

QSAR模型应用域的表征方法

王中钰, 陈景文*, 傅志强 and 李雪花

Citation: [科学通报](#); doi: 10.1360/TB-2021-0406

View online: <https://engine.scichina.com/doi/10.1360/TB-2021-0406>

Published by the [《中国科学》杂志社](#)

Articles you may be interested in

[参数区域化在全球水文模型FLEX-Global中的应用](#)

SCIENTIA SINICA Terrae **51**, 805 (2021);

[表征孔隙孔径分布的岩土体孔隙率模型及其应用](#)

Chinese Science Bulletin **55**, 2761 (2010);

[大气氧化性定量表征方法的建立及其应用](#)

Science in China Series B-Chemistry (in Chinese) **38**, 938 (2008);

[基于sEMG振子模型的骨骼肌等长收缩力与固有特性的能量核表征方法](#)

Chinese Science Bulletin **59**, 561 (2014);

[藻胆体模型复合物的合成及其表征](#)

Chinese Science Bulletin **42**, 1288 (1997);



QSAR模型应用域的特征方法

王中钰, 陈景文*, 傅志强, 李雪花

大连理工大学环境学院, 工业生态与环境工程教育部重点实验室, 大连市化学品风险防控及污染防治技术重点实验室, 大连 116024

* 联系人, E-mail: jwchen@dlut.edu.cn

2021-04-27 收稿, 2021-06-11 修回, 2021-06-11 接受, 2021-06-28 网络版发表

国家重点研究发展计划(2018YFC1801604, 2018YFE0110700)和国家自然科学基金(21661142001)资助

摘要 定量构效关系(QSAR)模型是填补化学品环境安全数据空缺的重要工具. QSAR模型需要明确定义的应用域, 才能合理地用于化学品管理. 本文回顾了应用域的3种概念: 描述符域、结构域和机理域. 基于案例, 重点介绍了基于分子指纹与相似性度量指标而计算结构域的方法、结构域的特点和优势. 讨论了结构-活性地貌(structure-activity landscape)中呈现的活性悬崖(activity cliffs)现象及其成因. 为了更好地理解描述符的适用性, 解释QSAR机制及合理选择应用域的特征方法, 有必要认识预测终点(endpoint)本质上所描述的系统, 该系统复杂性和空间异质性, 以及预测终点是否考察了系统行为的涌现.

关键词 定量构效关系(QSAR), 应用域, 描述符, 活性悬崖, 结构-活性地貌

在环境科学与工程领域, 定量构效关系(QSAR)指分子(或其聚集体, 如纳米粒子)的微观结构, 与其物理化学性质、环境暴露行为、工程设施中的行为、生物体内毒代动力学行为, 以及毒理学效应参数等终点(endpoints)之间的定量关系. 这里, 终点或预测终点泛指可以测定或建模的与化学结构有关的物理化学性质、生物学效应或环境行为参数, 在本文中统一泛称为“活性”. 近年来, 随着人工智能、机器学习和大数据分析技术的快速发展, 计算毒理学拓展了QSAR的外延. 广义上, 基于分子结构并结合其他类型的预测变量, 预测化学物质行为或效应的模型也属于QSAR. 例如, 结合化学结构与体外(*in vitro*)测试终点数据, 预测活体(*in vivo*)测试终点数据的定量结构-*in vitro*-*in vivo*外推模型^[1]. QSAR是对化学品进行风险预测和管理的必要工具^[2], 是计算毒理学^[3]研究的核心内容之一.

1 面向化学品管理的QSAR

化学品促进了社会发展, 极大地改善了人类生活

质量, 但也是新污染物的重要来源, 是影响人体与生态健康的重大风险源, 是人类社会可持续发展的重大挑战^[4]. 根据联合国环境规划署2019年发布的《全球化学品展望(第二版)》, “截至2020年, 最大限度地减少化学品和废物不利影响”的全球目标无法兑现, 需要加强科技对化学品管理的支撑作用, 填补关于化学品的数据空缺和未知领域(<https://www.unenvironment.org/resources/report/global-chemicals-outlook-ii-legacies-innovative-solutions>).

健全化学品管理的核心是化学品风险或安全性评价, 关键在于相关化学物质暴露与危害性数据的获取. 描述化学品环境分配、转化、毒代动力学及毒性效应的关键参数, 构成了支撑化学品风险评价的指标体系^[5]. 围绕这些指标体系, 欧美发达国家及国际组织发展了相应的标准测试方法, 如经济发展与合作组织的系列性化学品测试导则^[6]. 随着测试工作的展开, 化学品监管机构掌握了部分大宗贸易化学品的安全性数据, 也认识到一些测试(尤其毒性测试)项目周期长、物资

引用格式: 王中钰, 陈景文, 傅志强, 等. QSAR模型应用域的特征方法. 科学通报, 2021, 66

Wang Z Y, Chen J W, Fu Z Q, et al. Characterization of applicability domains for QSAR models (in Chinese). Chin Sci Bull, 2021, 66, doi: [10.1360/TB-2021-0406](https://doi.org/10.1360/TB-2021-0406)

(化学试剂、实验动物)消耗大、运行维护费用高. 2020年, 全球市场注册的化学品及其混合物超过35万种^[7]. 考虑化学品在复杂环境-生命系统中的转化产物, 实际暴露的化学污染物种类更多. 另一方面, 化学品产业也亟须开发更安全的分子和材料, 以替代法规禁止或限制使用的化学品^[8,9]. 这要求研发者在探索近乎无穷的化学空间时, 兼顾化学分子的工作性能和安全性^[10,11]. 单纯采用实验测试来获取全部化学品安全性数据, 会给监管机构或相关企业带来难以承受的经济压力, 低效的数据获取进程则妨碍健全化学品管理以及可持续发展^[12,13].

QSAR在填补相关的数据空缺方面, 具有不可替代的作用. QSAR模型基于化合物结构与活性均已知的数据集(训练集), 寻找表征分子结构的数值特征(描述符)与预测终点数据之间的规律, 并利用所揭示的规律, 实现从分子结构到定量(回归)或定性(分类)终点数据的预测^[14,15]. 近年来, 人工智能、机器学习、大数据分析技术飞速发展, 提供了强大的QSAR建模算法和工具^[16]. 然而, QSAR本质上仍是数据驱动模型, 训练集所呈现的规律未必适用于任意结构的化合物.

2004年, 经济发展合作组织(OECD)确立了面向风险评价的QSAR应该符合的5条标准: (1) 具有明确定义的终点; (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) (如果可能)进行机理解释(<https://read.oecd.org/10.1787/9789264085442-en?format=pdf>). 其中第3条恰恰体现出QSAR模型对训练数据的依赖性. QSAR模型的可靠性总是局限于特定的范围, 即应用域内^[5], 类似于机器学习领域所谈的模型“泛化”能力^[17].

2 应用域的特征

应用域的功能是界定能够被QSAR模型可靠预测的化合物, 也可理解为模型适用化合物的集合^[2]. 化学品风险管控中所用到的QSAR模型, 不论是定量回归模型还是定性分类模型, 也不论训练集所包含的是结构相似的系列物还是结构多样的分子, 其中化合物的结构特征一般由描述符表示^[18]. 因此, 训练集化合物所覆盖的描述符空间可以作为应用域, 亦称描述符域^[19].

2.1 描述符域

早期QSAR建模多利用实验测定数据(如正辛醇/水分配系数(log K_{OW})、谱学数据、摩尔折射率)或其折算

值(如取代基电子效应(σ)和立体效应(E_s))作为描述符^[20-22]. 然而, 实验测定需要足量的化学标样, 实验流程与分析方法也受到各类限制与干扰. 相比而言, 利用程序计算分子描述符更具优势. 分子结构可以由不同的理论模型表示, 如结构式、原子三维排布以及电子结构等, 得到的描述符也各有特色.

