Eco-Environmental Knowledge Web

生态毒理学报
Asian Journal of Ecotoxicology

第 17 卷 第 3 期 2022 年 6 月
Vol. 17, No.3 Jun. 2022

DOI: 10.7524/AJE.1673-5897.20210907001

郑玉婷, 王宝成, 于洋, 等. 一种筛选具有潜在持久性、迁移性和毒性(PMT)新污染物的计算毒理学模型工具[J]. 生态毒理学报,2022, 17(3): 111-120

Zheng Y T, Wang B C, Yu Y, et al. A computational toxicological modeling tool for screening potentially persistent, mobile, and toxic (PMT) emerging contaminants [J]. Asian Journal of Ecotoxicology, 2022, 17(3): 111-120 (in Chinese)

一种筛选具有潜在持久性、迁移性和毒性(PMT)新污染物的计算毒理学模型工具

郑玉婷[1]，王宝成[2]，于洋[1,*]，黄怡[2]，张丽丽[1]，杨先海[3]，金彪[4]，林军[1]，张干[4]

1. 生态环境部固体废物与化学品管理技术中心,北京 100029
2. 北京市污染源管理事务中心,北京 100089
3. 南京理工大学环境与生物工程学院,南京 210094
4. 国家有机地球化学国家重点实验室,中国科学院广州地球化学研究所,广州 510640

收稿日期:2021-09-07　　**录用日期**:2021-12-07

摘要: 具有持久性、迁移性和毒性(PMT)的化学物质,可能会对生态环境及人类健康造成危害,正受到世界各国化学品管理机构的关注。近年来,我国化学品环境管理机构也开始关注化学物质的 PMT 危害特性,并逐步开展潜在 PMT 物质的筛选及环境风险评估工作。然而,筛选工具的缺乏已成为制约我国开展有毒有害物质以及新污染物筛选等工作的重要因素。为服务于我国潜在 PMT 物质的环境管理工作,本研究基于 14 770 条数据信息,构建了能够预测化学物质 PMT 特性,且能快速筛选出潜在 PMT 物质的高通量计算毒理学工具。该工具包含 26 个定性和定量模型,模型表征结果显示,定性模型均具有较好的分类性能,定量模型均具有较好的拟合优度、稳健性和预测能力。

关键词: 新污染物;持久性;迁移性;毒性;PMT;(定量)结构-活性关系;计算毒理学
文章编号: 1673-5897(2022)3-111-10　　**中图分类号**: X171.5　　**文献标识码**: A

A Computational Toxicological Modeling Tool for Screening Potentially Persistent, Mobile, and Toxic (PMT) Emerging Contaminants

Zheng Yuting[1], Wang Baocheng[2], Yu Yang[1,*] , Huang Yi[2], Zhang Lili[1], Yang Xianhai[3], Jin Biao[4], Lin Jun[1], Zhang Gan[4]

1. Solid Waste and Chemicals Management Center, Ministry of Ecology and Environment, Beijing 100029, China
2. Beijing Municipal Pollution Source Management Center, Beijing 100089, China
3. School of Environmental and Biological Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
4. State Key Laboratory of Organic Geochemistry, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou 510640, China

Received 7 September 2021　　**accepted** 7 December 2021

Abstract: Chemical substances with persistence, mobility and toxicity (PMT), which may cause harm to the eco-

基金项目:具有地域特征的优控有毒有害大气污染物动态识别和筛选研究总理基金课题(DQGG0305-02);国家重点研发计划课题(2016YFD0200208,2017YFD0800701);有机地球化学国家重点实验室开放基金课题(SKLOG2020)
第一作者:郑玉婷(1989—),女,工程师,研究方向为化学物质环境风险评估,E-mail:zhengyuting@meescc.cn
* **通讯作者**(Corresponding author), E-mail: yuyang@meescc.cn

logical environment and human health, are attracting the attention of chemical management agencies all over the world. In recent years, the chemical environmental management agency of China is paying attention to the PMT hazard characteristics, and is also carrying out the screening and environmental risk assessment of potential chemical substances of PMT. However, the lack of screening tools has become an important factor restricting the screening of toxic and harmful substances and new pollutants in China. In order to serve the environmental management of potential PMT substances in China, a high throughput computational toxicological tool based on 14 770 data was developed to predict the properties of PMT and rapidly identify potential PMT substances. And 26 models were included in the tool, the qualitative model characterization results indicated that they had good classification performance, and the quantitative model characterization results indicated that they had satisfied goodness-of-fit, robustness and prediction ability.

Keywords: emerging contaminants; persistent; mobile; toxic; PMT; (Q)SAR; computational toxicology

PMT(persistent mobile toxic)类物质是一类具有持久性、迁移性和毒性的有机化学物质总称[1]。该类物质是一类新污染物,具有难降解、移动性强,不易被化学或者生物过程消减等特点,且难被土壤或活性炭等吸附去除,较易穿透土壤或水处理设施屏障,容易赋存于地表水、地下水和饮用水,对生态环境和人类健康产生未知风险。有研究表明,水环境已检测出潜在的 PMT 类物质,例如甲基叔丁醚(MTBE)、全氟烷基酸(PFAA)、三氯乙烯和四氯乙烯等[2-4]。

