分类号： 单位代码：10019

密 级： 学 号：S20213101981



学位论文

结合机器学习与代谢组学的方法再分析农药暴露对环境归趋、生态毒性和人类健康的影响

**Further analysis of the effects of pesticide exposure on environmental fate, ecotoxicity, and human health through a combination of machine learning and metabolomics methods**

研究生： 乐衣凡

指导教师： 朱文涛 副教授

申请学位门类级别： 理学硕士

专业领域名称： 农药学

研究方向： 农药分析与环境安全

所在学院： 理学院

2024年 4 月

独 创 性 声 明

本人声明所呈交的学位论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中已经注明引用和致谢的内容外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含本人为获得中国农业大学或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表达了谢意。

学位论文作者签名： 时间： 年 月 日

关于学位论文使用授权的说明

本人完全了解中国农业大学有关保留、使用学位论文的规定。本人同意中国农业大学有权保存及向国家有关部门和机构送交论文的纸质版和电子版，允许论文被查阅和借阅；本人同意中国农业大学将本学位论文的全部或部分内容授权汇编录入《中国博士学位论文全文数据库》或《中国优秀硕士学位论文全文数据库》进行出版，并享受相关权益。

**(保密的学位论文在解密后应遵守此协议)**

学位论文作者签名： 时间： 年 月 日

导师签名： 时间： 年 月 日

**摘 要**

农药暴露对环境和人类的影响在近年来受到广泛关注。农药的持久性、生态毒性、对人类健康的影响不可忽视。近年来研究人员使用各种方法研究农药暴露对诸如鱼、小鼠、大鼠等模式生物的影响。然而，这些研究都是针对单个或者几个农药。本文的第一部分基于PPDB数据库建立了QSAR模型，使用不同的分子描述符建立不同的机器学习分类模型（随机森林，支持向量机，逻辑回归，朴素贝叶斯，K近邻，梯度提升决策树），通过953个农药的分子结构预测农药暴露对环境归趋、生态毒性和人类健康的影响；同时建立了使用分子描述符建立了基于7种机器学习方法（支持向量回归、多元线性回归、随机森林、偏最小二乘回归、LASSO回归、岭回归、梯度提升回归）的QSRR模型，预测了367种代谢物在HILIC色谱柱上的保留时间，得到了准确的结果，辅助鉴定代谢物，使得农药暴露对模式生物代谢通路的研究更快速准确。在第二部分的研究中，创作了一个R包meTool以处理农药代谢组学的数据，另外有一个Python软件neuturalloss寻找质谱数据的中性丢失。同时以已经发表的文章“妊娠期烯啶虫胺暴露诱导后代雌性小鼠肠道菌群和粪便代谢物失衡”为例，详细解释了meTool包的使用方法。主要研究结果如下：

1. 有了全新的发现：农药对环境归趋和人类健康的影响很大程度取决于该农药中叠氮基团的数量；而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。
2. 农药的生态毒性主要取决于该农药的亲水性/亲脂性，这可能是因为PPDB数据库中的农药生态毒性由对鱼类或是藻类等水生生物的影响决定。
3. 代谢物在HILIC色谱柱上的保留时间和代谢物的分子量、分子中氮氧原子数量、分子表面上的电子状态等因素密切相关。
4. 随机森林建立的QSAR模型预测环境归趋和生态毒性的效果最好；由互信息的特征选择方式和随机森林的机器学习方法建立的QSRR模型能最好得预测367个代谢物在HILIC色谱柱上的保留时间，预测全部数据的平均绝对误差仅仅为0.616，并且仅仅用了63个分子描述符，便于解释。
5. R包meTool的优势在于：它清晰得显示了实验组和对照组数据PCA和PLS-DA等模型在主成分1和主成分2上的分布；可以和肠道菌群联合分析，这是其他代谢组学工具里没有的；比起SIMCA-P，它是免费的，且功能更多；比起R包Ropls，它作图更美观且自由度更高；比起metaboanalyst 5.0，它克服了网络卡顿的烦恼，而且同样的，自由度更高。当然，meTool也有局限性，它也有其他代谢组学工具没有的功能。
6. neuturalloss作为一款寻找质谱中性丢失的软件，它很好解决了原先寻找中丢失依赖全部人力的问题。

**关键词：**农药，生态毒性，叠氮基团，HILIC色谱柱，数据处理

**Abstract**

In recent years, the impact of pesticide exposure on the environment and human health has been widely recognized. The persistence of pesticides, ecotoxicity, and their effects on human health cannot be ignored. In recent years, researchers have used various methods to study the effects of pesticide exposure on model organisms such as fish, mice, and rats. However, these studies are mostly focused on individual or a few pesticides.

