Eco-Environmental Knowledge Web

环 境 化 学
ENVIRONMENTAL CHEMISTRY

第 40 卷第 5 期 2021 年 5 月
Vol. 40, No. 5 May 2021

DOI:10.7524/j.issn.0254-6108.2021011304

丁蕊, 陈景文, 于洋, 等. 基于集成学习算法构建有机化学品鱼体生物富集因子的 QSAR 预测模型[J]. 环境化学, 2021, 40(5): 1295-1305.
DING Rui, CHEN Jingwen, YU Yang, et al. Using ensemble learning algorithms to develop QSAR models on bioconcentration factors of organic chemicals in multispecies fish [J]. Environmental Chemistry, 2021, 40 (5): 1295-1305.

《环境化学》创刊 40 周年纪念专题

基于集成学习算法构建有机化学品鱼体生物富集因子的 QSAR 预测模型*

丁 蕊[1] 陈景文[1]** 于 洋[2] 林 军[2] 王中钰[1] 唐伟豪[1] 李雪花[1]

（1. 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，大连理工大学环境学院，大连，116024；2. 生态环境部固体废物与化学品管理技术中心，北京，100029）

摘 要 生物富集因子 (BCF) 是评价化学品生物累积能力的重要参数. 目前全球市场上使用的化学品数量已超过了 35 万种，但是只有一千多种化学品具有 BCF 值. 定量构效关系 (QSAR) 模型被认为是一种有效填补数据空缺的方法. 目前大多数预测 BCF 的 QSAR 模型为单一模型，而集成模型可能会对 BCF 的预测效果有所改进. 本研究建立了一个全面的鱼类 BCF 数据库，涵盖 1 300 多种有机化学品的 BCF 实测值. 基于此数据库，依据 QSAR 模型构建和验证导则，使用多种机器学习算法建立了预测鱼类 BCF 的 5 种单一模型和 11 种集成模型. 结果表明，与单一模型相比，集成模型具有更好的拟合能力、稳健性、预测准确性以及更广泛的应用域. 进一步使用最优集成模型对《中国现有化学物质清单》(IECSC) 中化学物质的 BCF 进行了预测，结果表明该清单中有 1 066 种化学物质具有生物累积性，86 种化学物质具有强生物累积性. 本研究所构建的模型可为化学品生物累积能力评估提供必要数据，支持化学品风险评价与管理工作.

关键词 生物富集因子，定量构效关系，机器学习，集成模型，应用域.

Using ensemble learning algorithms to develop QSAR models on bioconcentration factors of organic chemicals in multispecies fish

DING Rui[1] *CHEN Jingwen*[1]** *YU Yang*[2] *LIN Jun*[2] *WANG Zhongyu*[1] *TANG Weihao*[1] *LI Xuehua*[1]

（1. Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), Dalian Key Laboratory on Chemicals Risk Control and Pollution Prevention Technology, School of Environmental Science and Technology, Dalian University of Technology, Dalian, 116024, China；2. Solid Waste and Chemicals Management Center, Ministry of Ecology and Environment, Beijing, 100029, China）

Abstract Bioconcentration factor (BCF) is a key parameter characterizing bioaccumulation of chemicals in organisms. Nevertheless, only around one thousand chemicals have BCF values, in contrast to over 350 000 chemicals that have been registered for production and application in the global market. Quantitative structure-activity relationship (QSAR) models are regarded as an

2021 年 1 月 13 日收稿(Received：January 13，2021).

* 国家重点研究发展计划 (2018YFC1801604, 2018YFE0110700) 和国家自然科学基金 (21661142001) 资助.
Supported by the National Key Research and Development Program (2018YFC1801604, 2018YFE0110700) and the National Natural Science Foundation of China (21661142001).

** 通讯联系人 **Corresponding author**，Tel：0411-84706269，E-mail：jwchen@dlut.edu.cn

efficient method to fill the data gap. However, majority of QSAR models on BCF are individual models, while ensemble models may have improved capabilities on BCF prediction. In this study, a comprehensive fish BCF database was constructed, covering empirical BCF values of more than 1 300 organic chemicals. Based on the database, 5 individual QSAR models and 11 ensemble models were developed on BCF of organic compounds in fish using machine learning algorithms, following the guidelines on development and validation of QSARs proposed by the OECD. Results show the ensemble models have better goodness-of-fit, robustness, predictability and wider application domain than the individual models. The optimum ensemble model was further employed to predict BCF for chemicals in the inventory of existing chemical substances of China (IECSC), showing that 1 066 chemicals in the inventory are bioaccumulative, and 86 chemicals are very bioaccumulative. The models can provide necessary data for evaluating the bioaccumulation capacity of chemicals and support sound chemicals management.

Keywords bioconcentration factor (BCF), quantitative structure-activity relationship (QSAR), machine learning, ensemble model, applicability domain.

人工合成的有机化学品 (如杀虫剂、药物和各种工业化学品) 在促进社会发展、改善人类生活质量方面发挥了重要作用. Wang 等[1] 近期统计, 目前全球市场上使用的化学品数量已达 35 万种. 这些化学品在其整个生命周期中, 都可能被释放到环境中, 威胁生态系统和人类健康[1－2]. 具有持久性 (persistence)、生物累积性 (bioaccumulation)、毒性 (toxicity) 的化学品, 已经成为影响人体与生态健康的重要风险源[3－4]. 我国《新化学物质环境管理登记指南》中明确规定应当重点管控具有 PBT 属性的化学物质[5]. 其中, 生物累积是指生物从环境和膳食 (含吞食低营养级生物) 中积累化学物质, 使其体内该化学物质的浓度超过周围环境中浓度的现象[6]. 生物富集作为生物累积的类型之一, 是指生物从周围环境中摄取某种化学物质, 使其体内浓度超过周围环境中浓度的现象[6]. 生物富集常用生物富集因子 (BCF) 来表征, BCF 为化学物质在生物体内的浓度与其在环境介质中平衡浓度之比[7]. 欧盟化学品注册、评估、许可和限制 (REACH) 法规规定, BCF 是筛查生物累积性物质的重要指标之一[8].