当前,我国化学工业规模大于欧盟和美国总和,应对化学物质的环境释放已成为环境安全的重大挑战。新时代下,化学品环境管理战略也在不断创新。继 2016 年美国修订了《有毒物质控制法》,欧盟于 2020 年更新了《面向无毒环境的化学品可持续发展战略》,制定了“全新的欧洲化学品管理政策长期规划”,提出从生命周期的角度尽量减少 PMT 类新污染物对生态环境的影响,以确保生态环境的总体可持续。计划在《欧盟物质和混合物的分类、标签和包装法规》(CLP)中提出新的关于环境毒性、持久性、迁移性和生物积累性的危害等级和标准,并计划将内分泌干扰物、持久性、流动性、毒性以及高持久性和高迁移性的化学物质,列为高关注物质类别。德国联邦环境署(UBA)在欧盟 REACH 框架下,牵头建立了 PMT 物质的鉴定评判标准。我国于 2020 年提出了“重视新污染物治理”的新要求,国务院办公厅于 2022 年 5 月 4 日正式印发了《新污染物治理行动方案》。部分潜在 PMT 类物质已被纳入我国优先控制化学品名录管理,例如三氯乙烯、四氯乙烯等被列入我国《优先控制化学品名录(第一批)》。但是,仍有未知数量的 PMT 类物质还未受到关注及管控,我国缺乏服务于化学品环境风险管理的专业模型工具,计算毒理学工具逐渐成为了化学品环境管理的重要工具之一[5]。

为应对国际化学物质环境管理新趋势,贯彻落实新发展理念,认真执行新污染物治理行动方案,本研究运用计算毒理学技术,开发了一种能够筛选潜在 PMT 类新污染物的模型工具,辅助环境管理工作者从数以万计的化学物质中,快速识别出具有 PMT 危害特性的化学物质,以期服务于我国化学品环境管理及新污染物治理。

1 材料与方法(Materials and methods)

1.1 模型构建与验证方法

1.1.1 建模数据

本研究构建 PMT 模型的数据集包含了 14 770 条数据信息[6],P 模型包含 1 629 个化学物质的快速生物降解性数据,M 模型包含 9 961 个化学物质正辛醇-水分配系数数据,T 模型包含 946 个化学物质的鱼急性毒性数据,94 个化学物质的鱼慢性毒性数据,978 个化学物质的大型溞急性毒性数据,307 个化学物质的大型溞慢性毒性数据,445 个化学物质的绿藻急性毒性数据,410 个化学物质的绿藻慢性毒性数据。

1.1.2 建模方法

构建 PMT 模型时,建模数据均按照 3∶1 的比例,随机分为训练集和验证集。采用 PaDEL-Descriptor 软件[7],计算一维、二维分子结构描述符及 Pubchem 分子指纹描述符,其中一维和二维分子结构描述符用于建模,Pubchem 分子指纹描述符用于计算相似性指数(TS)[8],评估目标化学物质预测结果的可靠性。所有模型使用自编的 Python 程序[9-10]通过 *k*-最邻近分类(*k*NN)算法构建模型,采用 Euclidean 距离表征应用域。Euclidean 距离计算方法

如公式(1)所示:

$$D_E(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \quad (1)$$

式中:D_E 是 Euclidean 距离;x 和 y 是不同种化学物质;x_i 和 y_i 分别是化学物质 x 和化学物质 y 的第 i 个描述符的值。若目标化学物质的 Euclidean 距离小于对应模型训练集的 Euclidean 距离最大值,则判定目标化学物质在模型应用域内;反之,则判定其不在模型应用域内。

定性模型采用预测准确度(Q)、敏感性(S_n)和特异性(S_p)参数表征模型内部和外部预测能力,通过马修斯相关系数(MCC)、受试者工作特征曲线(ROC 曲线)下的面积(AUC)来表征分类性能。定量模型采用实测值与预测值之间的相关系数平方(r^2)来表征模型的拟合优度,去一法交叉验证系数(Q^2_{LOO})、去多法交叉验证系数(Q^2_{LMO})和 Bootstrapping 法验证系数(Q^2_{BOOT})表征模型的稳健性;并通过均方根误差(RMSE)、外部验证系数(Q^2_{EXT})、标准偏差(s)和平均绝对误差(MAE)等表征模型的内部和外部预测能力[11]。

1.1.3　毒性预测分类策略

毒性模型针对鱼类、大型溞和绿藻分别进行建模。由于不同类型的化学物质毒性作用差异较大,因此将目标化学物质进行了分类,分类策略如下。根据国标《化学品分类和标签规范 第 28 部分:对水生环境的危害》(GB 30000.28—2013)[12]以及《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》(GB/T 24782—2009)[13]关于毒性物质的判别标准,对于鱼急性毒性分类预测模型,以 L(E)C_{50} 为 0.01、0.1、1、10 和 100 mg·L^{-1}作为分类阈值,基于 946 个化学物质的鱼急性毒性数据,构建鱼急性毒性预测模型。但由于建模数据中 $LC_{50}\leqslant$0.01 mg·L^{-1}的物质数量少不足以建模,因此,本研究以 0.1、1、10 和 100 mg·L^{-1}为分类阈值,构建分类模型Ⅰ~Ⅳ,分类策略如图 1 所示。对于鱼慢性毒性分类预测模型,则以 NOEC 为 0.01、0.1 和 1 mg·L^{-1}为分类阈值,基于 94 个化学物质的鱼慢性毒性数据,构建鱼慢性毒性预测模型Ⅰ~Ⅲ,分类策略如图 2 所示。

与鱼急性/慢性毒性分类策略类似,对于大型溞急性毒性分类预测模型,是基于 978 个化学物质的大型溞急性毒性数据,以 EC_{50} 为 0.01、0.1、1、10 和 100 mg·L^{-1}作为分类阈值,构建分类预测模型Ⅰ~Ⅴ;