In the first part of this article, QSAR models were established based on the PPDB database. Different machine learning classification models (random forest, support vector machine, logistic regression, naive Bayes, K-nearest neighbors, gradient boosting decision tree) were built using different molecular descriptors to predict the effects of pesticide exposure on environmental fate, ecological toxicity, and human health using the molecular structures of 953 pesticides. Additionally, QSRR models were established using molecular descriptors and seven machine learning regression methods (support vector regression, multiple linear regression, random forest, partial least squares regression, LASSO regression, ridge regression, gradient boosting regression) to predict the retention times of 367 metabolites on HILIC columns. Accurate results were obtained, aiding in the identification of metabolites and facilitating faster and more accurate studies of pesticide exposure on metabolic pathways in model organisms.

In the second part of the study, an R package “meTool” was developed to process pesticide metabolomics data, along with a Python software “neuturalloss” for finding neutral losses in mass spectrometry data. Using the published article "Pregnancy pyrethroid exposure induces dysbiosis in offspring female mice gut microbiota and fecal metabolites" as an example, the usage of the meTool package was explained in detail. The main research findings are as follows:

1. A novel discovery was made: The impact of pesticides on environmental fate and human health largely depends on the number of azido groups in the pesticide. Azido compounds are toxic substances, and exposure to azides can lead to inhibition of cytochrome c oxidase in mitochondria, causing acute lethal poisoning. Furthermore, there are currently no antidotes for azido compounds.
2. The ecotoxicity of pesticides mainly depends on the hydrophilicity/lipophilicity of the pesticide, possibly because the ecological toxicity of pesticides in the PPDB database is determined by their effects on aquatic organisms such as fish or algae.
3. The retention time of metabolites on HILIC chromatographic columns is closely related to factors such as the molecular weight of the metabolite, the number of nitrogen and oxygen atoms in the molecule, and the electronic state on the molecular surface.
4. Random forest-based QSAR models performed the best in predicting environmental fate and ecotoxicity. QSRR models built using mutual information feature selection and random forest machine learning methods performed best in predicting the retention times of 367 metabolites on HILIC chromatographic columns, with an average absolute error of only 0.616, using only 63 molecular descriptors for easy interpretation.
5. The advantages of the meTool package include: It clearly displays the distribution of experimental and control group data on PCA and PLS-DA models in principal component 1 and principal component 2; it can be jointly analyzed with gut microbiota, which is not available in other metabolomics tools; compared to SIMCA-P, it is free and has more functions; compared to the R package Ropls, its plotting is more aesthetic and has higher freedom; compared to metaboanalyst 5.0, it overcomes the frustration of network lagging, and similarly, it has higher freedom. Of course, meTool also has limitations and features that other metabolomics tools do not have.

As a software for finding neutral losses in mass spectrometry data, neuturalloss effectively solves the problem of dependence on manual labor for finding neutral losses in the past.

**Key words:** Pesticides, Ecotoxicity, Azide, HILIC Column, Data Processing

**目 录**

摘要 [Ⅰ](#_Toc523987891)

第一章 绪论 [1](#_Toc523987891)

[1.1 结合机器学习模型研究农药的使用与环境归趋、生态毒性和人类健康 1](#_Toc523987892)

[1.2 烯啶虫胺类农药及农药生物标志物的发现 7](#_Toc523987893)

[1.3 代谢组学方法研究农药环境行为 9](#_Toc523987894)

[1.4 代谢组学的数据处理工具 11](#_Toc523987893)

