SAR 技术), 在毒害有机物生态风险评价中, 可以发挥弥补基础数据的缺失、降低昂贵的测试费用、减少动物实验、评估数据的不确定性等方面的作用. 经过几十年的发展, (Q)SAR 技术的发展, 呈现出目标导向性和应用性、多学科集成性、智能性三个特点, 并在发达国家, 得到了深入的研究和广泛的应用. 许多国家已经将该技术用于有机化学品的生态风险评价与管理工作中. 本文总结了(Q)SAR 的基本原理与方法, 以及面向生态风险评价的(Q)SAR 技术的最新进展, 包括模型预测的环境指标的确定, 实验数据的选择方法, 建立简单透明、可移植性好并可进行机理解释的(Q)SAR 模型的数学方法; 重点总结了模型应用域的定义与界定、离群点的诊断, 并从模型的拟合优度评价、稳定性分析和预测能力评估三方面介绍了模型性能评价方法. 本文对于促进面向毒害有机物生态风险性评价的(Q)SAR 技术的研究和发展, 具有一定的指导意义.

关键词
(Q)SAR
生态风险评价
应用域
稳定性
预测能力

1 合成有机化学品的生态风险性评价

据估计, 人类目前日常使用的合成有机化学品已达 8 万种之多, 并且这个数字正以每年 500~1000 种的速度增加. 截至 2007 年末, 美国化学文摘社(CAS: http://www.cas.org)登记的化学品已达 3300 多万种, 其中绝大多数是合成有机物. 合成有机化学品所导致的污染问题, 给人类带来了惨痛教训. 持久性有毒物质(PTS)的污染问题, 已经成为制约人类在 21 世纪生存和发展的重要环境问题. 显然, 对这些合成有机化学品进行生态风险性评价(ERA), 是预防和控制其污染的前提[1]. 美国环保局(US EPA)将ERA分为 3 个主要阶段: 问题表征(危害评估、确定评价指标、制定分析计划)、分析(暴露评价、效应评价及其关系)、风险表征(风险评估、描述与报告)[2]. 显然, 有机化合物的物理化学性质、环境行为与生态毒理学数据, 是进行生态风险性评价的基础. 但是, 这些数据存在 3 方面的问题:

(1) 数据缺失[3]. 例如, 对于 80%以上的日用合成有机化学品, 人类尚缺乏其环境行为和生态毒理方面的信息. 通过实验方法来测定这些数据, 在时间上是滞后的, 不能满足有毒有害化学品污染管理的“预先防范原则”.

(2) 测试费用昂贵[4]. 例如, 据欧盟于 2007 年 6 月开始全面实施的化学品管理新法规“化学品注册、评估、授权和限制法规(简称REACH法规)”估算, 每一种化学物质的基本检测费用约需 8.5 万欧元(不含长期环境影响的评估费用), 每一新物质全面检测费用约需 57 万欧元, 这意味着如果对每种化学品都开展实验测定, 需要巨额的费用. 进行全面的实验测试, 也不符合化学品管理中的减少实验(尤其是动物实验)的原则和趋势.

(3) 数据存在不确定性[5]. 例如, 美国地质调查所的科学家发现, 被全世界科学家广泛研究的农药DDT及其代谢产物DDE的辛醇/水分配系数(K_{OW})的实验测定值, 不同实验室的测定结果, 竟有几个数量级大小的误差. 如此大的不确定性, 很显然会导致生态风险评价结果更大的不确定性.

分子结构是决定有机物的物理化学性质在环境中迁移转化行为和生态毒理学效应的内因. 具有类似分子结构的物质, 也可能具有类似的物理化学性质、环境归趋和生态毒理学效应, 即: 有机物的物理化学性质、环境行为和生态毒理学参数, 与其分子结构之间存在内在联系; 这种联系是可以被认识、表征和应用的. 这种内在的联系, 以模型的方式表征出来, 就是结构-活性关系(SAR)和定量结构-活性关系(QSAR), 统称为(Q)SAR[6]. 因此, (Q)SAR可以弥补有机物环境行为与生态毒理数据的缺失, 大幅度降低实验费用, 有助于减少和替代实验(尤其动物实验). 此外, 由于这种内在的可表征的关系, 有机物尤其是系列化合物的物理化学性质、环境行为和生态毒理学参数的大小及其变化趋势, 必然与其分子结构的变化相一致, 所以(Q)SAR有助于评价实验数据的不确定性, 这也是 (Q)SAR技术在ERA中所发挥的重要作用之一[7]. 例如, 多氯联苯(PCB)系列物的物理化学性质一致性, 可以依据QSAR原理, 采用分子量和邻位氯取代基的数量进行阐明[8]. 综上所述, (Q)SAR技术对于有机污染物的生态风险性评价具有重要意义.

2 (Q)SAR 的基本原理与发展历程

2.1 (Q)SAR 的基本原理与方法

人类很早就认识到有机物的分子结构与其物理化学性质和生物活性之间存在内在的联系. 20 世纪 30 年代Hammett等人[9,10]所建立的线性自由能关系(LFER)理论, 为(Q)SAR奠定了热力学理论基础. Hammett等人[9,10]创造性地提出了表示取代基电子效应的参数σ, Taft[11]提出了表示取代基立体效应的参数Es. LFER属于超热力学关系, 即: 尽管热力学参数(分子结构参数)与活性之间的关系是客观存在的, 但热力学理论并不能推导出这种关系[12]. LFER在表征有机污染物在多介质环境中的平衡分配系数[13,14]和反应速率常数[15,16]中发挥了重要作用.

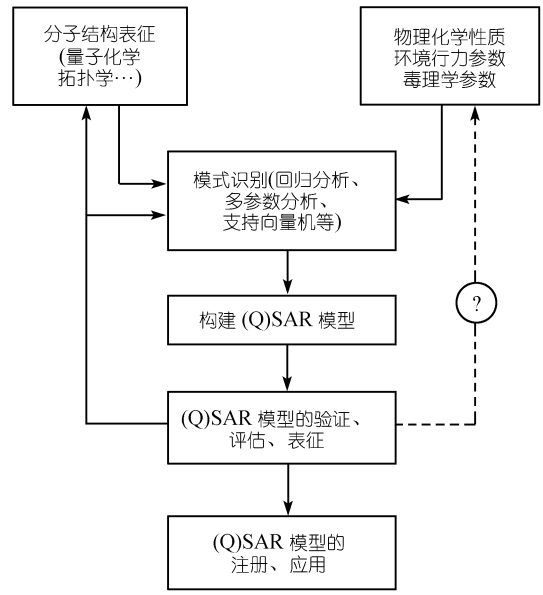

图 1 (Q)SAR 模型的建立流程

如图 1 所示, 获取和选取表征分子结构的参数(亦称为分子结构描述符), 是(Q)SAR模型构建的基础性工作. 主要有两种方法来选取分子结构参数, 第一种是借助于经验、分子的结构特征和物理-化学过程的机理来选取. 例如, 光致水解是卤代芳烃光解的主要途径之一, 因此建立卤代芳烃光解量子产率的QSARs时, 选用了各种表征碳-卤键性质的量子化学描述符[17,18]. 第二种方法是借助于模型来选取, 即所谓的模型方法. QSAR研究中经常使用的模型主要包括Hansch模型、线性溶解能相关模型、Free-Wilson 模型[19]以及三维QSAR分析方法(例如CoMFA)[20,21]等.

(1) Hansch 模型: 在LFER基础上, Hansch把QSAR 的研究范围扩大到了生物活性领域, 提出取代基对化合物生物活性(1/C)的影响主要是电性效应

(σ)、立体效应(E_s)以及疏水效应(π), 并且这些效应可以彼此独立加和[22~24]. Hansch方程存在线性和非线性形式[24,25], 在QSAR领域应用广泛[26,27].

(2) 线性溶解能相关(LSER)模型: Kamlet等人[28~31]发展的线性溶解能关系(LSER)是LFER的扩展, LSER模型包含空穴项、偶极项和氢键项, 并采用分子体积和溶剂化变色参数来表征溶质-溶剂相互作用. Abraham[32]进一步发展了新的LSER参数. Wilson和Famini[33,34]通过以理论计算的参数替代LSER模型中的经验性参数, 衍生出了理论线性溶解能关系(TLSER)模型. LSER和TLSER模型在有机污染物的水溶解度(S_w)[28,29], 正辛醇/水分配系数(K_{ow})[30,31]、高效液相色谱保留因子[35]以及非反应性毒性[36,37]的QSAR构建方面取得了很大成功.