结构式呈现了原子类型、分子拓扑(键连关系)等信息. 基于结构式可以推导相对分子质量(M_w), 估算分子体积. 分子拓扑对应于分子图(顶点和边分别表示原子和共价键), 其字符串格式(如简化分子线性输入系统(SMILES)和国际化学品标识符(InChI)^[23])常用于数据存储与检索, 其毗邻矩阵表示则衍生出各类拓扑指数和原子对描述符^[18,24]. 根据预设的(或某预设规则所能发现的)子结构或碎片模式的存在与否或出现频次, 可构造布尔或整型向量来表示化合物的分子指纹^[25]. 近年来, 图卷积神经网络有望实现从结构式到活性“端到端”的预测, 尽管描述符被隐含在算法内部, 但本质上仍基于分子拓扑结构进行数学变换^[26].

原子三维排布给出了分子的空间构象, 可以借助距离矩阵^[27]、分子形状分析^[28]等方法计算各种三维描述符. 此外, 比较分子力场(CoMFA)方法基于分子形状或药效团, 首先将分子与参照构象对齐, 进而考虑静电和范德瓦尔斯(van der Waals)等相互作用在三维格点上的差异, 则得到由格点数据展开的高维描述符^[29]. 在原子三维排布的基础上, 量子化学计算得到的电子结构模型则呈现了电子密度、轨道排布、能量等信息, 可提取出前线分子轨道能、原子部分电荷、偶极矩、分子极化率等描述符.

在实践中, n 种训练集化合物的 m 种描述符数据可以整理为一个 n 行 m 列的表格或矩阵 X . 每一行所表示的化合物 i , 都可以视为在 m 维描述符空间中的一个点, 对应于向量 $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,m})$. 考察每种描述符的取值范围, 即构造 m 维空间中的超矩形. 如果查询化合物所有描述符的值都落在对应训练集描述符的范围(或其他指定区间)内, 化合物也就位于超矩形应用域之内. 图1展示了美国Tox21项目^[30]8068种化合物的 M_w 和Wildman-Crippen log P 参数($SLogP$, 一种主要描述分子疏水性的参数)^[31]所张成的二维描述符空间. 其中, 图1(a)的虚线勾勒出了超矩形所涵盖的区域. 超矩形假设每个描述符的数据分布是均一的, 且不考虑描述符空间内部的空白以及不同描述符之间的相关性.

另一种评估描述符空间覆盖性的方法是凸包^[32].

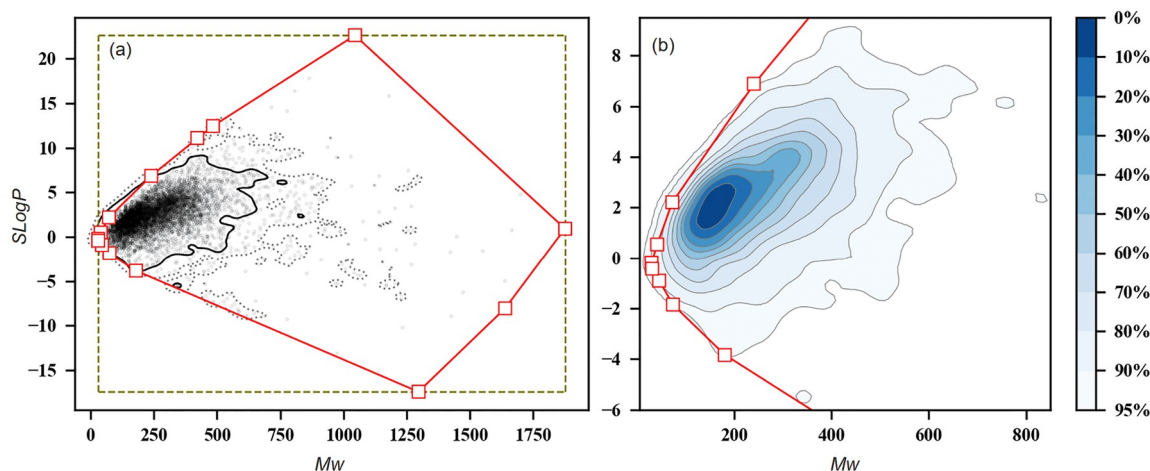


图1 (网络版彩色)超矩形、凸包(a)和Gaussian核概率密度(b)的描述符域。\$Mw\$: 相对分子质量, \$SLogP\$: Wildman-Crippen \$\log P\$

Figure 1 (Color online) Descriptor domain of hyper-rectangle, convex hull (a) and Gaussian kernel probability density estimation (b). \$Mw\$: Relative molecular mass, \$SLogP\$: Wildman-Crippen \$\log P\$

界定凸包的关键, 在于找到训练集在高维空间中最外围的点(图1, 方形点)。一些开源计算程序(如SciPy^[33])集成了界定凸包的算法, 在描述符空间维度很低时(\$m < 4\$)执行效率较高。凸包仅描绘数据在高维空间中的边界。位于凸包内部的点可被视为在应用域内。

利用核密度估计(KDE)^[34,35]方法, 构造训练集化合物在描述符空间中的非参数联合概率密度分布, 被认为是最为严格的描述符域界定方法^[36,37]。图1(b)展示了基于Gaussian核函数构造的联合概率核密度估计分布, 从中抽取足量的随机样本, 并以概率密度5%或10%(对应图1(b)中覆盖率95%或90%)分位点为阈值。当任意化合物的联合概率密度大于该阈值时, 则认为化合物在应用域内。然而, KDE算法在描述符空间维度较高时, 效率较低^[37]。从图1可知, 联合概率密度应用域所涵盖的描述符空间远小于超矩形与凸包(图1(a), 实线, 95%覆盖率)。

此外, 为了压缩描述符空间中的空白, 或应对凸包与KDE方法^[34]不适应高维空间的问题, 常结合数据降维(如主成分分析)技术, 以将原始描述符空间转换为彼此相互正交的主成分空间。并基于少数重要的主成分构造超矩形、凸包或联合概率密度分布类型的应用域。

除了超矩形、凸包或概率密度分布法, 化合物*i*与(对应于点 $\mathbf{x}_i$ 和 $\mathbf{x}$)在描述符空间中的距离, 也可用于判定描述符域边界。最常用的距离方法为欧式(Euclidean)距离:

$$D_E(i, j) = \sqrt{\sum_{k=1}^m w_k (x_{i,k} - x_{j,k})^2}, \quad (1)$$

式中, $x_{i,k}$ 与 $x_{j,k}$ 分别是化合物*i*和*j*的第*k*个描述符的值; w_k 为第*k*个描述符的权重, 一般取1。若描述符大致服从均匀分布, 则可使用街区距离:

$$D_C(i, j) = \sum_{k=1}^m |x_{i,k} - x_{j,k}|. \quad (2)$$

假定点*i*和*j*来自与训练集相同的分布, 也可计算马氏(Mahalanobis)距离:

$$D_M(i, j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{x}_j)}, \quad (3)$$

式中, $\mathbf{V}$ 为描述符矩阵 $\mathbf{X}$ 的协方差矩阵。在实践中, 通常会采用各个描述符平均值对应的点 $\boldsymbol{\mu}$ 代表整个训练集。待查询化合物($\mathbf{x}$)与均值点 $\boldsymbol{\mu}$ 的距离小于某个预设值时, 即认为其在应用域内^[37]。

马氏距离考虑了描述符之间的相关性, 在数值上与多元线性统计学中的杠杆值或帽子值相关。对于有*n*个化合物*m*个描述符的训练集, 可以构造*n*行(1+*m*)列的矩阵 $\mathbf{X}_+$ 。 $\mathbf{X}_+$ 比 $\mathbf{X}$ 多出来的一列表征线性回归模型中的截距项。化合物*i*的杠杆值或帽子值(h_i)为

$$h_i = \mathbf{x}_i (\mathbf{X}_+^T \mathbf{X}_+)^{-1} \mathbf{x}_i^T. \quad (4)$$

杠杆值的阈值(h^*)一般设定为 $3(1+m)/n$ 。如果某化合物的 h_i 比 h^* 大, 该化合物则被视为在应用域之外。QSAR研究领域往往直接使用 $\mathbf{X}$ (而非 $\mathbf{X}_+$)来计算杠杆值^[36,37]。注意到, 若原始描述符中的一些描述符彼此是

线性相关的, 则构成奇异阵, 从而无法求解出杠杆值或马氏距离.