[1.5 主要研究目的、主要内容及技术路线 1](#_Toc523987893)5

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 17

[2.1引言 17](#_Toc523987896)

[2.2材料和方法 17](#_Toc523987897)

[2.3实验结果 2](#_Toc523987898)7

[2.4讨论 40](#_Toc523987898)

[2.5小结 49](#_Toc523987898)

第三章 农药代谢物组学数据处理工具R包meTool和软件neutralloss 50

[3.1R包meTool 50](#_Toc523987900)

[3.2软件neutralloss 73](#_Toc523987901)

[第四章 结论与展望 77](#_Toc523987903)

[4.1结论 77](#_Toc523987904)

[4.2展望 78](#_Toc523987905)

[参考文献 80](#_Toc523987915)

[致 谢 92](#_Toc523987916)

[作者简介 93](#_Toc523987918)

**插图和附表清单**

插图清单：

[图1- 1 Mt QSAR建模中使用的方法示意图[4] 3](#_Toc164863339)

[图1- 2 QSRR模型的开发、验证、比较和选择示意图[2] 5](#_Toc164863340)

[图1- 3 QSRR模型开发和性能评估[5] 6](#_Toc164863341)

[图1- 4使用Ropls做的PLS-DA可视化[3] 12](#_Toc164863342)

[图1- 5 PlantMetSuite的可视化效果[1] 14](#_Toc164863343)

[图1- 6 技术路线 16](#_Toc164863344)

[图2- 1建立QSAR模型用到的数据，将警告程度High、Moderate和Low再次分组为High和Not High。这样两个组别的数据量相近 19](#_Toc164863345)

[图2- 2 2015-2023年课题组内发表的通过代谢组学方法研究环境农药暴露对模式生物影响的文章的随时间变化的流行话题（A）和词云图（B） 28](#_Toc164863346)

[图2- 3 2015-2023年课题组内发表的通过代谢组学方法研究环境农药暴露对模式生物影响的文章的引用次数（A）及发表的期刊（B） 29](#_Toc164863347)

[图2- 4不同机器学习模型（从上到下依次是：多元线性回归、支持向量回归、偏最小二乘回归、随机森林回归、LASSO回归、岭回归和梯度上升回归）的特征选择数与平均绝对误差的关系。 33](#_Toc164863348)

[图2- 5不同特征选择方法和机器学习方法（从左往右：互信MI+多元线性回归MLR，相关性统计量CS+支持向量回归SVR，MI+SVR）挑选出的最优特征 35](#_Toc164863349)

[图2- 6不同特征选择方法（相关性统计量CS、互信息MI）和随机森林RF的机器学习方法挑选出的最优特征。 36](#_Toc164863350)

[图2- 7相关性统计量CS+梯度提升回归GBR建立的模型预测的化合物的保留时间（深蓝）和实验所得的保留时间（浅蓝），按照二者MAE从小到大排列 37](#_Toc164863351)

[图2- 8互信息MI+随机森林回归RF建立的模型预测的化合物的保留时间（深绿）和实验所得的保留时间（浅绿），按照二者MAE从小到大排列 39](#_Toc164863352)

[图2- 9六种机器学习方法用分子描述符预测全集953种农药对环境归趋的警告程度的混淆矩阵（图A随机森林、B支持向量机、C逻辑回归、D朴素贝叶斯、EK近邻、F梯度提升决策树）和ROC曲线（G） 40](#_Toc164863353)

[图2- 10六种机器学习方法用分子描述符预测全集953种农药对生态毒性的警告程度的混淆矩阵（图A随机森林、B支持向量机、C逻辑回归、D朴素贝叶斯、EK近邻、F梯度提升决策树）和ROC曲线（G） 41](#_Toc164863354)

[图2- 11六种机器学习方法用分子描述符预测全集953种农药对人类健康的警告程度的混淆矩阵（图A随机森林、B支持向量机、C逻辑回归、D朴素贝叶斯、EK近邻、F梯度提升决策树）和ROC曲线（G） 42](#_Toc164863355)

[图2- 12随机森林建立QSAR模型预测不同农药对环境归趋的警告程度所用到的特征（按照ANOVA检验得分由高到低排列） 45](#_Toc164863356)

