第二章 使用QSAR和QSRR模型联合预测农药暴露对环境归趋、生态毒性和人类健康的影响

# 2.1 引言

定量结构活性关系（quantitative structure-activity relationships, QSAR) 旨在通过合理的数理统计方法建立起一系列化合物的生理活性或某种性质与其理化性质参数或者结构参数（包括二维分子结构参数、三维分子结构参数等）之间的定量关系。然后通过这些定量关系猜测化合物的相应特性，指导设计者有目的性地对生理活性物质进行结构改造，从而大大缩短高性能化合物的研发周期，节约研发成本。除了研发药物，它还能用于环境污染物对于各种生物（比如蚯蚓、大型蚤、黑呆头鱼等）毒性的预测166、167、168、169。

定量结构-保留关系(quantitative structure retention relationship, QSRR)模型常用于预测化合物（比如查尔酮）的色谱特性和保留时间170。通过预测代谢物的保留时间，可以更准确得度量农药对生物的影响。

基于农药性质数据库（Pesticide Property Database，PPDB，<https://sitem.herts.ac.uk/aeru/ppdb/en/index.htm>），通过外部（QSAR）和内部（QSRR）模型的预测，农药暴露对于环境、生态和人的影响将会得到更好的理解。

# 2.2材料和方法

## 2.2.1 化学品和仪器

标准品的HILIC柱保留时间测定

•LC-MS条件：采集模式MS1；阈值计数：200；m/z 50-1100；气体温度325°C；干燥气体10L/min；雾化器35磅/平方英寸；护套气体温度350°C；鞘气流量10L/min；TOF破片器100 V；撇沫器65 V；Oct1射频Vpp 750V；3个光谱/s；333.3毫秒/频谱

•LC系统：安捷伦1290 Infinity II作为流动相，安捷伦1260作为参考溶液

•MS系统：安捷伦6546 LC/四极飞行时间（Q-TOF）MS

LC柱：WATERS XBridge BEH Amide（15 cm x 2.1 mm；2.5µm）

缓冲液A:10mM乙酸铵和0.2%乙酸在95%H2O+3%ACN+2%MeOH中的溶液；

缓冲液B:10mM乙酸铵和0.2%乙酸在5%H2O+93%ACN+2%MeOH中的溶液；

注射液：5μL（+）-ESI和10μL（-）-ESI；洗涤：95%ACN+5%H2O，持续10s；流速（mL/min）：0.3

LC柱室温度：40°C；ESI模式：（+/-）

工作清单：在阳性和阴性ESI模式下注射每个样本（标记为-POS和-NEG）。样品制备的空白标记为空白-Prep。将血浆和组织样品结合起来，从血浆（QCp）和组织（QCt）中制备QC样品。

•梯度操作（分离）

最小 B%

0 95

3 95

8 50

12 50

13 95

35 95

## 2.2.2 数据处理

1. 使用R包bibliometrix总结了课题组2015-2023年间有关使用代谢组学研究农药暴露对模式生物影响的文献。
2. 使用QSAR模型农药暴露对环境归趋、生态毒性、人类健康的影响：
3. 搜索PPDB数据库（https://sitem.herts.ac.uk/aeru/ppdb/en/index.htm）
4. 使用python BeautifulSoup4和requests查看并获取数据库中所有农药的环境归趋、生态毒理、和人类健康警告情况以及该农药对应的SMILES式，删除不完整的数据，共有1145中农药具有这四项完整的属性，查找这1145种农药的分子描述符（MDs, molecular descriptors），能查到953种农药。使用这953种农药的分子描述符建立定量构效关系模型（QSAR,Quantitative Structure-Activity Relationship）。一共有953组数据；重新分组高级警告High Alert和非高级警告Not High Alert。

将数据以7：3分为训练集和测试集。

特征选择方式有ANOVA、随机森林，特征选择的标准是AUC（ROC曲线下面积）最大。ANOVA是基于统计学的方法，在处理线性相关的特征时比较慢。机器学习方法有随机森林（RF，100棵决策树）、核支持向量机（SVM，C=1）、逻辑回归（LR）、朴素贝叶斯（NB）、K近邻（KNN）、梯度提升树（GB）。