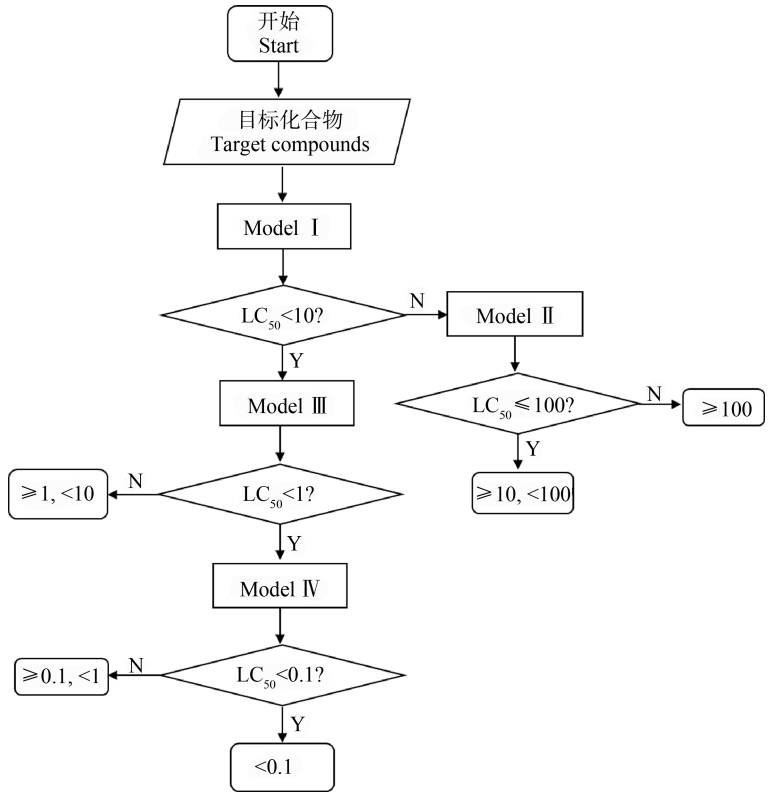

图 1　鱼急性毒性(LC_{50})预测模型分类策略示意图

Fig. 1　Schematic diagram of classification strategy for fish acute toxicity (LC_{50}) prediction model

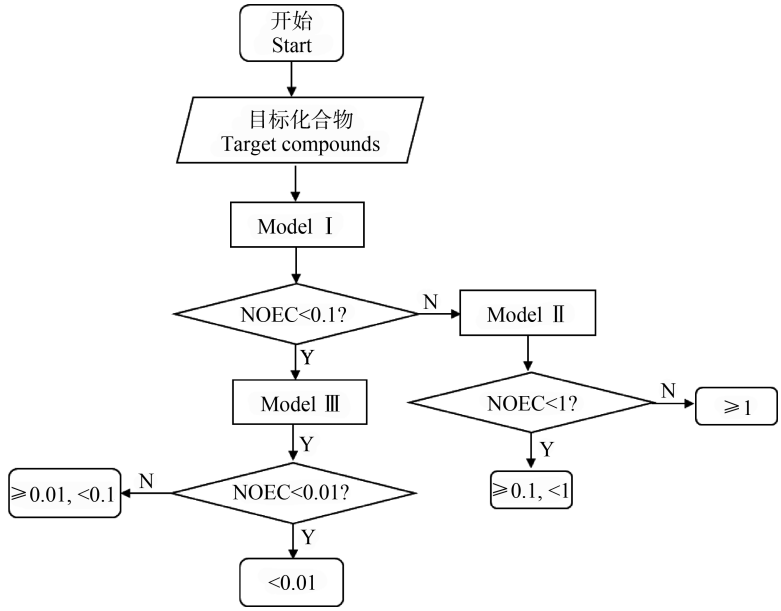

图 2 鱼慢性毒性(NOEC)预测模型分类策略示意图

Fig. 2 Schematic diagram of classification strategy for fish chronic toxicity (NOEC) prediction model

对于大型溞慢性毒性分类预测模型,是基于 307 个化学物质的大型溞慢性毒性数据,以 NOEC 为 0.01、0.1 和 1 mg·L^{-1}作为分类阈值,构建分类预测模型Ⅰ~Ⅲ;对于绿藻急性毒性分类预测模型,是基于 445 个化学物质的绿藻急性毒性数据,但由于 EC_{50} ≤0.01 mg·L^{-1}的物质个数少不足以建模,因此以 EC_{50} 为 0.1、1、10 和 100 mg·L^{-1}作为分类阈值,构建分类预测模型Ⅰ~Ⅳ;对于绿藻慢性毒性分类预测模型,是基于 410 个化学物质的绿藻慢性毒性数据,以 NOEC 为 0.01、0.1 和 1 mg·L^{-1}作为分类阈值,构建分类预测模型Ⅰ~Ⅲ。