鱼类是水生态系统的关键物种, 其体内污染物的积累程度对其他生物、甚至人类健康具有重要影响[9]. 传统上, 鱼体 BCF 的测定, 可遵循经济合作与发展组织 (OECD) 发布的“流水式鱼类生物富集测试指南 (OECD 指南 305)”[10]. 通过该方法, 虽可测得一些化学品的 BCF 数据, 但存在测试周期长、费用高、动物实验伦理等问题, 无法满足对大量商用化学品进行风险管理的现实需求[9]. 因此, 需要发展快速高效的替代方法来获取 BCF 数据.

定量构效关系 (QSAR) 模型, 作为计算毒理学技术的核心内容, 可以快速高通量地获取化学品环境暴露与危害性的相关信息[11]. QSAR 通过函数或映射关系将分子结构描述符 (描述分子结构特征的参数) 和预测终点联系起来[11]. 早期 BCF 的 QSAR 预测模型, 主要基于分子的理化参数、碎片参数、溶剂化参数等物理意义明确的描述符而构建, 多为线性模型[12－14]. 近年来, 各种机器学习算法被用于 QSAR 模型的构建[15－18]. 2019 年, Miller 等[19] 建立并比较了 24 种可用于预测 BCF 的线性模型 (如最小二乘回归、偏最小二乘回归和岭回归) 和非线性模型 (如随机森林、支持向量机和多层感知机), 发现大多数非线性模型对 BCF 的预测效果比线性模型好.

随着机器学习算法不断发展, 集成模型出现并得到应用. 集成模型通过投票法、平均法或学习法将多个单独模型的信息整合在一起, 有望产生更准确、更稳健的预测结果[20－22]. Valsecchi 等[20] 发现, 相对于单一模型, 集成模型具有减少预测不确定性、拓宽模型应用域等优点; Li 等[21] 发现集成模型能够增加模型多样性并减少过拟合. 集成模型在预测化学品毒性方面已有应用, 如鱼类半数致死浓度 (LC_{50}) 和无观测效应浓度 (NOEC) 的集成模型等[22]. 然而, 关于 BCF 的集成模型研究还不多见.

本研究搜集整理鱼体 BCF 数据并构建了数据库, 计算了 4 000 多种分子描述符, 选择 5 种机器学习算法建立了预测 BCF 的单一模型, 进而构建了集成模型. 依据 OECD 关于 QSAR 模型构建和验证的导则[23], 评价了模型的稳健性和预测能力, 并进行了应用域表征.

1　材料与方法 (Materials and methods)

1.1　数据库构建

从文献 [24 − 25] 和数据库 (NITE[26], CEFIC LRI[27], DSL[28], ECOTOX EPA[29], OECD Toolbox[30] 和 ECHA[31]) 中, 搜集有机化学品在不同种类鱼体的 BCF 测定值. 按以下规则对原始数据进行处理[25]: (1) 去除无机物、混合物以及金属配合物等; (2) 当 BCF 值不以 L·kg^{-1} 为单位, 不以鱼体全身测量为基础计算, 或不是在 OECD 推荐的物种 [鲤鱼 (*Cyprinus carpio*)、虹鳟鱼 (*Oncorhynchus mykiss*)、黑头呆鱼 (*Pimephales promelas*)、青鳉鱼 (*Oryzias latipes*)、斑马鱼 (*Danio rerio*)、蓝绿鳞鳃太阳鱼 (*Lepomis macrochirus*)、孔雀鱼 (*Poecilia reticulata*)、三刺鱼 (*Gasterosteus aculeatus*)] 上进行测试时, 则排除该值; (3) 当同一化合物有多个实测数据时, 取中值, 中值根据 ISO16269-7 规范计算得到[32]; (4) 确保每个化合物都有 CAS 号和 SMILES 码与之对应. 经过整理, 最终得到 1 384 种有机化学品在不同种类鱼体的 BCF 实测值 (单位为 L·kg^{-1}). 基于线性自由能关系的 QSAR 原理, 将 BCF 实测值以 10 为底取对数转换为 lgBCF, 作为预测终点[11].

1.2　分子描述符计算与筛选

使用 Dragon 6.0 软件计算分子结构描述符, 得到 4 885 种不同类型的描述符[33]. 为了使各描述符尺度处于同一数量级, 对其进行标准差法标准化处理[34]. 然后对描述符进行初步筛选: 去掉至少有一个缺失值的描述符, 去掉为常数的描述符. 以筛选得到的描述符为自变量, lgBCF 为因变量, 使用逐步回归分析构建多元线性回归模型. 去除方差膨胀因子大于 5, 显著性水平大于 0.001 的模型, 确保建模描述符之间不存在多重共线性且模型具有统计学意义[35]. 综合考虑经自由度调整后的决定系数和自变量个数 (通常应不超过样本个数的 1/5, 以避免过拟合), 确定用于构建 QSAR 模型的分子结构描述符.

1.3　模型构建与表征

将 1 384 个化合物以 4∶1 的比例随机划分为训练集 (1 107 个化合物) 和验证集 (277 个化合物), 训练集用于构建模型, 验证集用于对模型进行外部验证, 详细数据见附件.

综合考虑机器学习算法对数据的适应能力以及类型的多样性, 选择普通最小二乘 (OLS)[36]、支持向量机 (SVM)[37]、随机森林 (RF)[38 − 39]、梯度提升决策树 (GBDT)[40] 和极端梯度提升 (XGBoost)[41] 这 5 种算法, 先构建预测 BCF 的单一模型, 进而构建集成模型. 使用网格搜索交叉验证来调整模型参数, 确定最优模型[42]. 模型信息和相关参数见附件.

使用堆叠 (Stack) 方法构建集成模型[21 − 22, 43 − 46]. 堆叠集成模型通常包含两层, 第一层使用两个或两个以上模型对终点分别进行预测, 这些模型称为基学习器 (Base-learner), 为充分学习训练数据, 基学习器一般选择非线性模型[21 − 22, 46]; 第二层只有一个模型, 负责将第一层模型的预测结果进行融合, 称为元学习器 (Meta-learner), 为降低模型过拟合风险, 元学习器一般选择线性模型[22, 45]. 本研究将训练好的 SVM 模型、RF 模型、GBDT 模型、XGBoost 模型随机组合作为基学习器, OLS 模型作为元学习器构建堆叠集成模型.