上述描述符域的定义并非总是有效. 对于布尔向量型分子指纹转化而来的 m 维描述符, 由于每一维的描述符值($x_{i,k}$)非0即1, 其对应的超矩形是 m 维的单位超立方体. 且所有化合物点都不会位于单位超立方体之外. 从这个角度看, 超矩形、凸包与联合概率密度所定义的描述符域丧失了意义. 类似地, 在布尔指纹空间里, 各维度取平均后对应的点 μ 也难以直观地理解, 因此计算查询点与平均值点的距离似乎缺乏意义. QSAR模型的结构域, 有助于解决这些问题.

2.2 结构域

早期的QSAR研究, 主要围绕具有共同母体结构的系列化合物活性而展开^[38]. 这些研究仅仅适用于具有该母体结构的系列物, 因而相应的模型也被称为局域QSAR模型. 局域QSAR模型一般不适用于其他母体结构的系列物. 因此, 在考察待查询化合物与描述符域的关系之前, 已经隐含了对结构相似性的考虑.

2003年, Eriksson等人^[39]提出QSAR模型有效性的条件之一是数据均质性, 要求所研究的化合物处于相似的状态, 化合物数据集不应该有多组聚类或强影响的离群点. 在这些假设之下, 发展出了上文所介绍的描述符域表征方法. 近10年来, 众多公开可获取的数据库资源, 为QSAR建模提供了新的土壤, 大数据与人工智能技术在QSAR领域展露锋芒^[40-42]. 新的研究多基于化合物种类多、结构多样的大数据集构建全局QSAR模型, 实际上无视或打破了数据均质性假设.

尽管如此, 根据“具有相似结构的化合物往往具有相似活性”的相似性原理, 总是可以从结构相似性角度来表征QSAR模型的应用域, 即为结构域. 但是, 结构相似的判定并无绝对的标准, 而取决于具体的相似性度量方法^[43]. 例如, 距离(D)与相似度(S)间存在负相关关系(如 $S=1/(D+1)$), 查询点与训练集中任意化合物的距离提供了两个化合物相似性的度量. 在化学信息学领域, 基于分子指纹与相似性度量指标计算得到的指纹特异性的相似度值则更为常用. 度量化合物 u 和 v 的相似度, 常用Tanimoto系数^[44,45]:

$$Tc(u, v) = c / (a + b - c), \quad (5)$$

式中, a 、 b 分别为两化合物指纹各自具有的子结构或碎片模式的数目, c 为两化合物共享的结构模式的数目.

操作上, 可以考察待查询化合物与每种训练集化合物的相似性, 并给出一套判断规则, 例如 $Tc(u, v)$ 大于某相似度阈值(S_{cutoff})时, 认为待查询化合物落在结构域内^[46,47].

此外, 考察训练集化合物的两两相似性, 建立相似性网络图(NSG)^[48], 可实现结构多样性分布的可视化, 亦可借助社区识别^[49]或层级聚类分析来确定训练集化合物的结构聚类^[48]. 在典型的NSG中, 节点表示化合物; 若两个化合物结构相似度大于预设的阈值, 例如, 基于拓展连接性指纹(ECFP)^[25]的 Tc 值大于0.55时^[50], 则将对应的节点通过边连接起来. Tox21化学品^[51]的NSG如图2(a)所示, 遍布的连通子图直观地展示了其结构的多组聚类.

NSG节点的颜色一般标度活性值, 如图2(a)中的线粒体膜电位干扰性(PubChem AID: 720637)^[52,53]. 因此, NSG比描述符域额外考虑了响应值(y)的信息. 基于描述符 x , 也可以考虑结构特征与响应值的联合分布 $P(x, y)$. 例如, Tox21化学品^[51]的线粒体膜电位干扰活性与 Mw - $SLogP$ 描述符形成的三维空间如图2(b), (c)所示, 亦称结构-活性地貌(SAL). 注意到, SAL中呈现出密集的“山脊”, 意味着两个化合物的 Mw (或 $SLogP$)值接近时, 其活性值非但不接近, 还可能差异巨大, 即SAL是崎岖而非平缓, 跳跃而非连续的. 这种现象被形象地描述为“活性悬崖”^[54-56], 也违背了QSAR研究的基本假设: 相似性原理.

SAL的形貌特征取决于表征SAL的描述符. 出现如图2(b), (c)所示的密集的活性悬崖, 说明 Mw 或 $SLogP$ 都不适合作为预测线粒体膜电位干扰性的描述符. 在不确定采用何种描述符绘制SAL, 或描述符维度较高(如 $m>2$)而难以可视化SAL时, 可以通过NSG来考察SAL的性质. 在图2(a)所示的NSG中, 每个化合物节点(u)的大小, 由局域不连续得分值(LDS)^[48]表征:

$$LDS(u) = \frac{1}{Q} \sum_{\{v | Tc(u, v) \geq S_{\text{cutoff}}, u \neq v\}} Tc(u, v) D(u, v), \quad (6)$$

式中, 与 u 的相似度 $Tc(u, v) \geq S_{\text{cutoff}}$ 的化合物 v 构成集合 $\{v | Tc(u, v) \geq S_{\text{cutoff}}, u \neq v\}$, Q 是集合中元素的数目(即NSG中与 u 相连的节点个数), $D(u, v)$ 表示 u 与 v 活性的绝对差值. 节点越大, 说明其与相似结构化合物的平均活性差异越大, 围绕该节点的局域结构-活性关系违背相似性原理的程度也越大. 注意到, 图2(a)中部分具有共同结构特征的社区或连通子图的节点具有较小的 LDS , 说明