1.2 PMT 类物质筛选方法

本研究根据我国《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》(GB/T 24782—2009)判断化学物质是否具有持久性(P)和毒性(T)[13],根据德国联邦环境署有关标准判断化学物质是否具有迁移性(M)[1]。其中,P 通过化学物质的快速生物降解属性来确定,如果不能快速生物降解,则表明该物质具有持久性;M 通过 $\log K_{oc}$ 判断,如果 $\log K_{oc}$<4,则表明该物质具有迁移性;其中,$\log K_{oc}$ 通过 $\log K_{ow}$ 估算,该方法也是加拿大环境多介质模型工具[14](new equilibrium criterion)采用的方法之一,如公式(2)或(3)所示:

$$K_{oc}=0.35K_{ow} \quad (2)$$

$$\log K_{oc}=\log K_{ow}-0.456 \quad (3)$$

T 预测根据国标《化学品分类和标签规范 第 28 部分:对水生环境的危害》(GB 30000.28—2013)[12],通过化学物质对水生急性毒性或水生慢性毒性效应数据判断,如果鱼类急性毒性效应值(LC_{50})、大型溞急性毒性效应值(EC_{50})、绿藻急性毒性效应值(EC_{50})<0.1 mg·L^{-1}(或<0.01 mg·L^{-1}),或者水生慢性毒性效应数据(NOEC)<0.01 mg·L^{-1},则表明该化学物质具有水生生物毒性。

1.3 PMT 筛选工具开发

本研究基于 Python 语言开发了能够自动预测 PMT 属性的软件工具,即有毒有害化学物质高通量危害识别系统,以实现模型的高通量预测及筛选功能。该系统支持单一及批量化学物质的 SMILES 码、CAS 号等输入方式,通过输入化学物质的结构信息,即可高通量预测化学物质的快速生物降解性、吸附系数($\log K_{oc}$)、水生生物急慢性毒性,并根据筛选标准,评估识别潜在 PMT 类物质。

1.4 PMT 属性预测与对比

本研究利用有毒有害化学物质高通量危害识别系统,开展了 335 个化学物质 P、M 和 T 属性的预测,并将预测结果与 Huang 等[15]的研究成果进行了比对。Huang 等[15]的研究成果中包含了 432 个化学物质的 P、M 和 T 数据,同时具有 P、M 和 T 这 3 项指标的化学物质是 335 个,因此本研究对比验证的物质为 335 个。

2　结果与讨论(Results and discussion)

2.1　持久性(P)预测模型

快速生物降解最优模型包含了 MLFER_S、MLFER_BO、TSRW、MlogP 和 WTPT-4 这 5 个预测变量。模型 Q、S_n 和 S_p 分别介于 0.83 ~ 0.88、0.78 ~ 0.86 和 0.86 ~ 0.89;MCC 和 AUC 分别介于 0.64 ~ 0.75 和 0.86 ~ 0.96,说明模型具有较好的分类性能,表征结果如表 1 所示。模型应用域显示,目标化学物质的 Euclidean 距离≤1.24 时,在模型的应用域范围内。

2.2　迁移性(M)预测模型

正辛醇-水分配系数($\log K_{ow}$)最优模型包含了 CrippenLogP、XlogP 和 nHaaCH 这 3 个预测变量。如表 2 所示,模型的 $r^2_{训练集}$为 0.92,Q^2_{LOO} 为 0.84,Q^2_{EXT} 为 0.82,$RMSE_{训练集}$为 0.51,$RMSE_{验证集}$为 0.75,说明该模型具有较好的拟合优度、稳健性和外部预测能力。训练集和验证集化学物质 $\log K_{ow}$ 实验值和预测值关系如图 3 所示,可见实验值与预测值具有较好的拟合优度。模型应用域显示,目标化学物质的 Euclidean 距离≤0.99 时,在模型的应用域范围内。

表 1　快速生物降解模型表征结果

Table 1　Characterization results of a rapid biodegradation model

数据集 Data set	k	n	TP	TN	FN	FP	S_n	S_p	Q	MCC	AUC
训练集 Training set	3	1 221	401	675	64	81	0.86	0.89	0.88	0.75	0.96
验证集 Validation set		408	135	202	38	33	0.78	0.86	0.83	0.64	0.86

注:k 表示邻近数;n 表示化学物质数;TP 表示真阳性;TN 表示真阴性;FN 表示假阴性;FP 表示假阳性;S_n 表示敏感性;S_p 表示特异性;Q 表示预测准确度;MCC 表示马修斯相关系数;AUC 表示 ROC 曲线下面积。

Notes: k is the number of neighbors; n is number of chemicals; TP is true positive; TN is true negative; FN is false negative; FP is false positive; S_n is sensitivity; S_p is specificity; Q is predictive accuracy; MCC is Matthews correlation coefficient; AUC represents the area under the ROC curve.

表 2　$\log K_{ow}$ 模型表征结果

Table 2　Model characterization results of $\log K_{ow}$

指标 Indicators	数据集 Data set	参数 Parameters	数值 Numerical value
训练集数量 Number of training sets	训练集 Training set	$n_{训练集}$　$n_{training}$	7 470
拟合优度 Goodness of fit		$r^2_{训练集}$　$r^2_{training}$	0.92
内部预测能力 Internal predictive power		$RMSE_{训练集}$　$RMSE_{training}$	0.51
稳健性 Robustness		$s_{训练集}$　$s_{training}$	0.51
		$MAE_{训练集}$　$MAE_{training}$	0.38
		Q^2_{LOO}	0.84
		Q^2_{LMO}	0.82
		Q^2_{BOOT}	0.83
验证集数量 Number of verification sets	验证集 Validation set	$n_{验证集}$　$n_{validation}$	2 491
外部预测能力 External predictive power		Q^2_{EXT}	0.82
		$RMSE_{验证集}$　$RMSE_{validation}$	0.75
		$s_{验证集}$　$s_{validation}$	0.75
		$MAE_{验证集}$　$MAE_{validation}$	0.57