使用经自由度调整后的决定系数 (R^2_{adj})、均方根误差 (RMSE) 以及 10 折交叉验证系数 ($Q^2_{10\text{-fold}}$) 评价模型效果[47 − 48]. 训练集R^2_{adj}, RMSE 表征模型拟合优度; 验证集的R^2_{adj}, RMSE 表征模型预测能力; 训练集的$Q^2_{10\text{-fold}}$表征模型稳健性[35]. 模型的应用域表征采用 Williams 图, 即化合物的杠杆值 (h_i) 对标准残差 (δ) 作图[35, 47]. 相关计算公式详述于附件, 相关计算采用 Python3.7.0 软件实现[49].

2　结果与讨论 (Results and discussion)

经检索相关文献和数据库[12 − 19, 24 − 31, 35], 搜集得到 1 384 个有机化学品在不同种类鱼体的 BCF 实测值 (单位为 L·kg^{-1}), 构建了全面的鱼体 BCF 数据库, 其详细信息见附件.

2.1　描述符筛选结果

经初步筛选和逐步回归分析, 最终选择 8 个分子结构描述符 (D_1, D_2, ⋯, D_8) 用于构建模型, 其相关信息列于表 1 中.

表 1 分子描述符的类型及含义

Table 1 Type and description of the molecular descriptors

编号 Index	OLS模型中对应系数 Coefficient in OLS model	描述符名称 Descriptor name	类型及含义 Type and description
D_1	−0.933	BLTF96	与正辛醇/水分配系数相关的基本描述符
D_2	−0.438	SpPosA_Dz(m)	相对分子质量加权的2D矩阵描述符
D_3	0.342	Cl-089	与C(sp^2)相连的Cl原子中心碎片描述符
D_4	−0.325	SpMax1_Bh(s)	与分子中原子连接相关的2D矩阵描述符
D_5	0.217	B07[C-C]	表示拓扑距离7处是否存在C—C结构的2D原子对描述符
D_6	0.317	F02[C-O]	描述拓扑距离2处C—O结构出现频率的2D原子对描述符
D_7	−0.130	B04[O-Cl]	表示拓扑距离4处是否存在O—Cl结构的2D原子对描述符
D_8	−0.216	ATSC7m	相对分子质量加权的2D自相关描述符

2.2 模型构建结果

单一模型的相关统计参数汇总于表 2 中. 如表 2 所示, 线性模型 (OLS 模型) 在预测生物累积性这种复杂生物过程时误差较大, 非线性模型 (SVM 模型、RF 模型、GBDT 模型、XGBoost 模型) 的预测效果则有较大提升. 其中, RF 模型的$Q^2_{\text{10-fold}}$值最大, 模型最稳健; SVM 模型的$R^2_{\text{adj-test}}$值最大, 预测准确性最好.

表 2 单一模型相关统计参数汇总

Table 2 Summary of statistical parameters of individual models

Model	$R^2_{\text{adj-train}}$	$R^2_{\text{adj-test}}$	$Q^2_{\text{10-fold}}$	$\text{RMSE}_{\text{train}}$	$\text{RMSE}_{\text{test}}$
OLS	0.596	0.615	0.573	0.916	0.933
SVM	0.732	0.758	0.684	0.746	0.741
RF	0.839	0.751	0.700	0.579	0.751
GBDT	0.845	0.732	0.694	0.568	0.779
XGBoost	0.859	0.754	0.697	0.541	0.747

集成模型的相关统计参数汇总于表 3 中. 如表 3 所示, 多数集成模型的稳健性和准确性都比单一模型有提升. 图 1 比较了集成模型和稳健性、预测性最好的单一模型的$Q^2_{\text{10-fold}}$, $R^2_{\text{adj-test}}$值. 从图 1 可见, 虽然 Stack-6 模型的$Q^2_{\text{10-fold}}$值在所有集成模型中最低, 但仍与 RF 模型的稳健程度相当; 多数集成模型 (除 Stack-4, Stack-5, Stack-6, Stack-9 模型外) 的$R^2_{\text{adj-test}}$值高于 SVM 模型. 如表 3 所示, 使用不同类型基学习器的模型效果优于使用同种类型基学习器的模型效果.

表 3 集成模型相关统计参数汇总

Table 3 Summary of statistical parameters of ensemble models

Model	Base-learner	$R^2_{\text{adj-train}}$	$R^2_{\text{adj-test}}$	$Q^2_{\text{10-fold}}$	$\text{RMSE}_{\text{train}}$	$\text{RMSE}_{\text{test}}$
Stack-1	SVM, RF	0.800	0.766	0.706	0.644	0.728
Stack-2	SVM, XGBoost	0.808	0.769	0.707	0.632	0.723
Stack-3	SVM, GBDT	0.801	0.764	0.707	0.642	0.730
Stack-4	RF, XGBoost	0.855	0.756	0.703	0.548	0.744
Stack-5	RF, GBDT	0.849	0.745	0.702	0.559	0.760
Stack-6	XGBoost, GBDT	0.859	0.752	0.699	0.541	0.750
Stack-7	SVM, RF, XGBoost	0.821	0.770	0.708	0.610	0.723
Stack-8	SVM, RF, GBDT	0.815	0.764	0.708	0.620	0.731
Stack-9	RF, XGBoost,GBDT	0.856	0.755	0.703	0.547	0.745
Stack-10	SVM, XGBoost, GBDT	0.823	0.762	0.708	0.606	0.734
Stack-11	SVM, RF, XGBoost,GBDT	0.830	0.767	0.708	0.595	0.726