(3) Free-Wilson模型[19]: 由Free和Wilson于 1964 年提出, 认为系列化合物活性的变化取决于特定取代基在母体结构上数量和位置变化. 该方法计算简单, 但只适合存在多取代的情况.

(4) 三维QSAR分析方法: 最常见的是比较分子力场分析(CoMFA), 其核心是作用于同一受体的一系列生物活性分子, 与受体之间的各种作用力场应该有一定的相似性[20,21]. 因此, 在不了解受体三维结构的情况下, 研究生物活性分子周围作用力场的分布, 并与化合物分子的生物活性定量联系起来, 既可以推测受体的某些性质, 又可以设计新的化合物, 并定量预测化合物活性. 该方法在定量药物设计中应用广泛[38,39]; 在生态毒理学中亦得到应用, 例如内分泌干扰物的雌激素活性[40,41].

2.2 (Q)SAR 的发展趋势与特点

早期(Q)SAR 主要应用于药物设计领域. 20 世纪 70 年代以来, 出于对环境中大量的、不断增长的合成有机化学品的生态风险评价的需要, (Q)SAR 在环境科学中得到广泛应用, 并持续稳定发展. 纵观(Q)SAR 在过去几十年的发展历程, 可以发现其呈现如下 3 个趋势和特点:

(1) 目标导向性和应用性. 在环境科学技术领域, (Q)SAR研究一直主要围绕有机污染物生态风险评价中的暴露评价(污染物在多介质环境中的迁移和转化)与效应评价(污染物的生态毒理学效应)的目标而展开, 具有显著的目标导向性和应用性特点. 从所模拟的对象来看, 早期多针对有机污染物环境分配方面的参数(例如S_W[28,29]、K_{OW}[30,31,42]、生物富集因子(BCF)[43]、辛醇-空气分配系数(K_{OA})[44,45]、土壤(沉积物)吸附系数(K_{OC})[46,47]等)和对水生生物的急性毒性(半数致死浓度(LC_{50})[48]或效应浓度(EC_{50})[49]). 近期QSARs发展为模拟有机物污染物的环境内分泌干扰效应[50~52]以及反应速率常数(例如生物降解能力[53]、光解速率常数[54]与量子产率[17,18]、零价铁催化反应速率常数[15]、羟基自由基氧化反应速率常数[16])等.

1993 年, 期刊SAR and QSAR in Environmental Research在法国创刊. 自 1988 年起, 国际上每两年召开一次环境科学中QSAR学术讨论会[55]. 2003 年, 国际知名期刊Environmental Toxicology and Chemistry 的 22 卷第 8 期, 集中刊出了 23 篇QSAR的综述性文章, 涵盖有机化合物的物理化学性质、环境归趋、生物活性及生态效应等方面的内容, 集中而详细地介绍了QSAR 在环境领域的发展和应用. 这些都标志着环境科学中(Q)SAR的研究和应用方兴未艾.

(2) 多学科集成性. (Q)SAR 是多学科交叉的研究领域, 汇集化学信息学(化学计量学、计算化学)、物理化学、生物化学、毒理学、计算机科学、数学等多个学科的研究成果, 日益体现多学科集成性的特点.

从分子结构表征的角度看, 从早期通过实验测得的疏水性常数(π)[22,24,56]、电子效应常数(σ)[10,57,58]、立体效应常数(E_S)[11,59~61]及溶剂化参数[28~30,62,63]等经验分子结构描述符, 发展到目前广泛使用的拓朴学参数[64,65]、量子化学参数[66]等理论分子结构描述符. 例如, Dragon软件可以计算出 1000 余种分子结构描述符, 代表 0~3 维分子空间结构并且涵盖原子、化学键类型、连接性、电荷分布、原子空间坐标等信息[67]. 数学、分子拓扑学、量子化学、计算机数值计算等学科的融合发展, 使得对分子结构的表征更加细致全面, 为成功建立(Q)SAR模型奠定了良好基础.

从模型建立的角度看, 从最初的各种线性回归分析技术[68~70], 发展到综合应用各种多变量分析方法, 如: 因子分析与主成分分析(PCA)[71]、判别分析[71,72]、聚类分析[71,73]、偏最小二乘(PLS)回归分析[74]. 近年来还发展使用了一些非线性的建模技术[75], 如人工

神经网络(ANN)[71,76]、支持向量机(SVM)[77,78]等. 遗传算法(GA)[79,80]等优选方法亦用于变量的筛选之中. 同时, 产生了一系列的组合算法, 例如GA-PLS[81,82], GA-SVM[83], GA-BP[84], SVM-PLS[85]等. 这些方法的应用, 促进了模型建立技术的不断完善. 同时, 生物化学、毒理学等学科的发展, 使得对毒性作用机制的认识不断深入, 亦推动了(Q)SAR技术的不断发展.

(3) 智能性. 近年来, 由于计算机技术的发展, 一些政府部门、公司和研究机构开发了智能性较强、界面友好、面向不同用户、各具特色的(Q)SAR 应用软件. 经济合作与发展组织(OECD)统计了以有机化学品管理为宗旨的(Q)SAR 软件, 其中美国具有著作权的有 40 个, 英国有 3 个, 法国有 6 个, 加拿大有 8 个, 保加利亚 1 个. 如果包括各种(Q)SAR 软件, 保守估计有 200 个以上.

(Q)SAR未来的发展方向之一是建立决策支持系统 [86]. 该系统应该包含符合标准的模型, 实验测定和模型预测值数据库, 具有灵活的搜索引擎, 界面友好, 有合适的工具帮助进行模型选择, 并且可以通过互联网络获得. 通过这样体系的构建, 实现资源共享, 帮助非(Q)SAR研究人员正确使用这些模型, 在管理和决策领域发挥作用.

3 目前各国发展和应用(Q)SAR 技术的情况

由于(Q)SAR技术有助于实现有机化学品管理的“预先防范原则”, 能够替代相关的试验并可大幅降低测试费用, 因此, 世界各国纷纷开发和应用面向毒害有机物生态风险评价与管理的(Q)SAR技术 [7,87]. 截至 2002 年, 美国、加拿大、澳大利亚、德国、丹麦、日本和荷兰等国家, 均不同程度地应用(Q)SAR技术来预测有机化学品的物理化学性质、环境归趋和对水生生物的毒性 [88], 所涉及参数包括: K_{OW}、K_{OC}、S_W、沸点(B_p)、熔点(M_p)、蒸气压(P)、亨利定律常数(K_H)、在空气中的氧化速率、水解速率常数、*BCF*、生物降解性等 [88].

REACH法规提出了化学品监管的 3 条原则 [4]: ① “无安全信息便无市场”原则, 即: 在产品投放市场之前, 化学品公司必须提供产品安全信息; ② 减少实验尤其是动物实验的原则. 一方面为了降低实验的费用, 另一方面为了满足西方国家所倡导的动物保护理念; ③ 应用(Q)SAR技术的原则.

REACH法规规定, 如果(Q)SAR技术满足如下 4 方面的条件, 则(Q)SAR的预测结果就可以替代试验测试 [4,89]: 1) (Q)SAR模型的科学有效性已经得到证实; 2) 所预测的物质在(Q)SAR模型的应用域之内; 3) 所预测的结果足够用于化学品分类、标记和风险评价的目的; 4) 提供了足够和可靠的记录来描述所使用的方法. 欧盟的QSAR技术导则(TGD)中, 给出了(Q)SAR在化学物质生态效应和环境归趋预测方面的 4 个作用: 评估实验数据、决定是否进行进一步的测试实验、估计特定参数、确定潜在的数据需求 [7].