备注:$n_{训练集}$和 $n_{验证集}$分别表示训练集和验证集数量;$r^2_{训练集}$表示训练集中实测值与预测值的相关系数;$RMSE_{训练集}$和 $RMSE_{验证集}$分别表示训练集和验证集的均方根误差;$s_{训练集}$和 $s_{验证集}$分别表示训练集和验证集的标准偏差;$MAE_{训练集}$和 $MAE_{验证集}$分别表示训练集和验证集的预测平均误差;Q^2_{LOO}、Q^2_{LMO} 和 Q^2_{BOOT} 分别表示去一法交叉验证系数、去多法交叉验证系数和 Bootstrapping 法验证系数;Q^2 表示外部验证系数。

Note: $n_{training}$ and the $n_{validation}$ are the number of training set and verification set respectively; $r^2_{training}$ is the correlation between the measured and predicted values; $RMSE_{training}$ and $RMSE_{validation}$ are the root mean square error of the training set and the verification set respectively; $s_{training}$ and $s_{validation}$ are the standard deviation of the training set and the verification set respectively; $MAE_{training}$ and $MAE_{validation}$ are the prediction average error of the training set and the verification set respectively; Q^2_{LOO}, Q^2_{LMO} and Q^2_{BOOT} are the one-off cross-validation coefficient, the multi-method cross-validation coefficient and the Bootstrapping method validation coefficient respectively; Q^2_{EXT} is the external validation factor.

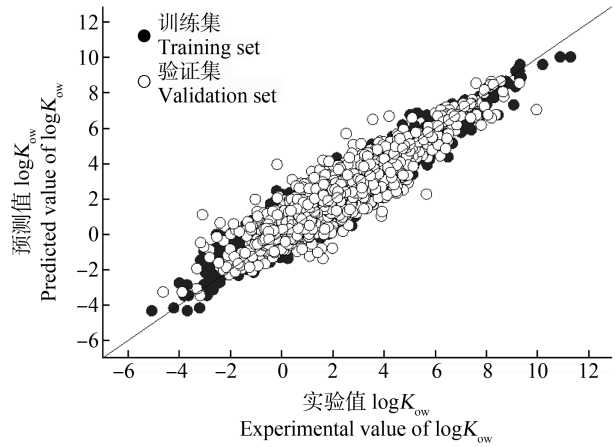

图 3 $\log K_{ow}$ 实验值和预测值拟合图

Fig. 3 $\log K_{ow}$ fitted graph of experimental and predicted values

2.3 毒性(T)预测模型

2.3.1 鱼急慢性毒性分类预测模型

2.3.1.1 鱼急性毒性分类预测模型

鱼急性毒性分类模型Ⅰ,以 LC_{50}=10 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$、GATS1p、SdCH2、nHBint3、nHAvin 和 maxssssC 这 6 个预测变量;鱼急性毒性分类模型Ⅱ,以 LC_{50}=100 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$、SIC0、maxHBint6、nHdCH2 和 minsssCH 这 5 个预测变量;鱼急性毒性分类模型Ⅲ,以 LC_{50}=1 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$、AATSC0v、MATS3p 和 VE1_DzZ 这 4 个预测变量;鱼急性毒性分类模型Ⅳ,以 LC_{50}=0.1 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$、AATSC1m、GATS2c 和 MATS1c 这 4 个预测变量。如表 3 所示,模型 Q、S_n 和 S_p 分别介于 0.85 ~0.92、0.70 ~0.92 和 0.81 ~0.92;MCC 和 AUC 分别介于 0.63 ~0.79 和 0.81 ~0.96,说明模型具有较好的分类性能。模型应用域表征结果显示,对于模型Ⅰ~Ⅳ,目标化学物质的 Euclidean 距离分别小于 1.23、1.04、1.05 和 1.07 时,在相应模型的应用域范围内。

2.3.1.2 鱼慢性毒性分类预测模型

鱼慢性毒性分类模型Ⅰ,以 NOEC=0.1 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$、bpol 和 minaasC 这 3 个预测变量;鱼慢性毒性分类模型Ⅱ,以 NOEC=1 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$、IC5 和 AATSC5p 这 3 个预测变量;鱼慢性毒性分类模型Ⅲ,以 NOEC=0.01 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$ 和 nHBint3 这 2 个预测变量。如表 4 所示,模型 Q、S_n 和 S_p 分别介于 0.88 ~1、0.89 ~1 和 0.85 ~1;MCC 和 AUC 分别介于 0.75 ~1 和 0.86 ~1,说明模型具有较好的分类性能。模型应用域表征结果显示,对于模型Ⅰ~Ⅲ,目标化学物质的 Euclidean 距离分别小于 0.73、0.75 和 1.04 时,在相应模型的应用域范围内。

2.3.2 大型溞急慢性毒性分类预测模型

2.3.2.1 大型溞急性毒性分类预测模型

大型溞急性毒性分类模型Ⅰ,以 EC_{50}=10 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$、minwHBa、ndsssP、SsSH 和 JGI6 这 5 个预测变量;大型溞急性毒性分类模型Ⅱ,以 EC_{50}=100 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$、MPC5、nBase、SRW6 和 naaN 这 5 个预测变量;大型溞急性毒性分类模型Ⅲ,以 EC_{50}=1 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$、BIC0、SdsssP 和 n6HeteroRing 这 4 个预测变量;大型溞急性毒性分类模型Ⅳ,以 EC_{50}=0.1 mg·L^{-1}为分类阈值,最优模型包含了 $\log K_{ow}$、bpol、AATSC0i