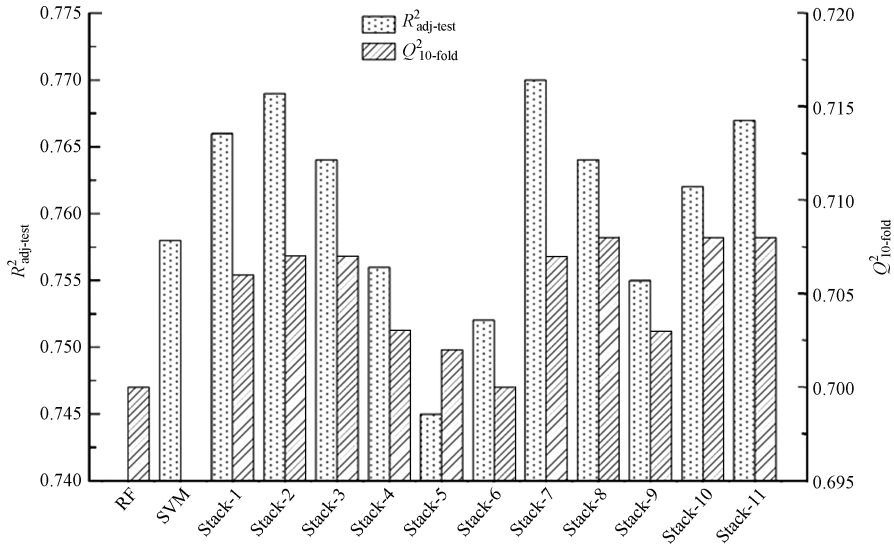


图 1　集成模型和单一模型的效果对比图

Fig.1　Comparison of performances between ensemble models and individual models

综合考虑各项评价指标, Stack-7 模型 (基学习器为 SVM, RF 和 XGBoost 模型, 元学习器为 OLS 模型) 在所有 11 个集成模型中表现最佳. Stack-7 模型具有最高$R^2_{adj\text{-}test}$和最低 $RMSE_{test}$. Roy 等[50] 建议评估 QSAR 模型的预测能力还应考虑以下标准:

(1) 若 MAE ≤ 0.10 × TR 并且 MAE + 3σ ≤ 0.20 × TR, 则模型具有良好的预测能力;

(2) 若 MAE > 0.15 × TR 或者 MAE + 3σ > 0.25 × TR, 则模型预测能力较差;

(3) 若不满足上述两个条件, 则模型预测能力中等.

MAE 表示验证集平均绝对误差; σ 值表示验证集数据绝对误差值的标准偏差; TR 为训练集数值范围. 按此方法评价 Stack-7 模型的预测能力, 训练集中 lgBCF 值范围为−1.22—6.60, TR = 7.82. 剔除验证集中 5% 高预测误差点后, 其 MAE = 0.482, σ = 0.351, MAE + 3σ = 1.535, 满足前述第一条标准, 故 Stack-7 模型预测能力良好, 选为最优模型, 该模型的 lgBCF 实测值和预测值拟合图如图 2a 所示.

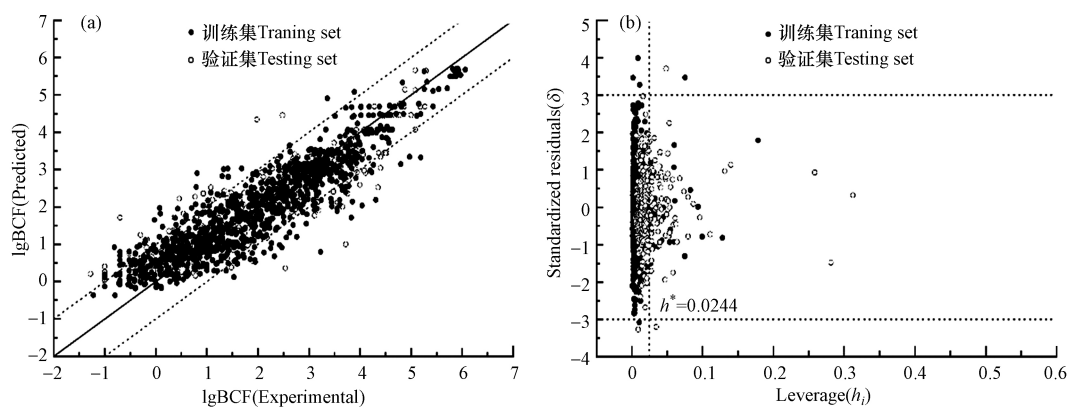


图 2　Stack-7 模型的 lgBCF 实测值/预测值拟合图 (a) 和表征应用域的 Williams 图 (b)

Fig.2　Plot of predicted versus observed lgBCF values (the left one) and Williams plot of Stack-7 model for applicability domain characterization (the right one)

2.3　最优模型误差分析

预测与实测值之间的差为预测残差, 主要由随机误差和系统误差两部分构成[51]. 随机误差由随机因素 (比如训练数据的扰动) 引起, 具有互相抵偿性; 系统误差通常来自算法本身, 会造成预测结果向特定方向偏离[52]. Roy 等[52] 认为, 如果模型满足以下条件之一, 则很可能出现系统误差:

(1) nPE/ nNE > 5 或者 nNE/ nPE > 5;

(2) ABS(MPE/MNE) > 2 或者 ABS(MNE/MPE) > 2;

(3) AAE − ABS(AE) < 0.5 × AAE;

(4) R^2(i^{th} vs $(i-1)^{th}$ residuals) > 0.5;

(5) R^2(Y vs residuals) > 0.5;

AE 为平均残差; AAE 为平均绝对残差; MPE 为平均正残差; MNE 为平均负残差; *n*PE 为正残差个数; *n*NE 为负残差个数; R^2(i^{th} vs $(i-1)^{th}$ residuals) 表示按实测值的递增对残差进行排序, 第 i 个残差值与第 $i-1$ 个残差值之间的相关性; R^2(Y vs residuals) 表示预测值和残差值之间的相关性. 基于上述标准对最优模型进行了评价, 相关评价指标值汇总在表 4 中. 结果表明, 上述 5 项条件预测误差均不满足, 说明最优模型不存在系统误差.

表 4 验证集预测误差的评价指标

Table 4 Evaluation indices of prediction errors from testing set

Data set	AE	AAE	MPE	MNE	*n*PE	*n*NE
Testing set	−0.010	0.551	0.575	−0.531	130	147

2.4 机理分析

一般认为生物富集过程实际上是有机化合物在水相和有机相的分配过程, 疏水性是生物富集过程中的主要驱动力[53], 因此根据疏水性参数可以较好地预测生物富集参数. 正辛醇/水分配系数 (K_{OW}) 常被用于预测 BCF[12−14]. Veith 等[12] 曾建立鱼类 lgBCF 与 lgK_{OW} 的线性模型, 模型 R^2 = 0.90, 但模型只包含 55 个疏水性化合物. BLTF96 是与疏水性参数 K_{OW} 相关的基本描述符. 本研究尝试仅使用了 BLTF96 描述符对数据库中 1384 个有机物建立线性模型, R^2 = 0.40, 说明仅靠疏水性这一性质难以准确估计数据库中大量化学品的 lgBCF 值.