欧洲化学品署(ECB)(http://ecb.jrc.it/)是欧盟负责有害化学品风险评价的核心官方机构, 负责实施 REACH 法规的技术支持. 近年来, ECB 围绕(Q)SAR 技术的开发和应用, 开展了大量的研究工作. 主要涉及 3 方面: 1) (Q)SAR 模型的报告格式、验证与评估方法; 2) 化学品分类技术; 3) 理化性质、环境行为或毒理参数的类比(Analogue 或 Read-Across)技术, 涉及 (Q)SAR 技术在不同目标层面上的应用.

OECD 也围绕化学品的安全性问题, 开展了(Q)SAR技术的应用研究. 2004 年 11 月, OECD提出了验证(Q)SARs模型的一些原则 [90]. 2007 年 2 月, OECD 发布了关于确认和验证(Q)SAR模型的指导文件 [91]. OECD围绕(Q)SARs在现有和新化学品管理中的应用, 组织开展了案例研究. 涉及的国家包括澳大利亚、加拿大、捷克共和国、丹麦、德国、意大利、日本、荷兰、美国、英国和欧盟委员会. 2006 年 8 月, OECD发布了该案例研究的报告 [92].

美国有多个政府部门研发和应用(Q)SAR技术, 包括: US EPA、空军(the U. S. Air Force)、有毒物质和疾病注册管理局(the Agency for Toxic Substances and Disease Registry, ATSDR)、有毒物质控制法案内部测试委员会(the Toxic Substance Control Act Interagency Testing Committee)、国家海洋大气管理局(the National Oceanic Atmospheric Administration, NOAA)、消费品安全委员会(Consumer Product Safety Commission, CPSC)、食品与药品管理局(Food and Drug Administration, FDA)、国立癌症研究所(National Cancer Institue, NCI)、国家毒理学计划(National Toxicology Program)等 [88].

2002 年, US EPA 在国会的要求下, 加强了计算毒理学的研究工作, 核心是开发和验证用于化学品筛选和确定优先污染物的非动物试验方法, 主要是(Q)SAR 方法. 在 US EPA 专门成立了国家计算毒理学研究中心(http://www.epa.gov/comptox/index.html). 近年来, US EPA 的总财政预算被削减了很多, 但在计算毒理学研究(主要为(Q)SAR)领域的预算却逐年有所增加.

US EPA 开发了 EPI Suite™ 软件(http://www.epa.gov/oppt/exposure), 包括预测 K_{OW}、K_{OC}、H、S_W、B_p、M_p、P、BCF、生物降解性、空气中的氧化速率、水解速率、污水处理厂去除效率等的子程序. US EPA 还应用 QSAR 技术预测大批量生产的化学品(HPV)和需要生产前告知(PMN)化学品的生物效应, 包括吸收、分配、代谢、排泄、急性效应、刺激性、致敏性、慢性或亚慢性效应、生殖效应、发育毒性、致癌性、致突变性等. 此外, US EPA 还应用 QSAR 预测化学品的雌激素效应. 关于其他国家应用 QSAR 技术的详情, 可以参阅文献[86,88].

(Q)SAR 的相关研究成果, 以论文形式发表的多于专利. 2006 年底, 以“QSAR”为关键词在标题和摘要中检索, 欧洲专利局(EPO)的 Worldwide Database 中检索到 22 个公开专利; 世界知识产权组织(WIPO)的专利数据库中检索得到 11 个公开专利; 美国专利商标局(USPTO)的数据库中, 检索到 8 个专利.

综上所述, 发达国家(Q)SAR技术的发展趋势可以概括为: 已经得到高度重视, 并在有机物生态风险评价与管理中日益得到应用. 针对其应用中的技术问题, 开展了大量的研究工作. 在环境科学技术领域, 在国家自然科学基金的资助下, 我国也开展了(Q)SAR方面的一些基础研究工作. 代表性的研究单位有南京大学 [93,94]、大连理工大学 [95,96]、湖南大学 [97]、兰州大学 [77,98]、长春应用化学研究所 [99]、东北师范大学 [100]等, 但总的来说, 开展的不系统, 也不深入, 在(Q)SAR技术的应用层面尚未开展实质性研究工作, 需要迎头赶上.

4 (Q)SAR 技术的新进展与前瞻

(Q)SAR技术的应用涉及多方面因素. 2002 年在Setubal召开的(Q)SAR研讨会对其应用和发展提出了初步指导意见, 即: 面向ERA的(Q)SAR应该符合如下标准 [7,86,89]: 1) 具有明确定义的环境指标; 2) 具有明确的算法; 3) 定义了模型的应用域; 4) 有适当的拟合度, 稳定性和预测能力; 5) 最好能够进行机理解释. 2004 年, OECD正式确定上述准则为(Q)SAR模型发展和使用的导则 [89], 符合这些条件的模型, 可以应用于化合物的ERA、化学品筛选以及优先控制等管理工作 [86,87]. 下面主要围绕上述问题, 对相关工作进行总结.

4.1 (Q)SAR 模型预测的环境指标

(Q)SAR 的环境指标(变量)是指任何能被测量和预测的物理化学、环境行为与生态毒理学参数. 这些指标可以在标准条件下, 采用规范的方法, 通过实验方法测定. 明确(Q)SAR 模型的环境指标, 可以判断模型的预测值是否适合于特定的 ERA.

研究表明, 高质量的实验数据是建立优秀(Q)SAR模型的重要基础 [101]. 最标准的数据应该是相同实验室相同工作人员采用统一的标准方法测定的 [102], 不同来源的实验数据间的系统差异, 会对(Q)SAR模型质量产生不可预知的影响 [103]. 同时, 应尽可能确保建立模型的训练集化合物有较大的结构差异性, 扩大训练集的物理化学空间, 增强模型的稳健性 [104]. 然而由于实验数据的限制, 实际工作中经常采用来自于不同文献的环境指标数据, 这样虽然会扩大数据范围, 提高结构差异, 但容易导致不精确的预测结果. 因此, 模型的拟合结果必须考虑实验数据误差, 保证拟合度要在环境指标数据的变化范围之内; 否则会不恰当地模拟误差信息, 造成模型过拟合.

4.2 建立 QSAR 模型的数学算法

应用于有机化学品管理和生态风险评价的QSAR模型, 最好具有简单、透明、容易解释、易于移植的数学算法. 所谓透明, 是指模型应基于基本的物理化学性质, 并具有清晰明确的表达形式 [101]. 一个透明的模型才有利于进行机理解释, 便于不同研究和管理人员之间的交互使用, 并且允许使用者查看和理解环境指标被预测的全过程. 这样的模型品质主要通过适当的统计数学方法来实现 [102].

模型所使用的统计分析方法应该具备一定的透

明性, 即通过该方法的实施, 获得相关的处理过程信息[102]. 研究表明[105~107], 不同方法的透明性依次为: 多元回归分析(MLR)>主成份和偏最小二乘分析(PCA&PLS) >人工神经网络(ANN) >遗传算法(GA). 然而, 模型的透明性又是与模型的稳健性相关联的, 后者是指模型应用范围和条件的相对自由程度, 且其顺序刚好与透明性相反[102]. 所以, 统计分析方法的选择, 应该综合模型的用途、考虑环境指标的需求、模型透明性和稳健性等相关指标.

4.3 (Q)SAR 模型的机理

(Q)SAR 模型的建立, 应该基于对机理的正确分析和解释; 反过来, 所建立的(Q)SAR 模型, 应该进一步有助于机理的解释. 机理解释可以明确影响化合物生态风险指标的分子结构因素, 进而判断是否可以用于新物质的 ERA. 模型的机理解释性, 主要通过如下两方面实现:

(1) 建立模型所使用的分子结构描述符, 应有利于模型的机理解释. 所以要尽可能选择具有明确物理化学意义的分子结构描述符[101,102]. 比较而言, 一些基础性质描述符(如分子量)和量子化学描述符较以原子和碎片为基础的结构和拓朴指数更易于解释[108].

(2) 与不断发展的生物化学、毒理学相结合, 深入对化合物毒性作用机理的认识, 提高模型的机理解释性.