表 3 鱼急性毒性分类模型表征结果

Table 3 Characterization of fish acute toxicity classification model

模型 Model	数据集 Data set	k	n	TP	TN	FN	FP	S_n	S_p	Q	MCC	AUC
模型Ⅰ	训练集 Training set	3	709	303	321	40	45	0.88	0.88	0.88	0.76	0.95
Model Ⅰ	验证集 Validation set		237	91	113	15	18	0.86	0.86	0.86	0.72	0.89
模型Ⅱ	训练集 Training set	3	372	217	110	21	24	0.91	0.82	0.88	0.74	0.94
Model Ⅱ	验证集 Validation set		125	67	42	6	10	0.92	0.81	0.87	0.74	0.90
模型Ⅲ	训练集 Training set	3	336	82	205	28	21	0.75	0.91	0.85	0.66	0.92
Model Ⅲ	验证集 Validation set		113	23	73	10	7	0.70	0.91	0.85	0.63	0.81
模型Ⅳ	训练集 Training set	3	107	25	73	3	6	0.89	0.92	0.92	0.79	0.96
ModelⅣ	验证集 Validation set		36	7	25	1	3	0.88	0.89	0.89	0.71	0.87

和 MATS7s 这 4 个预测变量;大型溞急性毒性分类模型 V,以 EC_{50}=0.01 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow}、GATS3c、mindCH2 和 SCH-3 这 4 个预测变量。如表 5 所示,模型 Q、S_n 和 S_p 分别介于 0.79 ~ 0.89、0.81 ~ 0.94 和 0.70 ~ 0.84;MCC 和 AUC 分别介于 0.57 ~ 0.72 和 0.77 ~ 0.92,说明模型具有较好的分类性能。模型应用域表征结果显示,对于模型Ⅰ ~ Ⅴ,目标化学物质的 Euclidean 距离分别小于 1.03、1.39、0.98、0.88 和 0.99 时,在相应模型的应用域范围内。

2.3.2.2　大型溞慢性毒性分类预测模型

大型溞慢性毒性分类模型Ⅰ,以 NOEC=1 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow}、AATSC0v、SHBint2 和 AATS2e 这 4 个预测变量;大型溞慢性毒性分类模型Ⅱ,以 NOEC=0.1 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow}、MAXDP、SHdsCH 和 ATSC6c 这 4 个预测变量;大型溞慢性毒性分类模型Ⅲ,以 NOEC=0.01 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow} 和 ATSC2p 这 2 个预测变量。如表 6 所示,模型 Q、S_n 和 S_p 分别介于 0.84 ~ 0.90、0.72 ~ 1 和 0.81 ~ 0.91;MCC 和 AUC 分别介于 0.63 ~ 0.76 和 0.86 ~ 0.95,说明模型具有较好的分类性能。模型应用域表征结果显示,对于模型Ⅰ ~ Ⅲ,目标化学物质的 Euclidean 距离分别小于 1.1、0.95 和 0.75 时,在相应模型的应用域范围内。

2.3.3　绿藻急慢性毒性分类预测模型

2.3.3.1　绿藻急性毒性分类预测模型

绿藻急性毒性分类模型Ⅰ,以 EC_{50}=10 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow}、SwHBa、nHBint6 和 MLFER_BO 这 4 个预测变量;绿藻急性毒性分类模型Ⅱ,以 EC_{50}=100 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow}、AATS4p、MPC10 和 ETA_dEpsilon_D 这 4 个预测变量;绿藻急性毒性分类模型Ⅲ,以 EC_{50}=1 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow}、SpMax_Dt 和 GATS2v 这 3 个预测变量;绿藻急性毒性分类模型Ⅳ,以 EC_{50}=0.1 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow}、AATSC0m 和 AATS6e 这 3 个预测变量。如表 7 所示,模型 Q、S_n

表 4　鱼慢性毒性分类模型表征结果

Table 4　Characterization of fish chronic toxicity classification model

模型 Model	数据集 Data set	k	n	TP	TN	FN	FP	S_n	S_p	Q	MCC	AUC
模型Ⅰ	训练集 Training set	3	70	25	38	3	4	0.89	0.90	0.90	0.79	0.95
Model I	验证集 Validation set		24	10	11	1	2	0.91	0.85	0.88	0.75	0.86
模型Ⅱ	训练集 Training set	3	41	21	18	2	0	0.91	1	0.95	0.91	0.97
Model II	验证集 Validation set		14	4	9	0	1	1	0.90	0.93	0.85	0.98
模型Ⅲ	训练集 Training set	3	29	8	19	1	1	0.89	0.95	0.93	0.84	0.96
Model III	验证集 Validation set		10	4	6	0	0	1	1	1	1	1

表 5　大型溞急性毒性分类模型表征结果

Table 5　Characterization of acute toxicity classification model of *Daphnia*

模型 Model	数据集 Data set	k	n	TP	TN	FN	FP	S_n	S_p	Q	MCC	AUC
模型Ⅰ	训练集 Training set	3	733	425	208	41	59	0.91	0.78	0.86	0.70	0.93
Model I	验证集 Validation set		245	138	61	21	25	0.87	0.71	0.81	0.58	0.81
模型Ⅱ	训练集 Training set	3	264	163	67	11	23	0.94	0.74	0.87	0.71	0.92
Model II	验证集 Validation set		89	64	15	7	3	0.90	0.83	0.89	0.68	0.83
模型Ⅲ	训练集 Training set	3	468	251	145	34	38	0.88	0.79	0.85	0.68	0.91
Model III	验证集 Validation set		157	77	47	13	20	0.86	0.70	0.79	0.57	0.83
模型Ⅳ	训练集 Training set	3	281	115	117	22	27	0.84	0.81	0.83	0.65	0.90
Model Ⅳ	验证集 Validation set		94	46	33	6	9	0.88	0.79	0.84	0.68	0.84
模型 V	训练集 Training set	3	141	64	57	9	11	0.88	0.84	0.86	0.72	0.91
Model V	验证集 Validationset		48	17	21	4	6	0.81	0.78	0.79	0.58	0.77