表 1 汇总了通过逐步回归分析得到的 8 个分子结构描述符的含义、类型以及它们在线性模型中对应的系数. BLTF96 的系数绝对值明显大于其他描述符. SpPosA_Dz(m) 和 ATSC7m 都是与相对分子质量相关的 2D 描述符. Lipinski[54] 发现, 相对分子质量小于 500 的小分子药物更容易被生物体吸收. Strempel 等[16] 也发现相对分子质量以及分子直径对生物累积性有重要影响. 综上, 分子的疏水性对生物累积性影响最为显著, 其次为相对分子质量和分子大小.

2.5 应用域表征

使用 Williams 图对 Stack-7 模型的应用域进行表征, 以确定集成模型的适用化合物范围. 如图 2b 所示, 横坐标表示杠杆值 (h_i), 纵坐标表示标准残差 (δ). 警戒杠杆值 (h^*) 为 0.0244, 认为 $h_i \le h^*$时的化合物适用于本模型; 当 $h_i > h^*$时, 认为该化合物超出训练集定义的描述符范围, 称其为模型的应用域外化合物. 模型方法的预测能力高度依赖于模型的应用域, 对于应用域内的化合物预测准确性较高, 而对于域外化合物的预测则存在较大不确定性[55]. 当化合物的 δ 值落在 (−3.0, +3.0) 以外时, 认为该点是离群点. $h_i > h^*$的化合物其 δ 值仍落在 (−3.0, +3.0) 以内, 说明模型具有一定的延展性[35]. Stack-7 模型的训练集和验证集中共有 5 个化合物 (CAS 号分别为 81-88-9, 112-27-6, 117-80-6, 14233-37-5, 4901-51-3) 的$|\delta| > 3$ 且 $h_i \le h^*$, 这些化合物为模型应用域内的离群点.

2,3,4,5-四氯苯酚 (CAS 号: 4901-51-3)、9-(2-羧基苯基)-3,6-双 (二乙氨基) 占吨翁氯化物 (CAS 号: 81-88-9) 的 lgBCF 预测值被高估. 二者都含有可解离基团 (酚羟基、羧基), 其酸解离常数 (p*K*a) 分别为 6.36 和 3.22[56−57]. 在 pH 值约为 7 — 9 范围的水环境中, 这两种物质均会以阴离子形态存在, 通常离子态比其中性分子更难通过生物膜而被生物富集[58−59], 所以实验测定的 BCF 会比仅考虑分子形态的预测值低. 在将来关于 BCF 的 QSAR 预测模型构建中, 应该考虑分子的解离形态[60−61].

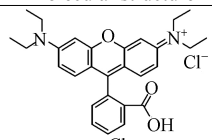
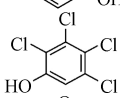
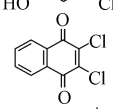
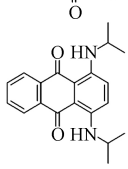
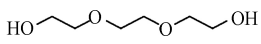
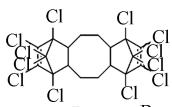
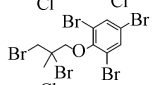
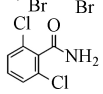
2,3-二氯-1,4-萘醌 (CAS 号: 117-80-6) 和 1,4-二 (1-异丙胺基) 蒽醌 (CAS 号: 14233-37-5) 的 lgBCF 预测值被低估. 二者作为醌类化合物, 容易发生亲电加成, 还原生成二元酚[62−63]. 有研究发现, 萘醌类化合物能够与生物亲核试剂发生反应, 而生物体内的白蛋白是一种普遍存在的亲核试剂, 可与含有至少一个未取代的醌碳的萘醌类化合物发生结合[64]. 这类物质进入生物体后, 不仅会在脂质中发生富集, 还可能与蛋白质等非脂肪组织发生特定相互作用, 从而造成实验测定值高于预测值的现象.

此外, 有 3 个化合物 (CAS 号分别为 2008-58-4, 13560-89-9, 36065-30-2) 的$|\delta| > 3$, 但它们落在模型

应用域外 ($h_i > h^*$), 因此模型对其预测的不确定性较大是可以理解的. 表 5 列出了上述离群点以及域外化合物的分子结构、标准偏差等.

表 5　Stack-7 模型离群点及域外化合物

Table 5　Outliers and out-of-domain compounds in Stack-7 model

CAS	中文名称 Chinese name	标准残差 Standardized residual	分子结构 Molecular structure
81-88-9	9-(2-羧基苯基)-3,6-双(二乙氨基)占吨翁氯化物	−3.300	
4901−51-3	2,3,4,5-四氯苯酚	−3.118	
117-80-6	2,3-二氯-1,4-萘醌	3.305	
14233−37-5	1,4-二(1-异丙胺基)蒽醌	3.493	
112-27-6	三甘醇	4.027	
13560−89-9	双(六氯环戊二烯)环辛烷	−3.228	
36065−30-2	2,4,6-三溴苯基(2,3-二溴-2-甲基丙基)醚	3.501	
2008-58-4	2,6-二氯苯甲酰胺	3.734	

2.6　模型比较

关于预测鱼类 lgBCF 的集成模型研究还较少. Zhao 等[65] 使用普通最小二乘回归 (OLS), 径向基函数神经网络 (RBF-NN) 和支持向量机 (SVM) 方法, 基于 473 种有机化学品的 lgBCF 数据集建立了多个 QSAR 模型, 并将两个使用不同描述符的 RBF 模型组合成一个集成模型. Gissi 等[66] 将两个常用预测 BCF 的 QSAR 模型按照特定规则进行了集成, 其一为上述 Zhao 等建立的集成模型; 另一为基于分子碎片的模型, 该模型使用实验测得或预测的 $\lg K_{OW}$ 作为唯一的描述符, 并增加特定结构碎片相关的校正因子对模型进行校正[13].

表 6 比较了本研究的集成模型与上述集成模型. 从该表可以看出, 本研究所构建的 Stack-7 模型在保证预测效果的基础上, 应用范围更加广泛, 模型表征更加严格, 集成策略更加简洁, 因此其在化学品 BCF 预测中的应用潜力更大.