4.4 (Q)SAR 模型的应用域

(1) (Q)SAR 模型应用域的表征

在ERA中应用(Q)SAR技术需要克服的难点之一, 就是表征模型的应用域(AD). 经验的(Q)SAR模型仅在验证的域内是有效的, 应用于域外的物质会导致严重的预测错误[109]. 模型的AD与模型的确认和验证密切相关. 所谓模型的确认与验证, 就是针对模型的某个预测功能, 证明在其AD内具有令人满意的预测准确度[110]. 因此, AD可以定义为: 经确认和验证, 某模型所适用的化合物集合[111]. 在实践中, 需要一个可操作的、可用计算机程序执行的方法来具体定义模型的应用域[112].

对应用域的研究, 首先可以从建立模型所使用描述符的角度来展开, 即训练集化合物所覆盖的描述符空间的组合, 也称之为描述符域[113]. 训练集的选择会直接影响模型描述符的空间范围[113].

其次, 考虑训练集和预测集化合物之间的结构相似性[114], 得到结构域. 结构域是基于分子相似性概念的, 对于预测来讲, 与训练集化合物分子相似性高的化合物会比相似性低的化合物得到更准确的预测结果[114]. 有些情况下, 模型的结构相似性是基于经验知识或假定的作用模式的[115]. 所以, 基于不同的定义结构相似性的方法, 可能得到不同的结构域.

分子结构描述符包含在模型的描述符空间中, 并且结构与训练集化合物的结构相似, 这两个条件是判断化合物是否处于模型应用域之中的必要条件[116]. 然而满足这两个条件并不能确保预测的可靠性和正确性, 还需要引入机理域的概念, 即测试集化合物的化学反应或毒性作用机理应该与训练集化合物相一致. 机理域的定义通常需要描述分子的亚结构, 并认为分子结构类似的物质具有类似的反应或毒性机理[116]. 机理域是保证模型预测准确度和精确度的最严格标准.

此外, 如果在毒性作用过程中发生了新陈代谢, 那么还应该从模拟代谢的角度定义代谢域[110]. 忽略代谢作用会给毒理作用指标的判断带来困难, 这也是传统的(Q)SAR模型中经常出现的问题[110].

综上, 可从 4 方面来表征模型的应用域: 1) 描述符变化范围; 2) 结构相似性; 3) 机理相似性; 4) 新陈代谢. 这 4 方面的交集, 构成了(Q)SAR模型最保守的应用域. 在实际应用中, 可根据(Q)SAR模型的实验数据的质量、所模拟的环境指标与实际应用目标, 确定(Q)SAR应用域的最佳表征方式[110].

(2) (Q)SAR 模型离群值的诊断

模型离群值(离域点)的诊断是十分重要的, 因为离域点的存在会给模型带来很多问题. 从模型的角度来讲, 典型离域点表现为: 化合物对于数据集是非稳定性的, 或表现在生物学上的不同作用机制, 或者表现为化学上的相异性, 偶尔可能表现为错误的数据. 从统计学角度讲, 离域点分为 3 类[109]: 1) X离域点: 物质的分子结构描述符不在其他物质的描述符空间之内; 2) Y离域点, 即实验数据的异常值; 3) X/Y关系离域: 描述符X与环境性质Y的关系方面, 与训

练集中其他物质不同, 即呈现不同的作用机制 [109].

判断模型的离域点, 对精确确定模型的应用域具有重要意义. 但三类离域点中, Y离域点只能根据经验判断, X/Y离域点也不能直接检测, 所以研究重点是判断X离域点, 主要有以下两种方法 [117]: 1) Hotelling's T^2: 是Student's *t*-test的多变量形式 [118]. 2) DModX: 表示化合物在X方向上到模型超平面的距离. 如果该距离大于模型设定的极限值, 则认为所代表的化合物为X离域点 [109]. 这两种诊断方法经常联合使用. 其主要区别在于: Hotelling's T^2 方法来自于可解释的变量信息, 判断结果为强烈离群的数据点, 而DModX方法来自于未解释的变量信息 [117], 所判断的离域点属于中等程度离域. 另外基于回归分析的模型, 也常采用标准残差做为离域点的判断标准.

值得注意的是, 离域点广泛存在于所有的环境指标中, 并对这些指标模型的发展起到了重要的推动作用 [101]. 分析离域点会加强对模型的深入理解, 促进作用机理的认识 [101]. 因此必须基于合理的原则和明确的算法来判断离域点. 可以通过去除离域点前后模型性能的变化进一步判断其性质. 如果离域点仅仅是由统计分析方法引起的, 那么去除后, 模型性能不会有显著提高 [101].

4.5 QSAR 模型的表征

关于QSAR模型的表征, 需要从三方面评价模型的性能 [119]. 首先是拟合效果的统计分析, 以表明模型解释训练集变化的能力; 然后通过交叉验证, 评估模型稳定性以及内部预测能力; 最后采用建立模型时未使用的数据, 进行外部预测能力的评价.

(1) (Q)SAR 模型拟合效果评价

传统使用的统计评价指标主要有以下几个:

1) 决定系数(R^2)/自由度调整后的决定系数(R_{adj}^2): R^2 是判定拟合优度的重要指标. 然而, 如果引入多余的预测变量会导致较低的自由度, 虽然R^2 较高, 但是模型的预测能力较差 [120]. 所以常采用经自由度校正的决定系数R_{adj}^2. 该值越大, 拟合优度越好.

2) 误差平方和(*SSE*): 反映了实测值与预测值之间的偏离, 该值依赖于数据点个数 [119].

3) 表示随机误差分散程度的均方根误差(RMSE)、表示实测值与拟合值之差的平均绝对残差(MAR)以及拟合值的标准误差(SE)/标准偏差(SD), 是衡量模型精确度的常用参数. 这些参数依赖于环境指标数据的范围和分布, 并受离域点的影响 [119].

4) *F*检验: 是对回归模型显著性水平的方差检验方法, 适用于基于MLR方法建立的模型 [119].

上述拟合优度参数常用于模型拟合效果的初步评价, 但不能鉴别模型的拟合不足或过度拟合问题. 所谓拟合不足, 是指模型没有充分揭示出训练集所包含的变量信息, 这样的问题会导致模型的预测能力降低; 过度拟合则是由于拟合了误差信息, 导致模型的拟合度高于环境指标数据和描述符结合的变化性. 后者是QSAR模型建立过程中经常出现的问题, 尤其对于采用非线性建模方法所得到的模型 [121]. 对于此类问题的判断, 需要通过模型的稳定性分析来解决.

(2) QSAR 模型的稳定性分析及内部验证

模型的稳定性分析是与模型拟合不足或过度拟合问题紧密相连的 [119,122]. 按照习惯, 常使用"不稳定性"这个概念, 其含义是模型受训练集中某些个别化合物或化合物子集的影响比较大. 如果化合物的预测值超出模型的置信区间, 就会导致模型不稳定 [119]. 直接对模型的不稳定性进行定量分析的研究比较少. Kolossov和Stanforth[119]从预测变量和预测值两个角度, 提出了模型不稳定性系数(MIC)和模型预测值不稳定性系数(MVIC). 如果MIC和MVIC 值小于 100%, 表明模型稳定, 反之则模型不稳定.

对于模型的不稳定性分析, 更常用的方法是通过内部验证来进行, 因为任何内部验证技术都能一定程度上评价模型的不稳定性. 内部验证技术主要包括以下几类:

1) 去多法(Leave-many-out)[120]: 将初始训练集中的n个数据点平均分成大小为m ($=n/G$) 的G个子集. 然后每次去除m个数据点, 采用剩下的$n-m$个数据点作为训练集重新建模并验证由m个数据点构成的验证集. 经G次计算, 得到交叉验证系数Q^2 来表征模型的稳定性和预测能力 [120]. 一般认为如果 Q^2大于0.5, 模型比较稳定; 大于 0.9, 模型的稳定性非常优秀 [117].

2) 去一法(Leave-one-out): 具体过程与去多法相似, 区别仅在于$m=1$. 统计学理论证明, 在变量选

择方面, 去多法比去一法效果要好 [123,124], 主要是因为去一法以及m值较小的去多法比m值较大的去多法容易包含更多的(潜在)变量信息, 导致模型过拟合, 对验证集的预测能力下降 [124].

3) Bootstrapping法 [120,125]: 从原始数据中随机选择m个数据点, 建模, 并预测其他未被选择的化合物. 重复G次, 得到平均Q^2. 同样, 较高的Q^2 值也表明模型的稳定性.