[图2- 13随机森林建立QSAR模型预测不同农药对生态毒性的警告程度所用到的特征（按照ANOVA检验得分由高到低排列） 45](#_Toc164863357)

[图2- 14支持向量机建立QSAR模型预测不同农药对人类健康的警告程度所用到的特征（按照ANOVA检验得分由高到低排列） 48](#_Toc164863358)

[图3- 1 R包meTool的功能总览：主成分分析、偏最小二乘分析、火山图注释、代谢物相关性分析、基因代谢物联合分析、肠道菌群/代谢物相对丰度热图、肠道菌群-代谢物联合分析以及一些计算功能 50](#_Toc164863359)

[图3- 2 meTool包需要的输入文件格式展示：（A）代谢物（B）肠道菌群（C）基因（D）代谢通路 51](#_Toc164863360)

[图3- 3 PCA图：横坐标为主成分1，纵坐标为主成分2 53](#_Toc164863361)

[图3- 4 PLS-DA图：CK-L和CK-H 54](#_Toc164863362)

[图3- 5 CK和L组以及CK和H组的差异代谢物火山图 55](#_Toc164863363)

[图3- 6 VIP值棒棒糖图 58](#_Toc164863364)

[图3- 7 差异代谢物的丰度热图 59](#_Toc164863365)

[图3- 8 差异肠道菌群的丰度热图 60](#_Toc164863366)

[图3- 9 所有代谢物和菌群的相对丰度热图 62](#_Toc164863367)

[图3- 10 代谢物相关性图 63](#_Toc164863368)

[图3- 11肠道菌群相关性图 65](#_Toc164863369)

[图3- 12 代谢物和肠道菌群的相关性图 66](#_Toc164863370)

[图3- 13 肠道菌群和代谢物mantel检验相关性图 67](#_Toc164863371)

[图3- 14肠道菌群和代谢物二维丰度分布图 69](#_Toc164863372)

[图3- 15批量输出代谢物相对丰度图 70](#_Toc164863373)

[图3- 16 受到影响的代谢通路图 71](#_Toc164863374)

[图3- 17 代谢物-基因关联分析 72](#_Toc164863375)

[图3- 18 软件neutralloss的初始界面 74](#_Toc164863376)

[图3- 19 输入软件的内容 75](#_Toc164863377)

附表清单：

[表2- 1建立QSRR和QSAR模型用到的分子描述符及其含义 26](#_Toc164863378)

[表2- 2不同机器学习方法建立QSAR模型对环境归趋、生态毒性、人类健康警告程度预测的效果 31](#_Toc164863379)

[表2- 3 不同特征选择和建模方式得到QSRR预测保留时间模型的效果。 32](#_Toc164863380)

**主要英文缩略词表**

|  |  |  |
| --- | --- | --- |
| 英文缩写 | 英文名称 | 中文名称 |
| QSAR | Quantitative structure-activity relationships | 定量结构-活性关系 |
| QSRR | Quantitative structure-retention relationship | 定量结构-保留关系 |
| PPDB | Pesticide Property Database | 农药性质数据库 |
| ROC curve | Receiver Operating Characteristic Curve | 受试者工作特征曲线 |
| AUC | Area under the curve | ROC曲线下面积 |
| MDs | Molecular descriptor | 分子描述符 |
| RF | Random Forest | 随机森林 |
| SVM | Support Vector Machine | 支持向量机 |
| LR | Logistic Regression | 逻辑回归 |
| NB | Naive Bayes | 朴素贝叶斯 |
| KNN | K-Nearest Neighbor | K近邻 |
| GB | Gradient Boosting Tree | 梯度提升树 |
| CS | Correlation Statistics | 相关性数据 |
| MI | Mutual Information | 互信息 |
| MLR | Multiple Linear Regression | 多元线性回归 |
| SVR | Support Vector Regression | 支持向量回归 |
| PLS | Partial Least Squares Regression | 偏最小二乘回归 |
| GBR | Gradient Boosting Regression | 梯度提升回归 |
| MAE | Mean Absolute Error | 平均绝对误差 |
| RT | Retention Time | 保留时间 |
| ANOVA | Analysis of Variance | 方差分析 |
| PCA | Principal Component Analysis | 主成分分析 |
| PLS-DA | Partial Least Squares Discriminant Analysis | 偏最小二乘法判别分析 |
| TN | True Positive | 真阳性 |
| TP | True Negative | 真阴性 |
| FN | False Negative | 假阴性 |
| FP | False Positive | 假阳性 |