表 6　本研究与其他集成模型的比较

Table 6　Comparison of the current model with previous ensemble models

模型 Model	描述符个数 $n_{descriptors}$	总数据量 n_{all}	训练集数据量n_{train}	R^2_{train}	$RMSE_{train}$	验证集数据量n_{test}	R^2_{test}	$RMSE_{test}$	交叉验证 Cross validation	应用域 Application domain
Zhao等[62]	8	473	378	0.830	0.560	95	0.800	0.590	有	—
Gissi等[63]	9	851	851	0.800	0.610	—	—	—	—	有
本研究	8	1384	1107	0.821	0.610	277	0.770	0.723	有	有

2.7　模型应用

利用 Stack-7 模型对《中国现有化学物质名录》(IECSC) 中化学物质的 lgBCF 进行了初步预测[67]. IECSC 中 21677 种化学物质含有分子 SMILES 码, 首先根据分子结构计算了所需 8 种描述符 (BLTF96,

Cl-089, SpPosA_Dz(m), SpMax1_Bh(s), B07[C-C], F02[C-O], B04[O-Cl], ATSC7m) 的数值，然后计算了每种化学物质的 h_i 值，确定有 21 174 种在 Stack-7 模型应用域范围内. 一般认为，当 BCF ≥ 2000(即 lgBCF ≥ 3.3) 时，该物质具有生物累积性；BCF ≥ 5 000(即 lgBCF ≥ 3.7) 时，具有强生物累积性[68]. IECSC 中 21 174 种化学物质的 lgBCF 预测值分布如图 3，其中 1 066 种化学物质具有生物累积性，86 种化学物质具有强生物累积性，该预测结果可为化学品风险评价与管理工作提供参考. IECSC 中化学品 lgBCF 预测值具体数据见附件.

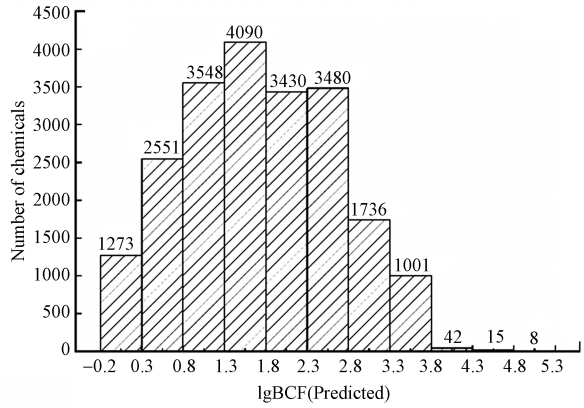


图 3 《中国现有化学物质名录》中 (21 174 种) 化学品 lgBCF 预测值分布图

Fig.3 Distribution of predicted lgBCF values for chemicals (21 174 molecules) included in the inventory of existing chemical substances of China

3 结论 (Conclusion)

本研究使用 OLS, RF, SVM, GBDT 和 XGBoost 建立了预测有机化学品鱼体 BCF 的 QSAR 模型，并进一步构建了堆叠集成模型. 依照 QSAR 模型构建和验证导则，对集成模型进行了评价和应用域表征. 结果表明，集成模型比单一模型的预测准确性更高，更稳健；相较以往研究，本研究所建集成模型应用域更广泛. 按照我国《新化学物质环境管理登记指南》中关于 QSAR 模型构建和使用的要求，进一步利用集成模型对《中国现有化学物质名录》中两万余种化学物质的 lgBCF 值进行了初步预测，预测结果可为化学品风险评价与管理工作提供参考. 此外，本研究还建立了关于有机化学品鱼类 BCF 实测值数据库，有助于后续相关研究和应用工作的开展.

参考文献 (References)

[1] WANG Z, WALKER G W, MUIRD C G, et al. Toward a global understanding of chemical pollution: A first comprehensive analysis of national and regional chemical inventories [J]. Environmental Science & Technology, 2020, 54(5): 2575-2584.

[2] Global Chemicals Outlook II: From legacies to innovative solutions: Implementing the 2030 agenda for sustainable development-Synthesis report[M]. Nairobi: United Nations Environment Programme, 2019: 1-88.

[3] KEITA-QUANE F. UNEP Chemicals' work: breaking the barriers to information access [J]. Toxicology, 2003, 190(1-2): 135-139.

[4] 罗孝俊, 麦碧娴. 新型持久性有机污染物的生物富集[M]. 北京: 科学出版社, 2017: 1-356.
LUO X J, MAI B X. Bioaccumulation of emergying persistent organic pollutants[M]. Beijing: Science Press, 2017: 1-356(in Chinese).

[5] 中华人民共和国生态环境部, 新化学物质环境管理登记指南[R]. 北京, 2020: 1-81.
Ministry of Ecology and Environment of the People's Republic of China, Guidelines for environmental management registration of new chemical substances[R]. Beijing, 2020: 1-81(in Chinese).

[6] 陈景文, 全燮. 环境化学[M]. 大连: 大连理工大学出版社, 2009: 1-387.
CHEN J W, QUAN X. Environmental chemistry[M]. Dalian: Dalian University of Technology Press, 2009: 1-387(in Chinese).

[7] GOBAS F A, WOLF W D, BURKHARD L P, et al. Revisiting bioaccumulation criteria for POPs and PBT assessments [J]. Integrated Environmental Assessment and Management: An International Journal, 2010, 5(4): 624-637.

[8] EU. Regulation(EC) No. 1907/2006 of the European parliament and of the council of 18 December 2006 concerning the registration, evaluation, authorization, and restriction of chemicals(REACH)[S]. Brussels: Official Journal of the EU, 2006.

[9] WOLF W D, COMBER M, DOUBENP, et al. Animal use replacement, reduction, and refinement: Development of an integrated testing strategy for bioconcentration of chemicals in fish [J]. Integrated Environmenta lAssessment and Management, 2007, 3(1): 3-17.

[10] OECD. OECD guideline for testing of chemicals 305: Bioconcentration: Flow-through fish test[R]. Paris, 1996: 1-23.

[11] 陈景文, 王中钰, 傅志强. 环境计算化学与毒理学[M]. 北京: 科学出版社, 2018: 1-274.
CHEN J W, WANG Z Y, FU Z Q. Environmental computational chemistry and toxicology[M]. Beijing: Science Press, 2018: 1-274(in Chinese).