4) Y的随机性检验 [120]: 这也是一种广泛用于表征模型稳健性的统计方法. 随机调整因变量Y形成新矩阵, 然后采用原来的自变量矩阵建立模型, 重复 50~100 次, 得到基于随机数据模型的R_{adj}^2 和交叉验证系数 Q^2 值. 如果这些值都比较低, 则证明原模型的稳定性比较好, 反之, 表明依目前的建模方法得到的模型不能被接受.

此外, 需要注意的是: 表示模型拟合能力的R_{adj}^2 比表示模型稳定性的交叉验证系数Q^2 值要高, $R_{adj}^2 - Q^2$ 的差值一般为 0.2~0.3, 如果超过 0.3, 表明模型可能存在如下问题: 模型过拟合、存在不相关的X变量或数据中存在离域点 [117].

(3) QSAR 模型预测能力的分析

QSAR 模型的预测能力以模型的拟合优度、稳定性为基础, 并高度依赖于模型的应用域. 评价模型预测能力的最有效方法是进行外部验证, 具体分为以下两类:

1) 外部数据集作为验证集: 采用独立的外部数据集作为验证集, 是最标准的外部验证方法, 不仅可以判断模型对新化合物的预测能力, 也是模型在ERA中得到认可和应用的基本的保证. 但是该方法的实施会受到可获得的外部数据的限制 [117].

2) 原始数据集的子集作为验证集: 鉴于获得外部数据的限制, 可采用将原始数据集的子集作为验证集的替代方法. 将原始数据集分为两个子集, 一个用于建模, 另外一个用于外部验证 [120]. 这种方法的潜在的要求是两个子集应该能分别代表整个数据集的描述符空间并且它们的结构域有一定相似性 [120]. 可应用的方法包括直接随机选择 [120,126]、通过神经网络 [120,127]等系统聚类技术选择、统计实验设计(D-优化)方法 [120,128]选取等.

外部验证的结果, 通过验证集化合物的实测值与预测值之间的交叉验证系数Q^2 和拟合系数R^2 来表示. Q^2与R^2没有相关性, 较高的Q^2值仅仅是模型具有较高预测能力的必要条件, 而非充分条件 [129,130].

以上总结了QSAR模型的拟合优度、稳定性以及预测能力的评价方法, 基于这三方面的模型质量评价, Kolossov和Stanforth[119]提出了一个表征QSAR模型质量的综合指数. 只有高质量的QSAR模型才能为有机化学品的筛选、管理和ERA提供决策依据. 所以, 须进一步发展模型的表征和验证等技术 [119,131,132].

5 (Q)SAR 模型登记和使用准则

为促进(Q)SAR模型的发展和应用, 应该建立登记制, 规范管理现有模型. 目前尚没有统一的登记格式, 根据Cronin等人 [101]的建议, 需要遵循以下几点原则: 1) 列出用于建立(Q)SAR模型的所有环境指标数据, 促进建立透明的模型并降低模型滥用的危险, 也利于进行其他研究; 2) 列出(Q)SAR模型中重要的物理化学描述符; 3) 对化合物结构进行全部描述, 列出IUPAC名、SMILES号或CAS号等相关信息. 所登记的(Q)SAR模型, 必须要有相应的拟合检验、稳定性分析和验证结果. 同时根据管理决策的具体需求, 制定模型不确定性的接受水平(考虑实验数据的变化), 平衡模型预测准确度与应用域之间的关系.

6 结语

在毒害有机物生态风险评价中, (Q)SAR 技术可以在弥补基础数据的缺失、降低昂贵的测试费用、减少动物实验、评估数据的不确定性等方面发挥重要作用. (Q)SAR 技术的发展, 具有很强的目标导向性和应用性、多学科集成性和智能性的特点. 发达国家对该技术进行了系统性的研究, 并在有机化学品监管等领域逐渐应用该技术. 欧盟 REACH 法规中对该技术的应用, 进行了明确的规定. (Q)SAR 技术研究, 应进一步侧重其目标导向, 完善和细化(Q)SAR 在毒害化学品 ERA 中的应用导则, 侧重模型表征方面的研究. 我国在(Q)SAR 技术研究和应用方面的工作, 需要迎头赶上国际先进水平.

参考文献

1 Macleod M, Mckone T E, Foster K L, Maddalena R L, Parkerton T F, Mackay D. Applications of contaminant fate and bioaccumulation models in assessing ecological risks of chemicals: A case study for gasoline hydrocarbons. Environ Sci Technol, 2004, 38(23): 6225—6233[DOI]

2 U. S. Environmental Protection Agency. Guidelines for ecological risk assessment. In: Risk Assessment Forum. Washington: U. S. Environmental Protection Agency, 1998, 63(93): 26846—26924

3 Verhaar H J M, Solbe J, Speksnijder J, Van Leeuwen C J, Hermens J L M. Classifying environmental pollutants: Part 3. External validation of the classification system. Chemosphere, 2000, 40(8): 875—883[DOI]

4 Enterprise & Industry Directorate General and Environment Directorate General. European Commission, REACH in brief. 2002, September. Available online at: http://ecb.jrc.it/REACH/

5 Linkov I, Ames M R, Crouch E A C, Satterstrom F K. Uncertainty in octanol-water partition coefficient: Implications for risk assessment and remedial costs. Environ Sci Technol, 2005, 39(18): 6917—6922[DOI]

6 Tunkel J, Mayo K, Austin C, Hickerson A, Howard P. Practical considerations on the use of predictive models for regulatory purposes. Environ Sci Technol, 2005, 39(7): 2188—2199[DOI]

7 Cronin M T D, Walker J D, Jaworska J S, Comber M H I, Watts C D, Worth A P. Use of QSARs in international decision-making frameworks to predict ecologic effects and environmental fate of chemical substances. Environ Health Persp, 2003, 111(10): 1376—1390

8 Li N Q, Wania F, Lei Y D, Daly G L. A Comprehensive and critical compilation, evaluation, and selection of physical-chemical property data for selected polychlorinated biphenyls. J Phy Chem Ref Data, 2003, 32(4): 1545—1590[DOI]

9 Hammett L P. Some relations between reaction rates and equilibrium constants. Chem Rev, 1935, 17 (1): 125—136[DOI]

10 Hammett L P. The effect of structure upon the reactions of organic compounds. Benzene derivatives. J Am Chem Soc, 1937, 59(1): 96—103[DOI]

11 Taft R M. Polar and steric substituent constants for aliphatic and *o*-benzoate groups from rates of esterification and hydrolysis of esters. J Am Chem Soc, 1952, 74(12): 3120—3128[DOI]

12 Kaliszan R. Quantitative structure-retention relationships applied to reversed-phase high-performance liquid chromatography. J Chromatogr A, 1993, 656(1-2): 417—435[DOI]

13 Goss K -U, Schwarzenbach R P. Linear free energy relationships used to evaluate equilibrium partitioning of organic compounds. Environ Sci Technol, 2001. 35(7): 1—9[DOI]

14 Nguyen T H, Goss K -U, Ball W P. Polyparameter linear free energy relationships for estimating the equilibrium partition of organic compounds between water and the natural organic matter in soils and sediments. Environ Sci Technol, 2005, 39(4): 913—924[DOI]

15 Chen J W, Pei J, Quan X, Zhao Y Z, Chen S. Linear free energy relationships on rate constants for dechlorination by zero-valent iron. SAR QSAR Environ Res, 2002, 13(6): 597—606[DOI]

16 Yan C L, Chen J W, Huang L P, Ding G H, Huang X Y. Linear free energy relationships on rate constants for the gas-phase reactions of hydroxyl radicals with PAHs and PCDD/Fs. Chemosphere, 2005, 61(10): 1523—1528 [DOI]

17 Chen J W, Peijnenburg W J G M, Quan X, Chen S, Zhao Y Z, Yang F L. The use of PLS algorithms and quantum chemical parameters derived from PM3 Hamiltonian in QSPR studies on direct photolysis quantum yields of substituted aromatic halides. Chemosphere, 2000, 40(12): 1319—1326[DOI]