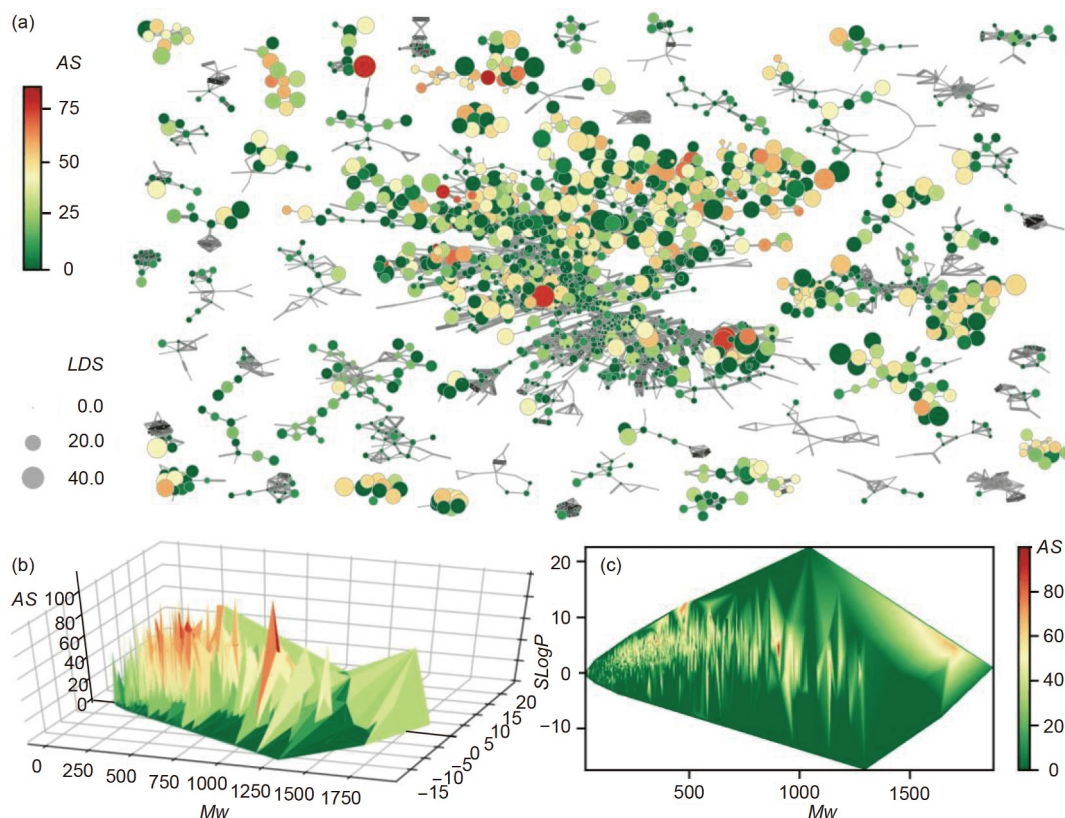


图2 (网络版彩色)Tox21化学品的结构-活性(线粒体膜电位干扰性^[52])地貌。(a) 相似性网络图; Mw , $SLogP$ 与终点数值呈现的三维地貌(b)及等高线图(c)。AS: 活性分值; LDS: 局域不连续性得分值; 未显示节点数小于10的连通子图

Figure 2 (Color online) Structure-activity (mitochondrial membrane potential disruption^[52]) landscape of the Tox21 chemicals. (a) Network-like similarity graph; 3D topography (b) and associated contour (c) presented by Mw , $SLogP$ and the endpoint values. AS: Activity score; LDS: Local discontinuity score; connected subgraphs with size < 10 are not shown

这些化合物的局域不连续性较小, 局域结构-活性地貌较为平滑。

结构-活性不连续性小的数据集更满足相似性原理, 也更可能被QSAR模型较准确地预测。以Tox21项目^[57]在PubChem上已经发表的59项终点活性为例(pubchem.ncbi.nlm.nih.gov/source/824, 截至2020年10月), 以1和0分别表示有活性和无活性, 使用不同的分子描述符体系构建随机森林分类器(利用Scikit-Learn程序^[58], 取1500棵树, 其余参数默认)。如图3所示, 尽管各终点训练集化合物LDS的平均值(LDS_{AVG} , 衡量该终点训练集的结构-活性不连续性)与受试者工作特性曲线下面积(AUC_{CV} , 衡量分类性能的指标, AUC 越大, 分类性能越好)^[61]没有显著相关性(图3(a)), 却与对数损失(亦称交叉熵损失, LL_{CV} , 衡量分类性能的指标, 对数损失越大, 分类效果越差)^[62]有显著的线性关系(图3(b))。因此, 在比较分析基于不同数据集训练的QSAR模型的交叉验

证效果时, 训练集的结构-活性不连续性提供了一种全新的解读视角。

此外, 局域不连续性也能够解释回归模型在交叉验证中对单一化合物的预测能力。例如, Wang等人^[47]基于2个训练集、5种机器学习算法以及3类描述符体系的不同组合, 训练了预测表征化学品与人体过氧化物酶体增殖物激活受体 γ (PPAR γ)结合能力参数的多个回归模型。对于每个模型, 单一化合物的LDS均与其交叉验证绝对预测误差之间呈现显著的线性正相关关系, 意味着训练集化合物的局域不连续性可以反映模型对其预测的不确定性^[47]。因此, 局域不连续性进一步丰富了基于相似性而定义的结构域内涵, 提供了一种量化评估针对应用域内单一化合物预测不确定性的新思路。

2.3 机理域

违背QSAR相似性原理的活性悬崖现象或结构-活

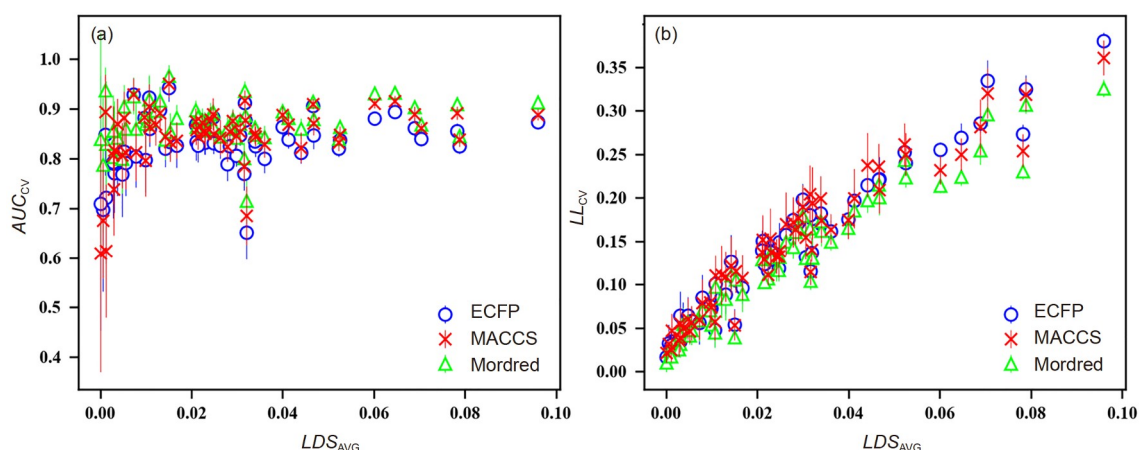


图3 (网络版彩色)随机森林分类器在Tox21项目59种生物检测终点的交叉验证性能与训练集网络相似性图节点平均局域不连续性得分值(LDS_{AVG})之间的关系。(a) 受试者工作特性曲线下面积(AUC_{CV}); (b) 对数损失(LL_{CV})。ECFP: 拓展连接性指纹; MACCS^[59]: 分子获取结构键; Mordred^[60]: 非三维描述符集

Figure 3 (Color online) Relationships between cross-validation performances of random forest classifiers on the 59 Tox21 bioassays and the average local discontinuity scores of associated training-set network-like similarity graphs (LDS_{AVG}). (a) Area under receiver operating characteristic curve (AUC_{CV}); (b) log-loss (LL_{CV}). ECFP: Extended connectivity fingerprints; MACCS^[59]: Molecular access structural keys; Mordred^[60]: A specific set of non-3D descriptors

性不连续性为什么存在? 这个问题需要从活性现象背后的机制来回答。以测试核受体激动剂活性的*in vitro*报告基因体系^[63]为例, 激动剂分子首先要穿过细胞膜, 进入核受体的配体结合口袋, 进而改变配体结合域构象, 促使共抑制蛋白离去和/或共激活蛋白的募集, 最终组成转录机器表达耦合在下游的荧光蛋白基因, 翻译为荧光蛋白而被检测器读取。其中与小分子结构直接关联的两个子过程是: (1) 小分子在*in vitro*体系中的动态分布; (2) 小分子与受体蛋白结合复合体的功能性变构^[64]。

