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

共测了514个化合物，其中367个有明确的峰和保留时间，被用于下一步的QSRR建模。

测得数据可见：<https://github.com/YiFanYUE99/master-s-thesis-project/blob/main/QSRR/QSRR3/chem_database.xlsx>

## 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，使用Python RDkit模块计算这1145种农药的210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（图13）。分子描述符和它们对应的含义见表3

将数据以7：3分为训练集和测试集。这些953种农药包括了我们组曾研究过的甲霜灵、腈菌唑、六康唑、氟环唑、α-氯氰菊酯、苯霜灵、氯丹、DDT、苯醚甲环唑、硫丹、氟酰胺、戊菌唑、丙硫菌唑、戊唑醇、咪草烟、烯啶虫胺、吡虫啉、噻虫嗪、噻虫胺、嘧菌酯、百菌清、环唑醇。

以ANOVA检验获得特征评，特征个数选择的标准是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，由于分子描述符（MDs）含有负值，卡方检验等特征选择方式并不适用。机器学习方法有随机森林（RF，100棵决策树）、核支持向量机（SVM）、逻辑回归（LR）、朴素贝叶斯（NB）、K近邻（KNN）、梯度提升树（GB）。特征的选择和模型的建立主要使用了python Scikit-learn库。如无特殊说明，参数都是该库的默认参数。

1. 保留时间的预测：
2. 共有367个可用数据：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，210个）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。
3. 注意：这些模型不是一定要标准化特征

所有代码和原始数据都在<https://github.com/YiFanYUE99/master-s-thesis-project/>

# 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6（A）显示2016的最流行话题是degradation（降解）；2017年的流行话题是induction（诱导）和liver（肝）；2018年的话题是gene-expression（基因表达）和toxicity（毒性）；2019年的流行话题是exposure（暴露），oxidative stress（氧化应激）和metabolism（代谢）；2020年的最流行话题pesticides（农药）和residues（残留）；2021年的最流行话题是inflammation（炎症）。在这些关键话题之中，毒性、氧化应激、暴露和农药是一直都被关注的话题。图6（B）的词云图则表示在这些已发表的文章的关键词，关心的话题主要围绕在toxicity（毒性）、exposure（暴露）、induction（诱导）、oxidative stress（氧化应激）、degradation（降解）、metabolism（代谢）、metabolomics（代谢组）、inflammation（炎症）、pesticides（农药）、difenoconazole（苯醚甲环唑）、fungicide penconazole（杀菌剂戊菌唑）等。

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。已经有十篇文章被引用次数超过35，发表在AQUAT TOXICOL上的文章（cite14）已被引用79次；在这些文章中，发表在Environmental Pollution上的是最多的，多达13篇。

## 2.3.2 使用MDs建立不同QSAR模型的预测效果

通常，分类模型以ROC曲线下面积AUC为标准，也有使用分类正确率为标准的。表2显示了六种机器学习（随机森林、支持向量机、逻辑回归、朴素贝叶斯、K近邻、梯度提升树）建立的QSAR模型预测农药对环境归趋、生态毒性、人类健康的影响。虽然随机森林在预测环境归趋和生态毒性时具有高AUC（0.947和0.934）和不错的正确率（0.861和0.868），然而在测试集上的AUC却比较低，在训练集上又过于高（均为0.999），所以模型存在过拟合的问题。当预测对人类健康的影响时，在环境归趋与生态毒性上表现较好的随机森林和梯度提升树显得不好了，在测试集上的AUC和正确率均不超过0.5，表明这两个算法还不如瞎猜；SVM预测农药对人类健康的影响效果稍微好一些，全数据集AUC达到了0.777，准确率达到了0.709，但也没有很准确，因此使用这几种方法建立的QSAR模型难以准确预测农药对人类健康的影响。当关注最佳特征选择数时，预测生态毒性的模型占据了优势，以比较少的特征数（所有模型均不到20），达到了该模型的最佳预测正确率。因此，在2.4中主要讨论了决定农药生态毒性的分子描述符。