18 Chen J W, Quan X, Schramm K -W, Kettrup A, Yang F L. Quantitative structure-property relationships (QSPRs) on direct photolysis of PCDDs. Chemosphere, 2001, 45(2): 151—159[DOI]

19 Free S M, Wilson J M. A mathematical contribution to structure-activity studies. J Med Chem, 1964, 7(4): 395—399[DOI]

20 Cramer R D, Patterson D E, Bunce J D. Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. J Am Chem Soc, 1988, 110(18): 5959—5967[DOI]

21 Marshall G R, Cramer Ⅲ R D. Three-dimensional structure-activity relationships. Trends Pharmacol Sci, 1988, 9(8): 285—289[DOI]

22 Hansch C, Maloney P P, Fujita T, Muir R M. Correlation of biological activity of phenoxyacetic acids with Hammett substituent constants and partition coefficients. Nature, 1962, 194(4824): 178—180[DOI]

23 Hansch C, Muir R M, Fujita T, Maloney P P, Geiger F, Streich M. The correlation of biological activity of plant growth regulators and

chloromycetin derivatives with Hammett constants and partition coefficients. J Am Chem Soc, 1963, 85(18): 2817—2824[DOI]
24 Fujita T, Iwasa J, Hansch C. A new substituent constant, π, derived from partition coefficients. J Am Chem Soc, 1964, 86(23): 5175—5180[DOI]
25 Hansch C, Clayton J M. Lipophilic character and biological activity of drugs II: the parabolic case. J Pharm Sci, 1973, 62(1): 1—21[DOI]
26 Hermens J, Leeueanch P, Musch A. Quantitative structure-activity relationships and mixture toxicity studies of chloro- and alkylanilines at an acute lethal toxicity level to the guppy (*Poecilia reticulata*). Ecotox Environ Safe, 1984, 8(4): 388—394[DOI]
27 Schultz W T, Bryant S E, Lin D T. Structure-toxicity relationships for tetrahymana: aliphatic aldehydes. B Environ Contam Tox, 1994, 52(2): 279—285
28 Kamlet M J, Abraham M H, Doherty R M, Taft R W. Solubility properties in polymers and biological media. 4. Correlations of octanol/water partition coefficients with solvatochromic parameters. J Am Chem Soc, 1984, 106(2): 464—466[DOI]
29 Kamlet M J, Doherty R M, Abboud J -L M, Abraham M H, Taft R W. Solubility: a new look. Chemtech, 1986, 16(9): 566—576
30 Kamlet M J, Doherty R M, Carr P W, Mackay D, Abraham M H, Taft R W. Linear solvation energy relationships. 44. Parameter estimation rules that allow accurate prediction of octanol/water partition coefficients and other solubility and toxicity properties of polychlorinated biphenyls and polycyclic aromatic hydrocarbons. Environ Sci Technol, 1988, 22(5): 503—509[DOI]
31 Leahy D E. Intrinsic molecular volume as a measure of the cavity term in linear solvation energy relationship: octanol-water partition coefficients and aqueous solubilities. J Pharm Sci, 1986, 75(7): 629—639[DOI]
32 Abraham M H, Ibrahim A, Zissimos A M. Determination of sets of solute descriptors from chromatographic measurements. J Chromatogr A, 2004, 1037(1-2): 29—47[DOI]
33 Wilson L Y, Famini G R. Using theoretical descriptors in quantitative structure-activity relationships: some toxicological indices. J Med Chem, 1991, 34(5): 1668—1674[DOI]
34 Famini G R, Renski 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[DOI]
35 Reta M, Carr P W, Sadek P C, Rutan S C. Comparative study of hydrocarbon, fluorocarbon, and aromatic bonded RP-HPLC stationary phases by linear solvation energy relationships. Anal Chem, 1999, 71(16): 3484—3496 [DOI]
36 Kamlet M J, Doherty R M, Veith G D, Taft R W, Abraham M H. Solubility properties in polymers and biological media. 7. An analysis of toxicant properties that influence inhibition of bioluminescence in *Photobacterium phosphoreum* (the Microtox test). Environ Sci Technol, 1986, 20(7): 690—695[DOI]
37 Kamlet M J, Doherty R M, Abraham M H, Veith G D, Abraham D J, Taft R W. Solubility properties in polymers and biological media. 8. An analysis of the factors that influence toxicities of organic nonelectrolytes to the Golden Orfe Fish (*Leuciscus idus melanotus*). Environ Sci Technol, 1987, 21(2): 149—155[DOI]
38 Balakrishnan A, Polli J E. Apical Sodium Dependent Bile Acid Transporter: A Potential Prodrug Target. Mol Pharmaceutics (Review), 2006, 3(3): 223—230[DOI]
39 Webb S R, Durst G L, Pernich D, Hall J. C. Interaction of cyclohexanediones with acetyl coenzyme-a carboxylase and an artificial target-site antibody mimic: a Comparative molecular field analysis. J Agric Food Chem, 2000, 48(6): 2506—2511 [DOI]
40 Yu S J, Keenan S M, Tong W, Welsh W J. Influence of the structural diversity of data sets on the statistical quality of three-dimensional quantitative structure-activity relationship (3D-QSAR) models: predicting the estrogenic activity of xenoestrogens. Chem Res Toxicol, 2002, 15(10): 1229—1234 [DOI]
41 Tong W, Lowis D R, Perkins R, Chen Y, Welsh W J, Goddette D W, Heritage T W, Sheehan D M. Evaluation of quantitative structure-activity relationship methods for large-scale prediction of chemicals binding to the estrogen receptor. J Chem Inf Comput Sci, 1998, 38(4): 669—677 [DOI]
42 Chen J W, Quan X, Zhao Y Z, Yan Y L, Yang F L. Quantitative structure-property relationship studies on *n*-octanol/water partitioning coefficients of PCDD/Fs. Chemosphere, 2001, 44(6): 1369—1374[DOI]
43 Pavan M, Worth A P, Netzeva T I. Review of QSAR Models for Bioconcentration. JRC report EUR EN I-21020. 2006
44 Chen J W, Harner T, Ding G H, Quan X, Schramm K W, Kettrup A. Universal predictive models on octanol-air partition coefficients at different temperatures for persistent organic pollutants. Environ Toxicol Chem, 2004, 23(10): 2309—2317[DOI]
45 Li X H, Chen J W, Zhang L, Qiao X L, Huang L P. The Fragment constant method for predicting octanol-air partition coefficients of persistent organic pollutants at different temperatures. J Phys Chem Ref Data, 2006, 35(3): 1365—1384[DOI]
46 Meylan W M, Howard P H, Boethling R S. Molecular topology/fragment contribution method for predicting soil sorption coefficients.