# 第一章 绪论

# 结合机器学习模型研究农药的使用与环境归趋、生态毒性和人类健康

### 1.1.1农药的环境归趋、生态毒性和对人类健康的影响

农药的环境归趋（environmental fate）是研究农药在环境中的行为和归宿的过程，即在自然环境中（例如土壤、水体、大气等介质）的分布、转化和去向。农药特性数据库（Pesticide Properties DataBase, PPDB，<http://sitem.herts.ac.uk/aeru/ppdb/en/>）在2007年作为一个免费访问的网站推出, 它包括了2300种农药活性物质和700多种代谢物的数据[6, 7]。在PPDB数据库中，使用持久性，排水流量移动性、地下水普遍性得分（Groundwater Ubiquity Score）等来评估农药的环境归趋警告程度。

生态毒性(Ecotoxicity)，是指外源化学物质进入生态系统后对系统中非人类生物体的受伤能力。是植物毒性、土壤动物毒性和特殊毒性的总称。在PPDB数据库中，使用鱼类和水藻急性/慢性毒性来评估农药的生态毒性。

农药对人类健康的影响是不可忽视的，在PPDB数据库中，常用哺乳动物急性/慢性毒性，生殖、发育、神经影响来评估。

### 1.1.2 分子描述符和QSAR模型

定量构效关系（QSAR）是一种常用的药物设计方法[8]。QSAR建模的有用性和实用性很大程度上取决于其特征，即分子描述符（molecular descriptors, MDs）的能力，然后优化描述符的选择，形成最佳QSAR模型[9]。分子描述符可以分为定量描述符和定性描述符。定量描述包括基于分子图论、各种理论或实验光谱数据（如紫外光谱）、分子组成（如氢键供体数、化学键数）、理化性质（如酯水分布系数）描述符、分子场描述符以及分子形状描述符等；定性描述符一般称为分子指纹，即将分子的结构、性质、片断或子结构用某种编码来表示。用于计算MDs的软件很多，比如PaDEL-Descriptor[10]，Mordred[11]，Dragon software[12]。也有一些基于变成语言的开源库用于计算MDs，比如基于JAVA的CDK[13]，基于R的rcdk包[14]，和基于python的rdkit[15]。而农药环境风险评估中常用的计算毒理学软件有EPI Suite[16]、QSAR Toolbox[17]、PBT Profiler[18]、PRZM-GW[19]、China-PEARL[20]、

在农药方面，QSARs已经被用于预测农药代谢物急性鱼类毒性（基于LC50值）[7]；早在2000年，有研究预测农药对虹鳟*Oncorhynchus mykiss*的急性毒性[21]；在2020年，又有研究建立了新的QSAR模型预测农药对虹鳟的急性毒性[22]；在2001年，预测了农药对蓝鳃太阳鱼*Lepomis macrochirus*的急性毒性[23]；其他一些研究使用QSAR模型预测不同杀虫剂对蜜蜂的毒性[24-26]；预测包含农药在内的诸多有机化学品对鲤科鱼类的毒性并得到氯氟氰菊酯毒性极强的结论[27]；模拟了农药对羊头鱼*Sheepshead minnow*的毒性；建立有机磷水生和陆生生物生态毒性的QSAR模型预测了新烟碱类药物对蜜蜂、家养麝香、美洲蟑螂和蚜虫（*Aphis craccivora* 和 *Myzus persicae*）的毒性[28]；使用基于17个分子描述符的QSAR模型预测了329种农药对于大鼠的急性口服毒性[29]；更有研究使用多尺度的QSAR模型预测农药的生态毒性并正确预测了超过75%的数据[30, 31]；其他研究使用Levenberg-Marquardt（LM）算法构建了易于解释的QSAR模型，用于同时预测哺乳动物和鸟类LD50和LD50[4]。QSAR模型也可以用于验证新型农药的药效：比如，最近有研究通过QSAR模型验证新型含肟醚香豆素衍生物可能作为新型杀菌剂[32]。也有通过QSAR模型寻找对蚯蚓更低毒的新烟碱类农药替代品[33]，辅助开发白纹伊蚊萜类驱虫剂[34]，指导开发新型肉桂酸衍生物作为杀菌剂[35]，验证几种化合物作为几丁质脱乙酰酶（CDA）抑制剂的新型杀菌剂的活性[36]。其他研究开发了一对一与全量构效关系（OvA-QSAR）模型用于估计化学农药的LD50值，使用朴素贝叶斯（NB），顺序最小优化（SMO），随机森林（RF）等[37]。还有使用2D分子描述符建立基于偏最小二乘（PLS）的QSAR模型预测农药对狗的亚慢性和慢性毒性[38]。还可以使用QSAR模型结合其他机器学习方法预测各种化学物质对心脏的毒性[39]。