和 S_p 分别介于 0.82 ~ 0.90、0.78 ~ 0.95 和 0.64 ~ 0.94;MCC 和 AUC 分别介于 0.60 ~ 0.79 和 0.79 ~ 0.95,说明模型具有较好的分类性能。模型应用域表征结果显示,对于模型Ⅰ ~ Ⅳ,目标化学物质的 Euclidean 距离分别为小于 1.25、1.17、1.03 和 0.98 时,在相应模型的应用域范围内。

2.3.3.2 绿藻慢性毒性分类预测模型

绿藻慢性毒性分类模型Ⅰ,以 NOEC = 1 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow}、piPC7、AATSC5p、VP-7、SHsSH 和 MDEC-34 这 6 个预测变量;绿藻慢性毒性分类模型Ⅱ,以 NOEC = 0.1 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow}、SpMax_Dt、SsOm、GATS3v 和 MATS8e 这 5 个预测变量。绿藻慢性毒性分类模型Ⅲ,以 NOEC = 0.01 mg·L^{-1}为分类阈值,最优模型包含了 logK_{ow}、nAtomP、nAtomLAC 和 GATS8p 这 4 个预测变量。如表 8 所示,模型 Q、S_n 和 S_p 分别介于 0.84 ~ 0.94、0.75 ~ 0.90 和 0.84 ~ 0.96;MCC 和 AUC 分别介于 0.68 ~ 0.86 和 0.82 ~ 0.96,说明模型具有较好的分类性能。模型应用域表征结果显示,对于模型Ⅰ ~ Ⅲ,目标化学物质的 Euclidean 距离分别为小于 1.28、1.05 和 1.06 时,在相应模型的应用域范围内。

表 6 大型溞慢性毒性分类预测模型表征结果

Table 6 Characterization of a classification prediction model for chronic toxicity of *Daphnia*

模型 Model	数据集 Data set	k	n	TP	TN	FN	FP	S_n	S_p	Q	MCC	AUC
模型Ⅰ	训练集 Training set	3	230	130	74	11	15	0.92	0.83	0.89	0.76	0.95
Model Ⅰ	验证集 Validation set		77	48	21	3	5	0.94	0.81	0.90	0.76	0.94
模型Ⅱ	训练集 Training set	3	144	54	73	7	10	0.89	0.88	0.88	0.76	0.93
Model Ⅱ	验证集 Validation set		48	13	29	3	3	0.81	0.91	0.88	0.72	0.86
模型Ⅲ	训练集 Training set	3	57	13	35	5	4	0.72	0.90	0.84	0.63	0.88
Model Ⅲ	验证集 Validation set		20	3	14	0	3	1	0.82	0.85	0.64	0.91

表 7 绿藻急性毒性分类预测模型表征结果

Table 7 Characterization of a predictive model for acute toxicity classification of green algae

模型 Model	数据集 Data set	k	n	TP	TN	FN	FP	S_n	S_p	Q	MCC	AUC
模型Ⅰ	训练集 Training set	3	333	174	111	21	27	0.89	0.80	0.86	0.70	0.93
Model Ⅰ	验证集 Validation set		112	57	37	5	13	0.92	0.74	0.84	0.68	0.86
模型Ⅱ	训练集 Training set	3	141	94	30	5	12	0.95	0.71	0.88	0.70	0.93
Model Ⅱ	验证集 Validation set		48	34	7	2	4	0.94	0.64	0.87	0.62	0.80
模型Ⅲ	训练集 Training set	3	192	73	85	17	17	0.81	0.83	0.82	0.64	0.89
Model Ⅲ	验证集 Validation set		65	19	34	3	9	0.86	0.79	0.82	0.63	0.84
模型Ⅳ	训练集 Training set	3	84	25	51	5	3	0.83	0.94	0.90	0.79	0.95
Model Ⅳ	验证集 Validation set		28	7	16	2	3	0.78	0.84	0.82	0.60	0.79

表 8 绿藻慢性毒性分类模型表征结果

Table 8 Characterization of chronic toxicity classification model of green algae

模型 Model	数据集 Data set	k	n	TP	TN	FN	FP	S_n	S_p	Q	MCC	AUC
模型Ⅰ	训练集 Training set	3	307	136	128	28	15	0.83	0.90	0.86	0.72	0.92
Model Ⅰ	验证集 Validation set		103	47	44	8	4	0.85	0.92	0.88	0.77	0.90
模型Ⅱ	训练集 Training set	3	163	62	75	12	14	0.84	0.84	0.84	0.68	0.89
Model Ⅱ	验证集 Validation set		55	18	29	3	5	0.86	0.85	0.85	0.70	0.87
模型Ⅲ	训练集 Training set	3	71	19	48	2	2	0.90	0.96	0.94	0.86	0.96
Model Ⅲ	验证集 Validation set		24	6	15	2	1	0.75	0.94	0.88	0.71	0.82