## 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。

# 2.4讨论

当预测农药暴露对生态毒性（鱼类和水藻急性/慢性毒性来评估）的影响，梯度提升决策时和随机森林都有不错的预测效果；随机森林在全数据集上正确预测的个数是827（TP+TN，图15（A）），错误预测个数是126（FP+FN）。评估生态毒性用到的分子描述符也很少，仅有13个，且ANOVA检验得分均高于6。与生态毒性相关性最强的分子描述符是Chi2v，它表示二级相邻原子对的吸电子能力之和的平均值；然后是NumHDonors，表示分子中的氢键给体原子数；Chi0v、Chi1v、Chi3v、Chi4v和Chi2v差不多，表示分子中不同种类相邻原子对的价电子对数；BCUT2D\_LOGPHI表示分子的最高的脂水分配系数LogP的对数对应的负荷特征值；MolLogP表示分子的亲水性/疏水性；qed则是分子的药用价值；NHOHCount是分子中的羟基氮原子数；SlogP\_VSA8表示分子上脂水分配相关的区域的性质；fr\_aniline表示分子中含有苯胺基团的数量、fr\_COO2表示分子中含有二元羧基的数量。以上的相关分子描述符表明农药的生态毒性和分子中原子对的吸电子能力以及亲水性/亲脂性有极大关系，这可能是因为PPDB数据库对于生态毒性的评估是基于农药对于水生生物的毒性，而水溶性的农药明显更容易对鱼和藻类产生毒性。比如，2017年的文章报道了可溶性砷杀虫剂会导致鱼类砷的生物蓄积，特别是在肝脏和肾脏中171；2023年的研究表明暴露于可溶性有机磷农药——马拉硫磷会导致一种杂交鱼体内检出马拉硫磷残留物且肠道内Lactobacillus上升172。

虽然QSAR模型预测农药对环境归趋（由持久性，排水流量移动性、地下水普遍性得分评估）和人类健康（由哺乳动物急性/慢性毒性，生殖、发育、神经影响来评估）的结果不是那么好（环境归趋稍微准确一点），但图17和图19显示fr\_azide是和它们非常相关的分子描述符，得分遥遥领先于其它的分子描述符。fr\_azide表示含有叠氮基团的数量。叠氮化合物是有毒物质，暴露于叠氮化物化合物会导致线粒体中细胞色素 c 氧化酶的抑制，导致急性致命中毒；并且直至现在叠氮化合物都没有解毒剂173、174。当然，2020年的一项研究显示提前注射一种钴希夫碱复合物，可以防止叠氮化合物中毒175。这个发现在日后化学农药的合成中可能有指导意义。

使用不同的机器学习方法建立QSRR模型预测不同农药在HILIC色谱柱上的保留时间得到了不错的结果（如图11、12所示）；使用相关性统计量加梯度提升决策树的方法可以极好地预测皮质醇、脱氧胸腺嘧啶核苷酸(Dtmp)、肉豆蔻酸、异己酸等代谢物的保留时间，误差范围小于0.01；而使用互信息和随机森林方法可以极好地预测十一烷酸、十七烷酸、中酒石酸、4-羟基苯甲酸、硬脂酸、十五烷酸、D-甘油-3-磷酸、L-2-氨基丁酸、亚油酸、棕榈酸、生物蝶呤、4-乙基苯甲酸、肉豆蔻酸、L-缬氨酸、L-缬氨酸这些代谢物的保留时间，误差范围小于0.01分钟。使用不同的特征选择方法和机器学习方法的组合建立QSRR模型，可以得到和HILIC色谱柱上息息相关的分子描述符。如图9所示，使用互信息+多元线性回归的QSRR模型只需要三种分子描述符（MolMR、PEOE\_VSA10、NOCount）就可以预测。SVM预测保留时间则仅用到了NOCount和MolMR这两个分子描述符。MolMR是分子的相对质量、NOCount是分子中氮氧原子数量、PEOE\_VSA10是分子表面处于10 Å范围内的正电子密度。HILIC（亲水相互作用色谱柱）是一种色谱分离技术，主要用于分离强极性化合物；近几年有不少文章使用各种机器学习方法预测化合物在HILIC色谱柱上的保留时间176、177，但这些研究的重点都在于如何使得结果更准确。在这篇文章里，还关注了哪些因素会影响预测结果。

尽管多元线性回归和SVM以很少的分子描述符预测了保留时间，随机森林预测保留时间却是更准确的一种方法。图10表示了随机森林预测保留时间用到的分子描述符，除了MolMR、PEOE\_VSA10、NOCount，NumHeteroatoms（分子中的杂原子数）、Estate\_VSA（分子表面上的电子状态）、BCUT2D\_MRHI（分子的最高的摩尔折射率对应的负荷本征值）也是很大的影响因素。

# 小结

在这项研究中，我们首先通过QSAR建模，发现比起其他分子描述符，农药中叠氮基团的个数很显著地影响了环境归趋和人类健康，这是一个很新的发现；同时，农药的亲水性/亲脂性决定了它的生态毒性，可能是因为生态毒性大部分时候由对鱼类或是藻类等水生生物的影响决定。

然后，我们通过QSRR模型，预测了367个代谢物在HILIC色谱柱上的保留时间。保留时间受到分子量、分子中氮氧原子数量、分子表面上的电子状态等因素的影响；使用互信息的特征选择方式加随机森林的建模方法预测的保留时间被认为是最准确的。预测这些代谢物的时间可以便于做农药代谢组实验时，更好地鉴定代谢物，以准确的分析农药对模式生物产生具体影响。

最后，结合QSAR模型和QSRR模型，可以更好地预测农药对环境或是人体的影响。