Environ Sci Technol, 1992, 26(8): 1560—1567[DOI]
47 Schüürmann G, Ebert R -U, Kühne R. Prediction of the sorption of organic compounds into soil organic matter from molecular structure. Environ Sci Technol, 2006, 40(22): 7005—7011[DOI]
48 Hermens J L M, Leeuwangh P, Musch A. Quantitative structure——activity relationships and mixture toxicity studies of chloro- and alkylanilines at an acute lethal toxicity level to the guppy (*Poecilia reticulata*). Ecotoxicol Environ Safe, 1984, 8: 388—394[DOI]
49 Bradbury S P, Russom C L, Ankley G T, Schultz T W, Walker J D. Overview of data and conceptual approaches for derivation of quantitative structure-activity relationships for ecotoxicological effects of organic chemicals. Environ Toxicol Chem, 2003, 22(8): 1789—1798[DOI]
50 Tong W, Fang H, Hong H, Xie Q, Perkins R, Anson1 J, Sheehan D M. Regulatory application of SAR/QSAR for priority setting of endocrine disruptors: A perspective. Pure Appl Chem, 2003, 75: 2375—2388[DOI]
51 Asikainen A, Ruuskanen J, Tuppurainen K. Consensus kNN QSAR: A versatile method for predicting the estrogenic activity of organic compounds in silico. A comparative study with five estrogen receptors and a large, diverse set of ligands. Environ Sci Technol, 2004, 38(24): 6724—6729[DOI]
52 Liu H X, Papa E, Gramatica P. QSAR prediction of estrogen activity for a large set of diverse chemicals under the guidance of OECD principles. Chem Res Toxicol, 2006, 19(11): 1540—1548[DOI]
53 Raymond J W, Rogers T N, Shonnard D R, Kline A A. A review of structure-based biodegradation estimation methods. J Hazard Mater, 2001, 84(2-3): 189—215[DOI]
54 Chen J W, Peijnenburg W J G M, Quan X, Chen S, Martens D, Schramm K W, Kettrup A. Is it possible to develop a QSPR model for direct photolysis half-lives of PAHs under irradiation of sunlight?. Environ Pollut, 2001, 114(1): 137—143[DOI]
55 Walker J D. International workshops on QSARs in the environmental sciences——The first 20 years. QSAR Comb Sci, 2003, 22(4): 415—421[DOI]
56 Nys G G, Rekker R F. Statistical analysis of a series of partition coefficients with special reference to the predictability of folding of drug molecules. The introduction of hydrophobic fragmental constants (f values). Eur J Med Chem, 1973, 8: 521—535
57 Taft R W, Lewis I C. The general applicability of a fixed scale of inductive effects. Ⅱ. Inductive effects of dipolar substituents in the reactivities of *m*- and *p*-substituted derivatives of benzene. J Am Chem Soc, 1958, 80(10): 2436—2443[DOI]
58 Hansch C, Leo A, Taft R W. A survey of Hammett substituent constants and resonance and field parameters. Chem Rev, 1991, 91(2): 165—195[DOI]
59 Hancock C K, Meyers E A, Yager B J. Quantitative separation of hyperconjugation effects from steric substituent constants. J Am Chem Soc, 1961, 83(20): 4211—4213[DOI]
60 Charton M. The nature of the ortho effect. Ⅱ. Composition of the Taft steric parameters. J Am Chem Soc, 1969, 91(3): 615—618[DOI]
61 Ghose A K, Crippen G M. Atomic physicochemical parameters for three-dimensional-structure-directed quantitative structure-activity relationships. 2. Modeling dispersive and hydrophobic interactions. J Chem Inf Comput Sci, 1987, 27(1): 21—35
62 Kamlet M J, Taft R W. The solvatochromic comparison method. Ⅰ. The beta-scale of solvent hydrogen-bond acceptor (*HBA*) basicities. J Am Chem Soc, 1976, 98(2): 377—383[DOI]
63 Taft R W, Kamlet M J. The solvatochromic comparison method. 2. The alpha-scale of solvent hydrogen-bond donor (*HBD*) acidities. J Am Chem Soc, 1976, 98(10): 2886—2894[DOI]
64 Balaban A T. Using real numbers as vertex invariants for third-generation topological indexes. J Chem Inf Comput Sci, 1992, 32(1): 23—28[DOI]
65 Kier L B, Hall L H. The nature of structure-activity relationships and their relation to molecular connectivity. Eur J Med Chem, 1977, 12: 307—312
66 Karelson M, Lobanov V S, Katritzky A R. Quantum-chemical descriptors in QSAR/QSPR studies. Chem Rev, 1996, 96(3): 1027—1043[DOI]
67 Todeschini R, Consonni V. Handbook of Molecular Descriptors. Weinheim: Wiley-VCH, 2000
68 Ren R E, Wang H W. Multivariate Statistical Analysis-Theory, Method, Case (in Chinese). Beijing: National Defence Industry Press, 1999
69 Livingstone D J, Salt D W. Judging the significance of multiple linear regression models. J Med Chem, 2005, 48(3): 661—663[DOI]
70 Dudek A Z, Arodz T, Galvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): A review. Comb Chem High T Scr, 2006, 9(3): 213—228

71 Xu L, Shao X G. Methods of Chemometrics (in Chinese). Beijing: Science Press, 2004
72 Guha R, Jurs P C. Determining the validity of a QSAR Model-a classification approach. J Chem Inf Model, 2005, 45(1): 65—73[DOI]
73 Barnard J M, Downs G M. Clustering of chemical structures on the basis of two-dimensional similarity measures. J Chem Inf Comput Sci, 1992, 32(6): 644—649[DOI]
74 Wang H W. Partial Least-Squares Regression-Method and Applications (in Chinese). Beijing: Defense Industry Press, 1999
75 Vapnik V. An overview of statistical learning theory. IEEE T Neural Networ, 1999, 10(5): 988—999[DOI]
76 Kövesdi I, Dominguez -Rodriguez M F, Ôrfi L, Naray-Szabo G, Varro A, Papp J G, Matyus P. Application of neural networks in structure-activity relationships. Med Res Rev, 1999, 19(3): 249—269[DOI]
77 Luan F. Application of Support Vector Machines (SVM) and Radial Basis Function Neural Networks (RBFNN) in Chemistry, Environmental Chemistry and Medicinal Chemistry (in Chinese). Doctoral Dissertation. Lanzhou: Lanzhou University, 2006
78 Yang S, Lu W, Chen N. Support vector regression based QSPR for the prediction of some physicochemical properties of alkyl benzenes. J Mol Struct, 2005, 719(1-3): 119—127
79 O'Hara-Mays P. Genetic Algorithms in Molecular Modeling. Edited by James Devillers. Principles of QSAR and Drug Design, Vol. 1. New York: Academic Press, Harcourt Brace & Company. 1996. 1—327
80 Leardi R. Genetic algorithms in chemometrics and chemistry: a review. J Chemometr, 2001, 15(7): 559—569[DOI]
81 Liu H X, Zhang R S, Yao X J, Liu M C, Hu Z D, Fan B T. Prediction of the isoelectric point of an amino acid based on GA-PLS and SVMs. J Chem Inf Comput Sci, 2004, 44 (1): 161—167[DOI]
82 Wanchana S, Yamashita F, Hashida M. QSAR analysis of the inhibition of recombinant CYP 3A4 activity by structurally diverse compounds using a genetic algorithm-combined partial least squares method. Pharm Res, 2003, 20(9): 1401—1408[DOI]
83 Liu J J, Cutler G, Li W, Pan Z, Peng S, Hoey T, Chen L, Ling X B. Multiclass cancer classification and biomarker discovery using GA-based algorithms. Bioinformatics, 2005, 21(11): 2691—2697[DOI]
84 McInerney M, Dhawan, A P. Use of genetic algorithms with back propagation in training of feed-forward neural networks. In: IEEE International Conference on Neural Networks, 1993. 203—208
85 Wang H, Yu J. Application study on nonlinear dynamic FIR modeling using hybrid SVM-PLS method. In: Proceedings of the World Congress on Intelligent Control and Automation (WCICA) 4, 2004. 3479—3482
86 Jaworska J S, Comber M, Auer C, Van Leeuwen C J. Summary of a workshop on regulatory acceptance of QSARs for human health and environmental endpoints. Environ Health Persp, 2003, 111(10): 1358—1360
87 Cronin M T D, Jaworska J S, Walker J D, Comber M H I, Watts C D, Worth A P. Use of QSARs in international decision-making frameworks to predict health effects of chemical substances. Environ Health Persp, 2003, 111(10): 1391—1401
88 Walker J W L, Carlsen E, Simon-Hettich B. Global government applications of analogues, SARs and QSARs to predict aquatic toxicity, chemical or physical properties, environmental fate parameters and health effects of organic chemicals. SAR QSAR Environ Res, 2002, 13(6): 607—616[DOI]
89 Worth A P, Bassan A, De Bruijn J, Saliner A G, Netzeva T, Patlewicz G, Pavan M, Tsakovska I, Eisenreich S. The role of the European Chemicals Bureau in promoting the regulatory use of QSARs methods. SAR QSAR Environ Res, 2007, 18(1-2): 111—125[DOI]
90 Organisation for Economic Co-Operation and Development (OECD). Report from the Expert Group on (Q)SARs on the Principles for the Validation of (Q)SARs, 2004. Available online at: http://appli1.oecd. org/olis/2004doc.nsf/linkto/env-jm-mono(2004)24
91 Organisation for Economic Co-Operation and Development (OECD). Guidance document on the validation of (Quantitative) Structure-Activity Relationships [(Q)SARs] models, 2007. Available online at: http://www.oecd.org/dataoecd/55/22/ 38131728.pdf
92 Organisation for Economic Co-Operation and Development (OECD). Testing and assessment Report on the regulatory uses and applications in OECD member countries of (Quantitative) Structure-Activity Relationship[(Q)SARs] models in the assessment of new and existing chemicals, 2006. Available online at: http://appli1.oecd.org/olis/ 2006doc.nsf/linkto/env-jm-mono(2006)25
93 Wang L S, Han S K. Quantitative Structure-Activity Relationships of Organic Compounds (in Chinese). Beijing: China Environmental Science Press, 1993
94 Wang L S. Chemistry of Organic Pollution (in Chinese). Beijing: Higher Education Press, 2004
95 Chen J W. Quantitative Structure-Property Relationships and Quantitative Structure-Activity Relationships of Organic Pollutants (in Chinese). Dalian: Dalian University of Technology Press, 1999
96 Ding G H. Application of PLS and GA on QSAR of Selected Organic Pollutants (in Chinese). Doctoral Dissertation. Dalian: Dalian University of Technology, 2006