分子在*in vitro*体系中的动态分布, 本质上是(亚)细胞水平的毒代动力学过程^[65]。其热力学平衡状态取决于分子在不同溶剂(水、脂质等)内的溶解能、在不同界面(器皿壁、膜系统)的吸附能, 以及与不同生物大分子(酶、受体蛋白、功能RNA)结合能的相对高低^[66,67]。这些凝聚态系统的复杂性和空间异质性差异巨大。均相溶剂系统的空间异质性, 随着溶剂分子的流动扩散而降低, 尽管溶剂分子尺寸和柔性会影响其流动扩散的自由性, 总体上均相溶剂系统复杂度和空间异质性都很低(图4(a))。相比而言, 界面相引入了不能自由流动扩散的表面而增加了其复杂性, 非对称性和/或柔性的表面也提升了其空间异质性(图4(b))。蛋白质口袋则是界面相的极端案例, 其内表面主要由非对称、高柔性的氨基酸侧链组成, 进一步减少了配体与自由流转

的溶剂分子直接接触的机会(图4(c))。

空间异质性强的界面可以通过特异性作用(如氢键、盐桥、形状互补、前线轨道分布), 实现选择性吸附或结合特定结构的分子, 或选择性催化特殊类型的反应。而分子结构看似细微的改变(包括手性翻转), 则可能破坏或构建特异性作用, 从而导致界面相对分子的吸附或特定反应的催化能力发生突变。因此, 复杂系统中的空间异质界面, 对特定分子结构和/或特定化学反应的选择性, 是出现活性悬崖的原因之一。

另一方面, 小分子与受体蛋白复合体的功能性变构(图4(c))^[64,68,69], 本质上是分子水平的毒效动力学过程^[70]。核受体这类需要空间变构来发挥功能的生物大分子, 本身就是复杂系统, 亦可抽象为网络: 节点表示原子, 具有空间位置和速度等属性; 边表示原子间的相互作用。给定描述原子相互作用的规则, 即可考察系统构象动态变化与分布规律。事实上, 基于经验力场的分子动力学模拟就是模拟该复杂系统的经典方法^[71]。

根据复杂系统混沌理论, 尽管原子之间所有相互作用规则都有明确数学物理定义, 但系统层级行为, 仍然可能随着初值的细微变化而剧变^[72]。因此, 不能总是从局部、静态的相互作用分析出发, 确定性地预测小分子结构细微改变(即调整初值)对于复合体构象分布的影响。进一步, 核受体、共调节蛋白以及转录机器相关的蛋白共同作用, 形成细胞信号通路网络^[73-76]。蕴含

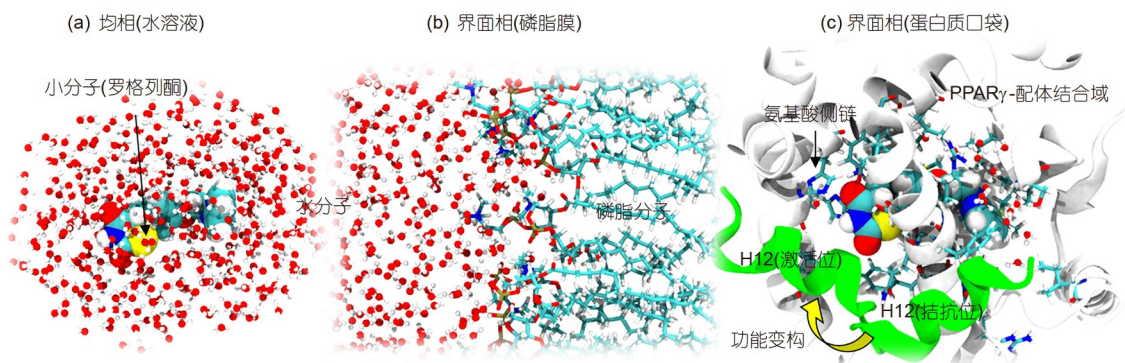


图4 (网络版彩色)不同凝聚态系统的复杂性和空间异质性。(a) 抗糖尿病药物分子罗格列酮溶解在水溶液中;(b) 磷脂膜与水相之间的界面;(c) 罗格列酮结合在过氧化物酶体增殖物激活受体 γ (PPAR γ)配体结合口袋内。H12肽段的位置变化调控着PPAR γ 对共调节蛋白的募集
Figure 4 (Color online) Complexity and spatial heterogeneity of different condensed state systems. (a) Anti-diabetic drug rosiglitazone solved in water solution; (b) interface between phospholipid membrane and water solution; (c) rosiglitazone bound inside the ligand binding pocket of peroxisome proliferator-activated receptor γ (PPAR γ). Position shift of the H12 peptide segment regulates the recruitment of coregulators by PPAR γ

于其中的非线性,又孕育出新的混沌^[77]。因此,复杂系统行为的涌现性^[78]与混沌性^[79],也是出现活性悬崖的原因之一。

综上,活性终点描述了什么系统,该系统复杂性和空间异质性如何,是否考察了系统层级的涌现行为。弄清这3个问题,有助于选择合适的描述符,解释QSAR机理,并推断应用域范围。例如,与小分子均相溶解行为有关终点的预测模型一般具有较简单的线性关系,在结构较广的化合物集上有较好的预测效果^[20,80,81]。在这些模型中,描述符通常为表征分子整体性质的体积、分子极化率、溶剂化色散参数^[20]等,而静电、原子部分电荷、氢键相互作用等锚定在局部基团上的描述符,在溶剂分子自由扩散的条件下,几乎不会受到所考察溶质分子取向和构象的影响。

相比而言,毒代动力学以及毒性效应相关的终点涉及生物系统。因此,相关QSAR模型的应用需要考虑特定的毒性机制。水生毒性测试中,引发鱼类急性毒性症状的毒性机制分为非极性麻醉、极性麻醉、氧化磷酸化解偶联等^[82]。基于此,Verhaar分类框架^[83]将化学品分为惰性、弱惰性、反应性、特异作用等类别^[83]。其中,惰性化学品无反应性,不与受体蛋白特异性作用,对应于非极性麻醉。此时,水生急性毒性(LC_{50})与化学品的疏水性($\log K_{OW}$)呈现良好的线性关系。随着分子结构中添加氢键供/受体、反应性基团等,毒性机制涉及系统的复杂性和空间异质性增加, LC_{50} 与 $\log K_{OW}$ 之间的线性关系逐渐不再适用。

对于化学反应性主导的终点,如引发致突变性的

DNA加成反应,反映电子结构的轨道能量和原子部分电荷等描述符更加适用^[84]。对特异性作用于受体蛋白的终点,表征参与特异性作用的关键结构模式的碎片,以及分子在蛋白口袋的取向、构象、诱导契合效应等,如结构警示^[85,86]、CoMFA及高维QSAR方法^[87,88],以及借助分子对接、分子动力学模拟来计算构象系综的描述符^[89],都有助于改善QSAR预测效果。这些操作,本质上添加了关于复杂空间异质性界面的描述信息,其效果则依赖于配体-受体复合体初始构象的正确性以及分子模拟方法可靠性,多限于共享母体结构或结构非常类似的系列化合物。

综上,若预测终点是非特异性的,相关QSAR模型应用域覆盖面较宽泛,模型跨结构外推能力较强;若预测终点涉及特异性作用机制,相关QSAR模型则局限于与训练集化合物结构相似的分子^[39]。基于毒性机制定义应用域,亦称机理域,曾被认为是保证模型预测可靠性的最严格标准^[2]。然而,机理域的概念仅在讨论相对宏观的毒性效应(如致突变性、水生急性毒性)才有意义。2010年,有害结局路径(AOP)概念框架被提出,将宏观毒性效应解构为以化学分子与生物大分子相互作用(即分子起始事件, MIEs)为起点的关键事件序列^[90,91]。此后,围绕MIEs的毒性测试方法开发与验证的研究和技术导则显著增多^[92-94]。例如,Tox21项目^[57]的测试终点几乎全是MIEs,很难再拆分为更基元的机理。尽管如此, QSAR预测效果会受到终点系统复杂性、空间异质性、或系统涌现^[78]和混沌^[79]的影响,对其微观机制的思考总是裨益良多。