[12] VEITH G D, DEFOE D L, BERGSTEDT B V. Measuring and estimating the bioconcentration factor of chemicals in fish [J]. Journal of the Fisheries Board of Canada, 1979, 36(9): 1040-1048.

[13] MEYLAN W M, HOWARD P H, BOETHLING R S, et al. Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient [J]. Environmental Toxicology and Chemistry, 1999, 18(4): 664-672.

[14] PAVAN M, NETZEVA T I, WORTH A P. Review of literature-based quantitative structure–activity relationship models for bioconcentration [J]. QSAR & Combinatorial Science, 2008, 27: 21-31.

[15] DEARDEN J C, HEWITT M. QSAR modelling of bioconcentration factor using hydrophobicity, hydrogen bonding and topological descriptors [J]. SAR and QSAR in Environmental Research, 2010, 21(7/8): 671-680.

[16] STREMPEL S, NENDZA M, SCHERINGER M, et al. Using conditional inference trees and random forests to predict the bioaccumulation potential of organic chemicals [J]. Environmental Toxicology and Chemistry, 2013, 32(5): 1187-1195.

[17] YUAN J, XIE C, ZHANG T, et al. Linear and nonlinear models for predicting fish bioconcentration factors for pesticides [J]. Chemosphere, 2016, 156: 334-340.

[18] AI H X, WU X W, ZHANG L, et al. QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods [J]. Ecotoxicology and Environmental Safety, 2019, 179: 71-78.

[19] MILLER T H, GALLIDABINO M D, MACRAE J I, et al. Prediction of bioconcentration factors in fish and invertebrates using machine learning [J]. Science of the Total Environment, 2019, 648: 80-89.

[20] VALSECCHI C, GRISONI F, CONSONNI V, et al. Consensus versus individual QSARs in classification: Comparison on a large-scale case study [J]. Journal of Chemical Information and Modeling, 2020, 60(3): 1215-1223.

[21] LI X, KLEINSTREUER N C, FOURCHES D. Hierarchical quantitative structure–activity relationship modeling approach for integrating binary, multiclass and regression models of acute oral systemic toxicity [J]. Chemical Research in Toxicology, 2020, 33(2): 353-366.

[22] SHEFFIELD T Y, JUDSON R S. Ensemble QSAR modeling to predict multispecies fish toxicity lethal concentrations and points of departure [J]. Environmental Science & Technology, 2019, 53(21): 12793-12802.

[23] OECD. Guideline document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models. Environment Health and Safety Publications Series on Testing and Assessment No. 69[R]. Paris: OECD, 2007: 1-154.

[24] ARNOT J A, GOBAS F A. A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms [J]. Environmental Reviews, 2006, 14(4): 257-297.

[25] LUNGHINI F, MARCOU G, AZAM P, et al. QSPR models for bioconcentration factor (BCF): Are they able to predict data of industrial interest? [J]. SAR and QSAR in Environmental Research, 2019, 30(7): 507-524.

[26] NITE (Japanese National Institute of Technology and Evaluation). Data from: Biodegradation and bioconcentration data under CSCL National Institute of Technology and Evaluation [DB/OL]. [2020-01-12]. https://www.nite.go.jp/en/index.html.

[27] CEFIC LRI (European Chemical Industry Council Long Range Initiative). Data from: Bioconcentration factor database, European Chemical Industry Council Long range research initiative [DB/OL]. [2020-01-12]. http://cefic-lri.org/.

[28] DSL (Canadian Domestic Substance List). Data from: Canadian domestic substances list (DSL), Environment and Climate Change Canada [DB/OL]. [2020-01-12]. https://www.canada.ca/en/environment-climate-change/services/canadian-environmental-protection-act-registry/substances-list.html#toc0.

[29] ECOTOX EPA (ECOTOXicology knowledgebase of the US Environmental Protection Agency). Data from: ECOTOX Knowledgebase, US Environmental Protection Agency [DB/OL]. [2020-01-12]. https://cfpub.epa.gov/ecotox/.

[30] QSAR Toolbox v 4.1. OASIS Laboratory of mathematical chemistry, Burgas, BG [DB/OL]. [2020-01-12]. http://oasis-lmc.org/products/software/toolbox.aspx.

[31] OECD (Organisation for Economic Co-Operation and Development). Data from: EChemPortal: Global portal to information on chemical substances, Organisation for Economic Co-operation Development [DB/OL]. [2020-01-12]. https://www.echemportal.org/echemportal/.

[32] ISO16269-7-2001, Statistical interpretation of data. Part 7: Median; Estimation and confidence intervals[S]. Geneva: International Organization for Standardization, 2001.

[33] DRAGON(SoftwareforMolecularDescriptorCalculation), Version 6.0[CP], 2012. http://www.talete.mi.it/.

[34] SINGH B K, VERMA K, THOKE A S. Investigations on impact of feature normalization techniques on classifier's performance in breast tumor classification [J]. International Journal of Computer Applications, 2015, 116(19): 11-15.

[35] 郑玉婷. 有机化学品鱼类生物富集因子QSAR模型的构建[D]. 大连: 大连理工大学, 2014: 1-60.
ZHENG Y T. Development of QSAR models on bioconcentration factors of chemicals in fish[D]. Dalian: Dalian University of Technology, 2014: 1-60(in Chinese).

[36] NATHANS L L, OSWALDF L, NIMON K. Interpreting multiple linear regression: A guidebook of variable importance [J]. Practical Assessment, Research, and Evaluation, 2012, 17(1): 1-19.

[37] CORTES C, VAPNIK V. Support-vector networks [J]. Machine Learning, 1995(20): 273-297.

[38] BREIMAN L. Random forests [J]. Machine Learning, 2001(45): 5-32.

[39] ATHEY S, TIBSHIRANI J, WAGER S. Generalized random forests [J]. Annals of Statistics, 2019, 47(2): 1148-1178.

[40] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine [J]. Annals of Statistics, 2001, 29(5): 1189-1232.

[41] CHEN T Q, GUESTRIN C. Xgboost: A scalable tree boosting system//Assoc Comp Machinery. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining[C]. 2016: 785-794.