1. 基于python sklearn库建立模型，建立逻辑回归（LR）、支持向量机（SVM）、随机森林（RF）模型，预测农药的影响是否高级警告。
2. 保留时间的预测：
3. 使用相关性数据（Correlation Statistics）和互信息（Mutual Information）为特征选择方式，分子描述符（MDs）为特征，建立多元线性回归（MLR）、支持向量回归（SVR）、偏最小二乘回归（PLS，2个主成分）、随机森林回归（RF，40棵决策树组成）、LASSO回归（alpha=1.0）、岭回归（alpha=1.0）、梯度提升回归（GBR），预测化合物的保留时间，以平均绝对误差(Mean Absolute Error)为评价标准，选择最优的特征选择和建模的组合。
4. 注意：这些模型不是一定要标准化特征

# 2.3实验结果

## 2.3.1 回望过去（bibliometrix cite 1-47）

图6显示了课题组2015-2023年发表的文章（共47篇）的话题趋势和这些文章的核心关键词。

图7显示了课题组2015-2023年发表的文章的应用次数和发表的期刊。

## 2.3.2 使用MDs建立不同QSAR模型的预测效果

## 2.3.3 不同QSRR模型对HILIC色谱柱上化学物质保留时间的预测

如表1所示，将平均绝对误差MAE作为评价标准，得到了不同特征选择和不同建模方式组合的模型效果。交叉验证是很好的验证模型好坏的方式，采用重复3次的5折交叉验证，发现MAE最小的是梯度提升回归模型(GBR)，仅仅为0.956和0.958；而使用相关性统计量作为特征选择标准的GBR效果又略优于使用互信息为特征选择方式的，不过相关性统计量会选出更多的特征（172个，图8）。

GBR也有自身的缺陷，从表1可以看出它在训练集上的MAE（0.222和0.198）远小于在测试集（1.037和1.039）上的，说明模型倾向于过拟合，这通常意味着模型过于复杂，能够记住训练数据的细节和噪声，而不是学习数据的通用模式。

相比之下，随机森林回归（RF）可能是更好的模型,重复三次的五折交叉验证后MAE仅为1.014和1.009，并且数据在测试集上的MAE（1.107和1.063）与在训练集（0.424和0.423）上的差异不大；更有优势的是，比起GBR（172和132个，图8），它仅需要的特征数量（49和63个，图8）不多，就可以很好的预测保留时间。

描述了好的模型，再来看看表现不是那么好的模型。岭回归、LASSO、偏最小二乘回归明显都不是很好的模型。岭回归不是重复3次的5折交叉验证MAE过大（1.755），就是最大MAE过大（8.535）。试想一下、预测的保留时间和实际的差距高达8分钟以上，那么预测将毫无意义。LASSO回归的劣势在于，它需要太多的特征（114和104个，图8），并且重复3次的5折交叉验证MAE也很大（1.559和1.500）。偏最小二乘回归集中了以上两个模型的劣势，不再重复叙述。

值得一提的是多元线性回归MLR和支持向量回归SVR，虽然它们的预测效果一般般（表1的最后一列），但是需要的特征数量极少（31、3、2、1个，图8），便于后续分析，并且在测试集和训练集上的MAE差距极小，除了相关性统计量的多元线性回归，其余的二者之差不高于0.060。

所以，使用互信息的MLR模型、SVR模型和RF模型挑选出的特征是2.4.3中的重点关注对象。图9显示了通过MLR和SVR方法挑选出的特征和这些特征在模型中的重要程度（Score）。图10显示了通过RF的方法挑选出的特征。注意：不同的特征选择方法（相关性统计量、互信息）产生的得分（Score）不能相比，图上仅仅是表现特征的相对重要情况。

如果仅仅以MAE为评价标准，忽略模型是否过拟合，相关性统计量CS的特征选择方式配合GBR的机器学习方法无疑是最优的预测保留时间的模型，他的预测结果如图11所示。然而考虑过拟合问题，互信息MI的特征选择方式配合随机森林回归RF似乎才是最佳模型，其预测结果如图12所示。