3 总结

QSAR模型能够实现化学品活性的虚拟高通量预测或筛选,与交叉参照(read-across)方法共同构成了快速填补化学品数据空缺的“非测试策略(non-testing strategy)”。然而,这些方法本质上都是在归纳并应用化学品的描述符(或性质)与终点数据之间的相关规律,依赖于特定的训练数据,受到应用域的约束。

在毒理学进入大数据的时代^[40-42], QSAR模型往往需要处理成千上万甚至上百万量级的化学结构数据集。相比计算缓慢的基于3D构象和电子结构的描述符体系,可快速计算的碎片、指纹和基于分子拓扑的描述符体系得到了更广泛的应用,尤其在深度学习对化学分子的建模中,后者占据主导地位^[26]。大数据和机器学习时代的QSAR建模情景中,描述符往往具有较高的维度(≥ 3),或者隐含在机器学习流程内部,难以用经典的描述符域表征^[26,95]。此时,本文介绍的基于分子指纹相似度的结构域^[46,47],提供了一种可行的应用域定义

方案。

仅采用小分子结构信息作为描述符,无法从复杂系统的空间异质性或涌现^[78]的角度揭示机理,因而产生了活性悬崖^[54]现象。此时,只有借助更多的数据提升SAL的分辨率,才有可能勾勒出活性悬崖边界,或识别出结构-活性关系较为连续的化合物群。注意到,为应对不平衡或小样本数据问题,机器学习领域常基于相似性原理构造人工样本,包括半监督学习或过采样方法^[96]。然而,活性悬崖的存在否定了相似性原理的绝对性。因此,需要谨慎使用构造人工样本的策略,以免产生大量错误训练样本,而训练出错误的模型。

随着中国对化学品环境管理的加强,对新污染物治理的重视^[4,97],预期QSAR模型的开发与使用将会增加。只有对终点机制及描述符有基本的理解,并充分认识到所开发或拟使用的QSAR模型的特点,针对性地选择应用域表征方法,才能够给出相对可靠的预测,从而对我国化学品环境管理起到正面推动作用。

参考文献

- 1 Sedykh A, Zhu H, Tang H, et al. Use of *in vitro* HTS-derived concentration-response data as biological descriptors improves the accuracy of QSAR models of *in vivo* toxicity. *Environ Health Perspect*, 2011, 119: 364–370
- 2 Chen J W, Li X H, Yu H Y, et al. (Q)SAR technology oriented for ecological risk assessment of hazardous organics: Advances and perspectives (in Chinese). *Sci Sin Chim*, 2008, 38: 461–474 [陈景文, 李雪花, 于海瀛, 等. 面向毒害有机物生态风险评价的(Q)SAR技术: 进展与展望. *中国科学: 化学*, 2008, 38: 461–474]
- 3 Wang Z Y, Chen J W, Qiao X L, et al. Computational toxicology: Oriented for chemicals risk assessment (in Chinese). *Sci Sin Chim*, 2016, 46: 222–240 [王中钰, 陈景文, 乔显亮, 等. 面向化学品风险评价的计算(预测)毒理学. *中国科学: 化学*, 2016, 46: 222–240]
- 4 Ruan T, Jiang G B. Basic theory and analytical methodology for identification of novel environmental organic pollutants (in Chinese). *Bull Chin Acad Sci*, 2020, 35: 1328–1336 [阮挺, 江桂斌. 发现新型环境有机污染物的基本理论与方法. *中国科学院院刊*, 2020, 35: 1328–1336]
- 5 Jaworska J S, Comber M, Auer C, et al. Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints. *Environ Health Perspect*, 2003, 111: 1358–1360
- 6 Buschmann J. The OECD guidelines for the testing of chemicals and pesticides. *Methods Mol Biol*, 2013, 947: 37–56
- 7 Wang Z, Walker G W, Muir D C G, et al. Toward a global understanding of chemical pollution: A first comprehensive analysis of national and regional chemical inventories. *Environ Sci Technol*, 2020, 54: 2575–2584
- 8 Zimmerman J B, Anastas P T. Toward designing safer chemicals. *Science*, 2015, 347: 215
- 9 Zimmerman J B, Anastas P T. Toward substitution with no regrets. *Science*, 2015, 347: 1198–1199
- 10 Reymond J L. The chemical space project. *Acc Chem Res*, 2015, 48: 722–730
- 11 Zimmerman J B, Anastas P T, Erythropel H C, et al. Designing for a green chemistry future. *Science*, 2020, 367: 397–400
- 12 Krinsky S. The unsteady state and inertia of chemical regulation under the US Toxic Substances Control Act. *PLoS Biol*, 2017, 15: e2002404
- 13 Krewski D, Andersen M E, Tyshenko M G, et al. Toxicity testing in the 21st century: Progress in the past decade and future perspectives. *Arch Toxicol*, 2020, 94: 1–58
- 14 Cherkasov A, Muratov E N, Fourches D, et al. QSAR modeling: Where have you been? Where are you going to? *J Med Chem*, 2014, 57: 4977–5010
- 15 Tang W, Chen J, Wang Z, et al. Deep learning for predicting toxicity of chemicals: A mini review. *J Environ Sci Health Part C*, 2018, 36: 252–271
- 16 Muratov E N, Bajorath J, Sheridan R P, et al. QSAR without borders. *Chem Soc Rev*, 2020, 49: 3525–3564
- 17 Zhou Z H. Machine Learning (in Chinese). Beijing: Tsinghua University Press, 2016 [周志华. 机器学习. 北京: 清华大学出版社, 2016]