97 Lv Q Z, Shen G L, Yu R Q. Genetic training of network using chaos concept: application to QSAR studies of vibration modes of tetrahedral halides. J Comput Chem, 2002, 23(14): 1357—1365[DOI]

98 Zhao C Y. Applications of QSAR in Life Analytical Chemistry and Environmental Chemistry (in Chinese). Doctoral Dissertation. Lanzhou: Lanzhou University, 2003

99 Yao Y Y, Xu L, Yang Y Q, Yuan X S. Study on structure-activity relationships of organic compounds: Three new topological indices and their applications. J Chem Inf Comput Sci, 1993, 33(4): 590—594[DOI]

100 Lu G H, Yuan X, Zhao Y H. QSAR study on the toxicity of substituted benzenes to the algae(*scenedesmus obliquus*). Chemsphere, 2001, 44(3): 437—440[DOI]

101 Cronin M T D, Schultz T W. Pitfalls in QSAR. J Mol Struct, 2003, 622(1-2): 39—51

102 Schultz T W, Cronin M T D. Essential and desirable characteristics of ecotoxicity quantitative structure-activity relationships. Environ Toxicol Chem, 2003, 22(3): 599—607[DOI]

103 Cronin M T D, Schultz T W. Validation of *Vibrio fischeri* acute toxicity data: mechanism of action-based QSARs for nonpolar narcotics and polar narcotic phenols. Sci Total Environ, 1997, 204(1): 75—88[DOI]

104 Walker J D, Jaworska J, Comber M H I, Schultz T W, Dearden J C. Guidelines for developing and using quantitative structure-activity relationships. Environ Toxicol Chem, 2003, 22(8): 1653—1665[DOI]

105 Livingstone D J. Data Analysis for Chemists: Applications to QSAR and Chemical Product Design. Oxford: Oxford University Press, 1995

106 Cronin M T D, Schultz T W. Development of quantitative structure-activity relationships for the toxicity of aromatic compounds to *Tetrahymena pyriformis*: comparative assessment of methodologies. Chem Res Toxicol, 2001, 14(9): 1284—1295[DOI]

107 Burden F R, Winkler D A. A quantitative structure-activity relationships model for the acute toxicity of substituted benzenes to *Tetrahymena pyriformis* using Bayesian-regularized neural networks. Chem Res Toxicol, 2000, 13(6): 436—440[DOI]

108 Kholodovych V, Smith J R, Knight D, Abramson S, Kohn J, Welsh W J. Accurate predictions of cellular response using QSPR: a feasibility test of rational design of polymeric biomaterials. Polymer, 2004, 45(22): 7367—7379[DOI]

109 Furusjö E, Svenson A, Rahmberg M, Andersson M. The importance of outlier detection and training set slection for reliable environmental QSAR prediction. Chemosphere, 2006, 63(1): 99—108[DOI]

110 Dimitrov S, Dimitrova G, Pavlov T, Dimitrova N, Patlewicz G, Niemela J, Mekenyan O. A stepwise approach for defining the applicability domain of SAR and QSAR models. J Chem Inf Model, 2005, 45(4): 839—849 [DOI]

111 EC (European Commission). Technical Guidance Document on Risk Assessment in support of Commission Directive 93/67/EEC on Risk Assessment for new notified substances and Commission Regulation (EC) No 1488/94 on Risk Assessment for existing substances, and Directive 98/8/EC of the European Parliament and of the Council concerning the placing of bio-cidal products on the market, Parts 3. 2003

112 Jaworska J S, Nikolova -Jeliazkova N. Aldenberg T. QSAR applicability domain estimation by projection of the training set in descriptor space: A review. Atla-Altern Lab Anim, 2005, 33(5): 445—459

113 Netzeva T I, Saliner A G, Worth A P. Comparison of the applicability domain of a quantitative structure-activity relationship for estrogenicity with a large chemical inventory. Environ Toxicol Chem, 2006, 25(5): 1223—1230[DOI]

114 Sheridan R P, Feuston B P, Maiorov V N, Kearsley S K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. J Chem Inf Comput Sci, 2004, 44(6): 1912—1928[DOI]

115 Dimitrov S, Koleva Y, Schiltz T W, Walker J D, Mekenyan O. Interspecies quantitative structure-activity relationships model for aldehydes: Aquatic toxicity. Environ Toxicol Chem, 2004, 23(2): 463—470[DOI]

116 Schultz T W, Hewitt M, Netzeva T I, Cronin M T D. Assessing applicability domains of toxicological QSARs: definition, confidence in predicted values, and the role of mechanisms of Action. QSAR Comb Sci, 2007, 26(2): 238—254[DOI]

117 Eriksson L, Jaworska J, Worth A P, Cronin M T D, McDowell R M, Gramatica P. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. Environ Health Persp, 2003, 111(10): 1361—1375

118 Jackson J E. A User's Guide to Principal Components. New York: John Wiley. 1991

119 Kolossov E, Stanforth R. The quality of QSAR models: problems and solutions. SAR QSAR Environ Res, 2007, 18(1-2): 89—100 [DOI]

120 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. QSAR Comb Sci, 2003, 22(1): 69—77[DOI]

121 Livingstone D J, Manallack D T, Tetko I V. Data modeling with neural networks: advantages and limitations. J Comput Aid Mol Des, 1997, 11(2): 135—142[DOI]
122 Hawkins D M. The problem of overfitting. J Chem Inf Comput Sci. 2004, 44(1): 1—12[DOI]
123 Zhang P. Model selection via multifold cross validation. Ann Statist, 1993, 21: 299—313[DOI]
124 Baumann K, Korff M, Albert H. A systematic evaluation of the benefits and hazards of variable selection in latent variable regression. Part Ⅱ. Practical applications. J Chemometr, 2002, 16(7): 351—360[DOI]
125 Wehrens R, Putter H, Buydens L M C. The bootstrap: a tutorial. Chemom Intell Lab Systems, 2000(1), 54: 35—52[DOI]
126 Yasri A, Hartsough D. Toward an optimal procedure for variable selection and QSAR model building. J Chem Inf Comput Sci, 2001, 41(5): 1218—1227[DOI]
127 Burden F R, Ford M G, Whitley D C, Winkler D A. Use of automatic relevance determination in QSAR studies using bayesian neural networks. J Chem Inf Comput Sci, 2000, 40(6): 1423—1430[DOI]
128 Mitchell T J. An algorithm for the construction of "D-optimal" experimental design. Technometrics, 2000, 42(1): 48—54[DOI]
129 Kubinyi H, Hamprecht F A, Mietzner T. Three-dimensional quantitative similarity-activity relationships (3D QSAR) from SEAL similarity matrices. J Med Chem, 1998, 41(14): 2553—2564[DOI]
130 Golbraikh A, Tropsha A. Beware of q^2!. J Mol Graph Model, 2002, 20(4): 269—276[DOI]
131 Schultz T W, Netzeva T I, Cronin M T D. Evaluation of QSARs for ecotoxicity: A method for assigning quality and confidence. SAR QSAR Environ Res, 2004, 15(5-6): 385—397[DOI]
132 Deardon J C, Roberts D W. Larger molecules penetrate membranes more readily. J Pharm Pharmacol, 2006, 58: 60