|  |
| --- |
| Fig1- 1 Schematic diagram of the methods used in Mt QSAR modeling.  图1- 1 Mt QSAR建模中使用的方法示意图[4] |

QSAR模型的建立如图1-1所示[4]，大致过程是：首先收集了229种杀虫剂的环境归趋、生态毒性、化学物质亲脂性（LogP）、生物浓缩因子(BCF)、致死剂量50（LD50小鼠和LD50鸟类）数据，计算这些农药的分子描述符、选择特定的分子描述符作为特征，拆分训练集和测试集，得到结果，最后选择最好的模型。

### 1.1.3 QSRR模型和代谢物保留时间的预测

定量结构-（色谱）保留关系Quantitative Structure-(Chromatographic) Retention Relationships被用来广泛得辅助处理色谱数据[40]，已经使用过的QSRR的具体方法有分类和回归树 （CART）、基于树的模型的随机梯度提升 （Treeboost）、随机森林 （RF）、无信息变量消除偏最小二乘法 （UVE-PLS） 和多元线性回归遗传算法 （GA-MLR）[41]。QSRRs和QSARs是将内部化学结构和特定生物活性联系起来的相对较新的方法，它们可以联合运用以计算不同取代的苯并咪唑衍生物的色谱特性和抗菌活性[42]、抗肿瘤药物的生物活性[43]、预测吖啶酮衍生物与DNA的理化相互作用[44, 45]。早在2010年，就有研究使用QSRR模型作为植物生物体农药残留测定和建模的分析工具[46]，近年来，又有研究使用7个分子描述符结合支持向量回归建立QSRR模型以准确预测反相液相色谱中农药保留时间 [2]。有研究使基于METLIN数据库使用不同的特征选择方法（遗传算法GA、逐步算法和Boruta算法）和不同的机器学习方法（支持向量机SVM、多元线性回归MLR、随机森林RF和XGBOOST）建立QSRR模型，提供了对色谱的保留时间有不错预测能力的模型，其中用GA的特征选择方式配合MLR的机器学习模型得到了最优的预测结果[47]。其他研究利用分析物的结构计算了800多个描述符，再使用这些描述符模拟四种色谱条件下分析物的保留时间，其中随机森林模型是最优的[48]。黄酮类化合物可以试用其氢键能（XAH）和溶出能（ES）的线性关系，预测其在超高效液相色谱-串联质谱（UHPLC-MS/MS）的保留时间[49]。基于SVM的QSRR模型可以很好得预测C-18色谱柱上抗糖尿病药物吡格列酮和和格列美脲的保留时间[50]；对芫荽和鼠尾草精油化合物的保留时间（RT）进行预测[51]。使用106种参比化合物，建立了基于梯度增强机的QSRR模型，可靠地鉴定了421种人参皂苷[52]。其他文献预测了非法添加剂的保留时间，在特征选择阶段比较了最小冗余最大相关性（MRMR）和F检验两种选择算法，使用回归树 （Reg-T）、支持向量机 （SVM）、高斯过程回归模型 （GPR）、树集成和核近似模型建模，其中指数高斯过程回归模型（E-GPR）精确预测了一些非法添加剂小分子的保留时间[53]。