- 18 Todeschini R, Consonni V. Molecular Descriptors for Chemoinformatics. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, 2009
- 19 Netzeva T I, Gallegos Saliner A, 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: 1223–1230
- 20 Abraham M H, Ibrahim A, Zissimos A M. Determination of sets of solute descriptors from chromatographic measurements. *J Chromatogr A*, 2004, 1037: 29–47
- 21 Hammett L P. The effect of structure upon the reactions of organic compounds. Benzene derivatives. *J Am Chem Soc*, 1936, 59: 96–103
- 22 Taft Jr R W. Linear steric energy relationships. *J Am Chem Soc*, 1953, 75: 4538–4539
- 23 Heller S R, McNaught A, Pletnev I, et al. InChI, the IUPAC international chemical identifier. *J Cheminformatics*, 2015, 7: 23
- 24 Carhart R E, Smith D H, Venkataraghavan R. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J Chem Inf Comput Sci*, 1985, 25: 64–73
- 25 Rogers D, Hahn M. Extended-connectivity fingerprints. *J Chem Inf Model*, 2010, 50: 742–754
- 26 Mater A C, Coote M L. Deep learning in chemistry. *J Chem Inf Model*, 2019, 59: 2545–2559
- 27 Crippen G M. A novel approach to calculation of conformation: Distance geometry. *J Comput Phys*, 1977, 24: 96–107
- 28 Hopfinger A J. A QSAR investigation of dihydrofolate reductase inhibition by Baker triazines based upon molecular shape analysis. *J Am Chem Soc*, 1980, 102: 7196–7206
- 29 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: 5959–5967
- 30 Attene-Ramos M S, Miller N, Huang R, et al. The Tox21 robotic platform for the assessment of environmental chemicals—From vision to reality. *Drug Discov Today*, 2013, 18: 716–723
- 31 Wildman S A, Crippen G M. Prediction of physicochemical parameters by atomic contributions. *J Chem Inf Comput Sci*, 1999, 39: 868–873
- 32 Chen D X. Convex calculation in the plane (in Chinese). *China Inf Technol Educ*, 2020, (21): 25–29 [陈道蓄. 平面上的凸包计算. 中国信息技术教育, 2020, (21): 25–29]
- 33 Virtanen P, Gommers R, Oliphant T E, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat Methods*, 2020, 17: 261–272
- 34 Scott D W. Multivariate Density Estimation: Theory, Practice, and Visualization. New York, Chichester: John Wiley & Sons, 1992
- 35 Heidenreich N B, Schindler A, Sperlich S. Bandwidth selection for kernel density estimation: A review of fully automatic selectors. *ASSt Adv Stat Anal*, 2013, 97: 403–433
- 36 Netzeva T I, Worth A P, Aldenberg T, et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. *Altern Lab Anim*, 2005, 33: 155–173
- 37 Jaworska J, Nikolova-Jeliazkova N, Aldenberg T. QSAR applicability domain estimation by projection of the training set in descriptor space: A review. *Altern Lab Anim*, 2005, 33: 445–459
- 38 Hansch C, Fujita T. ρ - σ - π analysis. A method for the correlation of biological activity and chemical structure. *J Am Chem Soc*, 1964, 86: 1616–1626
- 39 Eriksson L, Jaworska J, Worth A P, et al. Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs. *Environ Health Perspect*, 2003, 111: 1361–1375
- 40 Wang Y, Bolton E, Dracheva S, et al. An overview of the PubChem BioAssay resource. *Nucleic Acids Res*, 2010, 38: D255–D266
- 41 Mendez D, Gaulton A, Bento A P, et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res*, 2019, 47: D930–D940
- 42 Zhu H, Zhang J, Kim M T, et al. Big data in chemical toxicity research: The use of high-throughput screening assays to identify potential toxicants. *Chem Res Toxicol*, 2014, 27: 1643–1651
- 43 Maggiora G, Vogt M, Stumpfe D, et al. Molecular similarity in medicinal chemistry. *J Med Chem*, 2014, 57: 3186–3204
- 44 Bajusz D, Racz A, Heberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminformatics*, 2015, 7: 20
- 45 Rogers D J, Tanimoto T T. A computer program for classifying plants. *Science*, 1960, 132: 1115–1118
- 46 Wang Z, Chen J, Hong H. Applicability domains enhance application of PPAR γ agonist classifiers trained by drug-like compounds to environmental chemicals. *Chem Res Toxicol*, 2020, 33: 1382–1388
- 47 Wang Z, Chen J, Hong H. Developing QSAR models with defined applicability domains on PPAR γ binding affinity using large data sets and machine learning algorithms. *Environ Sci Technol*, 2021, 55: 6857–6866
- 48 Wawer M, Peltason L, Weskamp N, et al. Structure—Activity relationship anatomy by network-like similarity graphs and local structure—Activity relationship indices. *J Med Chem*, 2008, 51: 6075–6084
- 49 Newman M E J. Modularity and community structure in networks. *Proc Natl Acad Sci USA*, 2006, 103: 8577–8582
- 50 Stumpfe D, Bajorath J. Exploring activity cliffs in medicinal chemistry. *J Med Chem*, 2012, 55: 2932–2942

- 51 Tice R R, Austin C P, Kavlock R J, et al. Improving the human hazard characterization of chemicals: A Tox21 update. *Environ Health Perspect*, 2013, 121: 756–765
- 52 Attene-Ramos M S, Huang R, Michael S, et al. Profiling of the Tox21 chemical collection for mitochondrial function to identify compounds that acutely decrease mitochondrial membrane potential. *Environ Health Perspect*, 2015, 123: 49–56
- 53 Xia M, Huang R, Shi Q, et al. Comprehensive analyses and prioritization of Tox21 10K chemicals affecting mitochondrial function by in-depth mechanistic studies. *Environ Health Perspect*, 2018, 126: 077010
- 54 Maggiora G M. On outliers and activity cliffs: Why QSAR often disappoints. *J Chem Inf Model*, 2006, 46: 1535
- 55 Vogt M, Huang Y, Bajorath J. From activity cliffs to activity ridges: Informative data structures for SAR analysis. *J Chem Inf Model*, 2011, 51: 1848–1856
- 56 Wassermann A M, Bajorath J. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J Chem Inf Model*, 2010, 50: 1248–1256
- 57 Huang R, Xia M. Editorial: Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental toxicants and drugs. *Front Environ Sci*, 2017, 5: 3
- 58 Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*, 2011, 12: 2825–2830
- 59 MDL Information Systems Inc. Molecular ACCess System (MACCS). San Leandro: MDL Information Systems Inc., 2005
- 60 Moriwaki H, Tian Y S, Kawashita N, et al. Mordred: A molecular descriptor calculator. *J Cheminformatics*, 2018, 10: 4
- 61 Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett*, 2006, 27: 861–874
- 62 Bishop C M. *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006
- 63 Legler J, van den Brink C E, Brouwer A, et al. Development of a stably transfected estrogen receptor-mediated luciferase reporter gene assay in the human T47D breast cancer cell line. *Toxicol Sci*, 1999, 48: 55–66
- 64 Shang J, Mosure S A, Zheng J, et al. A molecular switch regulating transcriptional repression and activation of PPAR γ . *Nat Commun*, 2020, 11: 956
- 65 Wetmore B A, Wambaugh J F, Ferguson S S, et al. Integration of dosimetry, exposure, and high-throughput screening data in chemical toxicity assessment. *Toxicol Sci*, 2012, 125: 157–174
- 66 Seibert H, Mörchel S, Gülden M. Factors influencing nominal effective concentrations of chemical compounds *in vitro*: Medium protein concentration. *Toxicol Vitro*, 2002, 16: 289–297
- 67 Gülden M, Mörchel S, Seibert H. Factors influencing nominal effective concentrations of chemical compounds *in vitro*: Cell concentration. *Toxicol Vitro*, 2001, 15: 233–243
- 68 Bruning J B, Parent A A, Gil G, et al. Coupling of receptor conformation and ligand orientation determine graded activity. *Nat Chem Biol*, 2010, 6: 837–843
- 69 Nagy L, Schwabe J W R. Mechanism of the nuclear receptor molecular switch. *Trends Biochem Sci*, 2004, 29: 317–324
- 70 Jager T, Albert C, Preuss T G, et al. General unified threshold model of survival—A toxicokinetic-toxicodynamic framework for ecotoxicology. *Environ Sci Technol*, 2011, 45: 2529–2540
- 71 Pan L M, Lü X, Zhuang S L. The application of molecular dynamics simulations in mechanism of toxicity of organic contaminants (in Chinese). *Chin Sci Bull*, 2015, 60: 1781–1788 [潘柳萌, 吕翀, 庄树林. 分子动力学模拟在有机污染物毒性作用机制中的应用. 科学通报, 2015, 60: 1781–1788]
- 72 Wu T. Non-linear dynamics chaos theory and its significance (in Chinese). *J Tsinghua Univ (Phil Soc Sci)*, 2000, 15: 72–79 [吴彤. 非线性动力学混沌理论方法及其意义. 清华大学学报(哲学社会科学版), 2000, 15: 72–79]
- 73 Evans R M, Mangelsdorf D J. Nuclear receptors, RXR, and the big bang. *Cell*, 2014, 157: 255–266
- 74 Rosenfeld M G, Lunyak V V, Glass C K. Sensors and signals: A coactivator/corepressor/epigenetic code for integrating signal-dependent programs of transcriptional response. *Genes Dev*, 2006, 20: 1405–1428
- 75 Perissi V, Aggarwal A, Glass C K, et al. A corepressor/coactivator exchange complex required for transcriptional activation by nuclear receptors and other regulated transcription factors. *Cell*, 2004, 116: 511–526
- 76 Zomerdijs J. Pivotal findings for a transcription machine. *Nature*, 2013, 502: 629–630
- 77 Rivkind A, Schreiber H, Brenner N, et al. Scale free topology as an effective feedback system. *PLoS Comput Biol*, 2020, 16: e1007825
- 78 Atay F M, Jost J. On the emergence of complex systems on the basis of the coordination of complex behaviors of their elements: Synchronization and complexity. *Complexity*, 2004, 10: 17–22
- 79 Judd K. *Chaos in Complex Systems*. Boston: Birkhäuser, 1990
- 80 Kamlet M J, Abboud J L M, Abraham M H, et al. Linear solvation energy relationships. 23. A comprehensive collection of the solvatochromic parameters, π^* , α , and β , and some methods for simplifying the generalized solvatochromic equation. *J Org Chem*, 1983, 48: 2877–2887
- 81 Wilson L Y, Famini G R. Using theoretical descriptors in quantitative structure-activity relationships: Some toxicological indices. *J Med Chem*,