2.4　与现有潜在 PMT 物质对比

将 335 个化学物质 P、M、T 预测结果和 Huang 等[15]的研究成果对比可知,对于 P 有 299 个物质的结果一致,对于 M 有 299 个物质的结果一致,对于 T 有 70 个物质的结果一致。P 和 M 一致性比较高,分别为 89% 和 89% 。T 的一致性存在较大差异,是由于本研究模型服务于生态环境指标的预测,T 指的是藻、溞、鱼的急性与慢性毒性指标,而 Huang 等[15]的研究成果中,毒性指的是人体健康领域致癌、致突变和生殖毒性(CMR)等毒性指标,因此产生了较大的差异。

3　展望(Prospect)

新污染物治理是“十四五”期间我国深入打好污染防治攻坚战的主战场之一。PMT 类新污染物,可能会对人类健康构成威胁,对生态环境造成危害,进而产生影响气候变化、加速生态系统退化和加剧生物多样性锐减等全球性危机。当前,我国化学物质环境风险防控形势严峻,新污染物治理任务艰巨。党的十九届五中全会通过的《中共中央关于制定国民经济和社会发展第十四个五年规划和二〇三五年远景目标的建议》中提出了“重视新污染物治理”,明确了“健全有毒有害化学物质环境风险管理体制”。生态环境部高度重视新污染物治理工作,组织编制了《新污染物治理行动方案(征求意见稿)》及相关文件。

新污染物治理是一套系统工程,“筛、评、控”是核心内容。然而,新污染物数据相对缺失制约了筛查和评估过程。为了克服数据缺失的瓶颈,生态环境部固体废物与化学品管理技术中心面向社会开展了 2021 年计算毒理与暴露模型的征集工作,并尝试自主开发了多个计算毒理学模型工具,发挥计算毒理工具的预测优势,为我国新污染物治理、化学物质高通量危害筛查和风险评估提供技术支持。本研究着眼于新污染物治理及化学物质危害筛查,建立了我国 PMT 类新污染物筛选方法,基于 QSAR 方法开发了计算毒理学预测工具,首次实现了 PMT 类新污染物的高通量预测功能,旨在通过计算毒理学科学理论转化应用成果,探索新技术应用于新污染物治理实践,支撑我国化学物质环境管理中 PMT 类物质的筛选工作,提升我国 PMT 类新污染物环境风险管控能力,助力“十四五”深入打好污染防治攻坚战。

此外,PMT 类新污染物性质特殊,我国尚未建立相关监测标准,环境监管较为薄弱,仅依赖计算毒理工具解决 PMT 类新污染物的全部危害及暴露信息并不现实。尤其在环境监测技术方法和相关去除技术方面还需要社会各界更多的投入。同时,计算毒理工具的开发也依赖于高质量实测数据,随着未来建模数据与计算机技术的快速发展,PMT 属性的预测准确性也将不断提升。

通讯作者简介:于洋(1982—),男,博士,高级工程师,主要研究方向为化学物质环境管理技术方法和计算毒理学。

参考文献(References):

[1] Neumann M, Schliebner I. A revised proposal for implementing criteria and an assessment procedure to identify persistent, mobile and toxic (PMT) and very persistent, very mobile (vPvM) substances registered under REACH [R]. Dessau: German Environment Agency, 2019

[2] Achten C, Kolb A, Püttmann W. Occurrence of methyl tert-butyl ether (MTBE) in riverbank filtered water and drinking water produced by riverbank filtration. 2 [J]. Environmental Science & Technology, 2002, 36(17): 3662-3670

[3] Garnett J, Halsall C, Vader A, et al. High concentrations of perfluoroalkyl acids in Arctic Seawater driven by early thawing sea ice [J]. Environmental Science & Technology, 2021, 55(16): 11049-11059

[4] Pierri D. Actual decay of tetrachloroethene (PCE) and trichloroethene (TCE) in a highly contaminated shallow groundwater system [J]. Environmental Advances, 2021, 5: 100090

[5] 于洋, 林军, 郑玉婷, 等. 化学品环境管理的计算毒理学[M]. 北京: 中国农业出版社, 2021: 1-2

[6] Stituto di Ricerche Farmacologiche Mario Negri. IRCCS VEGA HUB [CP/OL]. [2021-08-26]. https://www.vegahub.eu/

[7] Yap C W. PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints [J]. Journal of Computational Chemistry, 2011, 32 (7): 1466-1474

[8] Yang X H, Ou W, Zhao S S, et al. Rapid screening of human transthyretin disruptors through a tiered *in silico* approach [J]. ACS Sustainable Chemistry & Engineering, 2021, 9(16): 5661-5672

[9] Liu H H, Yang X H, Lu R. Development of classification model and QSAR model for predicting binding affinity of endocrine disrupting chemicals to human sex hormone-binding globulin [J]. Chemosphere, 2016, 156: 1-7

[10] 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

[11] 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: 5-6
Zheng Y T. Development of QSAR models on bioconcentration factors of chemicals in fish [D]. Dalian: Dalian University of Technology, 2014: 5-6 (in Chinese)

[12] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. 化学品分类和标签规范 第 28 部分: 对水生环境的危害: GB 30000.28—2013[S]. 北京: 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会, 2013

[13] 国家质量监督检验检疫总局, 中国国家标准化管理委员会. 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: GB/T 24782—2009 [S]. 北京: 中国标准出版社, 2010

[14] Seth R, Mackay D, Muncke J. Estimating the organic carbon partition coefficient and its variability for hydrophobic chemicals [J]. Environmental Science & Technology, 1999, 33(14): 2390-2394

[15] Huang C, Jin B, Han M, et al. The distribution of persistent, mobile and toxic (PMT) pharmaceuticals and personal care products monitored across Chinese water resources [J]. Journal of Hazardous Materials Letters, 2021, 2: 100026 ◆