[42] VANDERPLAS J. Python data science handbook[M]. Sevastopol: O'Reilly Media Inc, 2018: 1-500.
[43] WOLPERT D H. Stacked generalization [J]. Neural Networks, 1992, 5(2): 241-259.
[44] BREIMAN L. Stacked regressions [J]. Machine Learning, 1996, 24(1): 49-64.
[45] ZENKO B, DZEROSKI S. Stacking with an extended set of meta-level attributes and MLR[A]. In: Elomaa T, Mannila H, et al. 13th European Conference on Machine Learning[C]. Springer, Berlin, Heidelberg, 2002: 493-504.
[46] SHARMA A, RANI R. Drug sensitivity prediction framework using ensemble and multi-task learning [J]. International Journal of Machine Learning and Cybernetics, 2020, 11(3): 1-10.
[47] GRAMATICA P. Principles of QSAR models validation: internal and external [J]. QSAR & Combinatorial Science, 2007, 26(5): 694-701.
[48] 覃礼堂, 刘树深, 肖乾芬, 等. QSAR模型内部和外部验证方法综述 [J]. 环境化学, 2013, 32(7): 1205-1211.
QIN L T, LIU S S, XIAO Q F, et al. Internal and external validations of QSAR model: Review [J]. Environmental Chemistry, 2013, 32(7): 1205-1211(in Chinese).
[49] Python, Version 3.7. 0[CP]. https://www.python.org/downloads/release/python-370/.
[50] ROY K, DAS R N, AMBURE P, et al. Be aware of error measures. Further studies on validation of predictive QSAR models [J]. Chemometrics and Intelligent Laboratory Systems, 2016, 152: 18-33.
[51] LARSEN R J, MARX M L. An introduction to mathematical statistics and its applications[M]. Upper Saddle River: Prentice-Hall Inc, 1981: 1-920.
[52] ROY K, AMBURE P, AHER R B. How important is to detect systematic error in predictions and understand statistical applicability domain of QSAR models? [J]. Chemometrics & Intelligent Laboratory Systems, 2017, 162: 44-54.
[53] 闻洋. 有机污染物生物富集与鱼体内临界浓度关系的研究[D]. 长春: 东北师范大学, 2015: 1-126.
WEN Y. Relationship between bioconcentration and critical body residues of organic pollutants[D]. Changchun: Northeast Normal University, 2015, 1-126(in Chinese).
[54] TICE C M. Selecting the right compounds for screening: does Lipinski's Rule of 5 for pharmaceuticals apply to agrochemicals? [J]. Pest Management Science: formerly Pesticide Science, 2001, 57(1): 3-16.
[55] 李超. 有机污染物与·OH气相反应动力学和机制的计算模拟预测[D]. 大连: 大连理工大学, 2015: 1-211.
LI C. Computational simulation to predict gaseous reaction kinetics and mechanism of organic pollutants with·OH[D]. Dalian: Dalian University of Technology, 2015: 1-211(in Chinese).
[56] WEN Y, HE J, LIU X, et al. Linear and non-linear relationships between bioconcentration and hydrophobicity: Theoretical consideration [J]. Environmental Toxicology and Pharmacology, 2012, 34(2): 200-208.
[57] MCHEDLOV-PETROSSYAN N O, VODOLAZKAYA N A, DOROSHENKO A O. Ionic equilibria of fluorophores in organized solutions: The influence of micellar microenvironment on protolytic and photophysical properties of rhodamine B [J]. Journal of Fluorescence, 2003, 13(3): 235-248.
[58] BRINKMANN M, ALHARBI H, FUCHYLO U, et al. Mechanisms of pH dependent uptake of ionizable organic chemicals by fish from oil sands process-affected water (OSPW) [J]. Environmental Science & Technology, 2020, 54(15): 9547-9555.
[59] 邰红巍, 闻洋, 苏丽敏, 等. 有机污染物在鱼体内临界浓度研究进展 [J]. 科学通报, 2015(19): 1789-1795.
TAI H W, WEN Y, SU L M, et al. Critical body residue to fish of organic pollutants [J]. Chinese Science Bulletin, 2015(19): 1789-1795(in Chinese).
[60] 席越, 杨先海, 张红雨, 等. 基于形态修正的描述符构建可电离化合物对大型溞急性毒性的QSAR模型 [J]. 生态毒理学报, 2019, 14(4): 183-191.
XI Y, YANG X H, ZHANG H Y, et al. Development of acute toxicity of daphnia magna QSAR models for ionogenic organic chemicals based on chemical from adjusted descriptors [J]. Asian Journal of Ecotoxicology, 2019, 14(4): 183-191(in Chinese).
[61] LIN S Y, YANG X H, LIU H H. Development of liposome/water partition coefficients predictive models for neutral and ionogenic organic chemicals [J]. Ecotoxicology and Environmental Safety, 2019, 179: 40-49.
[62] BOLTON J L, DUNLAP T L. Formation and biological targets of quinones: Cytotoxic versus cytoprotective effects [J]. Chemical Research in Toxicology, 2017, 30(1): 13-37.
[63] TERRENCE J M, DOUGLAS C J. The metabolism and toxicity of quinones, quinonimines, quinonemethides and quinone-thioethers [J]. Current Drug Metabolism, 2002, 3(4): 425-438.
[64] CHRASTINA A, WELSH J, RONDEAU G, et al. Plumbagin-serum albumin interaction: spectral, electrochemical, structure-binding analysis, antiproliferative and cell signaling aspects with implications for anticancer therapy [J]. ChemMedChem, 2020, 14(15): 1338-1347.
[65] ZHAO C, BORIANI E, CHANA A, et al. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF) [J]. Chemosphere, 2008, 73(11): 1701-1707.
[66] GISSI A, NICOLOTTI O, CAROTTI A, et al. Integration of QSAR models for bioconcentration suitable for REACH [J]. Science of the Total Environment, 2013, 456: 325-332.
[67] ZHANG X M, SUN X F, JIANG R F, et al. Screening new persistent and bioaccumulative organics in China's inventory of industrial chemicals [J]. Environmental Science & Technology, 2020, 54: 7398-7408.
[68] GB/T24782-2009. 持久性、生物累积性和毒性物质及高持久性和高生物累积性物质的判定方法[S]. 北京: 中华人民共和国国家质量监督检验检疫总局和中国国家标准化管理委员会, 2009.
GB/T24782-2009. Determination methods for persistent, bioaccumulative and toxic substances and highly persistent and highly bioaccumulative substances[S]. Beijing: General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China, Standardization Administration of China, 2009(in Chinese).