- 1991, 34: 1668–1674
- 82 McKim J M, Bradbury S P, Niemi G J. Fish acute toxicity syndromes and their use in the QSAR approach to hazard assessment. *Environ Health Perspect*, 1987, 71: 171–186
- 83 Verhaar H J M, van Leeuwen C J, Hermens J L M. Classifying environmental pollutants. *Chemosphere*, 1992, 25: 471–491
- 84 Townsend P A, Grayson M N. Density functional theory in the prediction of mutagenicity: A perspective. *Chem Res Toxicol*, 2021, 34: 179–188
- 85 Wedlake A J, Folia M, Piechota S, et al. Structural alerts and random forest models in a consensus approach for receptor binding molecular initiating events. *Chem Res Toxicol*, 2020, 33: 388–401
- 86 Allen T E H, Nelms M D, Edwards S W, et al. *In silico* guidance for *in vitro* androgen and glucocorticoid receptor ToxCast assays. *Environ Sci Technol*, 2020, 54: 7461–7470
- 87 Li M Y, Xia L. From 3D to 5D: Advances in QSAR (in Chinese). *J China Phar Univ*, 2003, 34: 586–591 [李敏勇, 夏霖. 从三维到五维: 定量构效关系(QSAR)的研究进展. *中国药科大学学报*, 2003, 34: 586–591]
- 88 Vedani A, Dobler M, Lill M A. Combining protein modeling and 6D-QSAR. Simulating the binding of structurally diverse ligands to the estrogen receptor. *J Med Chem*, 2005, 48: 3700–3703
- 89 Ash J, Fourches D. Characterizing the chemical space of ERK2 kinase inhibitors using descriptors computed from molecular dynamics trajectories. *J Chem Inf Model*, 2017, 57: 1286–1299
- 90 Ankley G T, Bennett R S, Erickson R J, et al. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environ Toxicol Chem*, 2010, 29: 730–741
- 91 Allen T E H, Goodman J M, Gutsell S, et al. Defining molecular initiating events in the adverse outcome pathway framework for risk assessment. *Chem Res Toxicol*, 2014, 27: 2100–2112
- 92 Ellison C M, Piechota P, Madden J C, et al. Adverse outcome pathway (AOP) informed modeling of aquatic toxicology: QSARs, read-across, and interspecies verification of modes of action. *Environ Sci Technol*, 2016, 50: 3995–4007
- 93 Patlewicz G, Kuseva C, Kesova A, et al. Towards AOP application—Implementation of an integrated approach to testing and assessment (IATA) into a pipeline tool for skin sensitization. *Regul Toxicol Pharmacol*, 2014, 69: 529–545
- 94 Villeneuve D L, Crump D, Garcia-Reyero N, et al. Adverse outcome pathway (AOP) development I: Strategies and principles. *Toxicol Sci*, 2014, 142: 312–320
- 95 Sanchez-Lengeling B, Aspuru-Guzik A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 2018, 361: 360–365
- 96 Li Y X, Chai Y, Hu Y Q, et al. Review of imbalanced data classification methods (in Chinese). *Control Decis*, 2019, 34: 673–688 [李艳霞, 柴毅, 胡友强, 等. 不平衡数据分类方法综述. *控制与决策*, 2019, 34: 673–688]
- 97 Yu F, Li Q S. Difficulties in the treatment of novel pollutants in China (in Chinese). *China Environment News*, 2020-11-23 [於方, 李秋爽. 我国新污染物治理难在哪? *中国环境报*, 2020-11-23]

Summary for “QSAR模型应用域的特征方法”

Characterization of applicability domains for QSAR models

Zhongyu Wang, Jingwen Chen^{*}, Zhiqiang Fu & Xuehua Li

Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), School of Environmental Science and Technology, Dalian University of Technology, Dalian 116024, China

^{*} Corresponding author, E-mail: jwchen@dlut.edu.cn

In the field of environmental science and engineering, quantitative structure-activity relationship (QSAR) means the quantitative relationship between the structure of molecules (or their aggregates e.g., nanoparticles) and certain endpoints. Herein, endpoints generally refer to physicochemical properties, biological effects or environmental behavior parameters, etc. that can be measured or modeled. Based on data sets of chemical structures and their known endpoint values (i.e., training set), QSAR models could, by means of specific algorithms, establish the mathematical relationships between the digital features that characterize the molecular structure (i.e., descriptors) and the endpoint values. Then, the established mathematical relationships can be employed to predict the endpoint values for given chemical structures.

QSAR models are important tools for filling the data gap in environmental safety of chemicals and addressing the issues from so-called “emerging pollutants” that are closely related to the improper management of chemicals. Notably, QSAR models are intrinsically data-driven models. The relationships presented in the training set is not necessarily applicable to arbitrary chemical structures. Reliability of QSAR models is always limited to certain applicability domains. Therefore, acceptance of QSAR models in sound management of chemicals requires clearly defined applicability domains. This study reviewed three concepts of the applicability domain: Descriptor domain, structural domain and mechanism domain.

For characterizing descriptor domain, methods based on hyper-rectangle, convex hull, joint probability density estimation and various types of distances were described. Notably, when Boolean fingerprints are used as descriptors, these methods become meaningless. Thus, implementation, characters and advantages of the structural domain based on fingerprints and similarity, were specially introduced. Moreover, structure-activity landscapes (SALs), as exemplified by a network-like similarity graph (NSG) and a 3D topography of the endpoint and the descriptor values, were also demonstrated. “Activity cliffs” manifested in the SALs were highlighted. Based on the NSG, local discontinuity score (*LDS*) was calculated to quantitatively describe the “smoothness” of the SAL. *LDS* was also demonstrated to be linearly correlated to the cross-validation log loss of random forest classifiers for 59 Tox21 bioassay endpoints, further implying that smoothness of the training SAL could be an indicator of the performance of the established QSAR models. The causes of the “activity cliff” were further discussed and assumed to be dependent on the complexity and spatial heterogeneity of the investigated endpoint systems. Particularly, the chaos and emergence behavior of complex systems could be sensitive to tiny but specific structural changes in the small molecules and thus cannot be satisfactorily predicted based purely on the molecular descriptors. Hence, the understanding of the mechanisms underlying the endpoints was emphasized, which corresponds to the concept of mechanism domain.

In conclusion, in order to understand the feasibility of the molecular descriptors, explain the mechanisms of the QSARs, and reasonably select defining methods for the applicability domain, it is essential to address three questions: (1) What system does the endpoint intrinsically describe? (2) How complex and spatially heterogeneous is the system? (3) Whether the endpoint considers the emergence of systems behavior? With the enhancement of chemicals management and the emphasis on the treatment of novel pollutants, development and use of QSAR models are anticipated to be increased. Only with basic knowledge of the mechanisms underlying the modeled endpoints, with sufficient recognition of the characters of the established QSAR models, and with properly selected applicability domain characterization methods, could the QSAR models yield relatively reliable predictions and thus benefit the environmental management of chemicals in China.

quantitative structure-activity relationship (QSAR), applicability domain, descriptor, activity cliffs, structure-activity landscape

doi: [10.1360/TB-2021-0406](https://doi.org/10.1360/TB-2021-0406)