除了预测保留时间，QSRR模型还可以用于解释和预测分子对磷脂的亲和力，这种亲和力主要取决于亲脂性和电[54]；预测拟除虫菊酯类化合物的色谱特性，反向传递人工神经网络(BP-ANN)是最优的QSRR模型[55]。通过QSRR模型，发现查尔酮衍生物在人血清白蛋白 （HSA）固定相上的保留特性取决于其结构和电子性质[52]；预测抗真菌药物异恶唑[3,4-b]吡啶-3（1H）-酮（Isoxazolo[3,4-b]pyridine-3(1H)-Ones to Phospholipids）对磷脂的亲和性[56]；作为色谱测定β-环糊精络合过程稳定性常数和热力学参数的潜在工具[57]。还可以通过QSRR模型提高中药小分子结构类似物的鉴定效率，比如黄柏和关黄柏生物碱类成分[58]。近年来，还有很多文献使用不同方法优化了QSRR模型，改进质谱分析种的特征标注过程[59, 60]。

|  |
| --- |
| Fig1- 2 Development, validation, comparison, and selection diagram of QSRR model  图1- 2 QSRR模型的开发、验证、比较和选择示意图[2] |

有研究详细描述了农药保留时间的QSRR模型的具体建立过程见图1-2：首先，收集了843个农药和它们保留时间的数据，然后阅读一些相关文献，接着挑选和计算了一些分子描述符，然后对描述符数据清洗和标准化，把数据分为训练集和测试集，然后建立模型（线性回归LR、多元线性回归MLR、判别分析回归PLS-R、支持向量回归SVM-R、和多层感知机回归MLP-R），比较预测结果，最后选择最优的QSRR模型来预测农药的保留时间[2]。其他研究也有描述使用QSRR模型预测农药的保留时间（图1-3），方法相似，只不过使用不同方法选择特征（Lasso，Pearson，RFE，PCA）以及不同的机器学习模型（深度神经网络DNN）建立QSRR模型[5]。

|  |
| --- |
| 图1- 3 QSRR模型开发和性能评估[5]  Fig1- 3 QSRR model development and performance evaluation |

# 烯啶虫胺类农药及农药生物标志物的发现

### 烯啶虫胺（nitenpyram）的使用概况

烯啶虫胺是新烟碱类杀虫剂包括七种商业上销售的活性成分之一[61]。新烟碱类杀虫剂作用于昆虫烟碱乙酰胆碱受体nAChRs[62]。烯啶虫胺对于湘北棉区大花蝉（半翅目：蝉科）的防治效果优于吡虫啉，劣于阿维菌素，有着不错的防治效果[63]。对于淡色金蝇（神经翅目：金蝇科）的防治，烯啶虫胺通过调控热休克蛋白、fluoxetine protein 6和prophenoloxidas达到杀虫效果[64]。烯啶虫胺还可以通过破坏褐稻虱(Nilaparvata lugens)的微生物群落来达到杀虫的目的[65]；但褐飞虱已经逐渐对烯啶虫胺产生抗性，可以通过补充多效唑来提升烯啶虫胺对褐飞虱的杀虫效率[66]。虽然绿盲蝽对很多农药产生了抗性，烯啶虫胺对绿盲蝽依旧有着不错的杀虫效果[67]。

因为烯啶虫胺是水溶性的，它会通过径流从农业土壤迁移到水中，从而导致蔬菜水果从土壤中吸取更多此类农药；而青少年与老年人相比，更容易摄取水果中的新烟碱类农药[68]。多篇文献报道，烯啶虫胺会导致肝损伤、生殖系统氧化应激标志物功能障碍和肠道毒性[69]。2024年的一项研究表明，中国青少年尿液中的烯啶虫胺含量和BMI z评分以及肥胖几率成正比[70]，而尿液中烯啶虫胺含量高的儿童BMI z 评分较低[71]。2023年广西的一项研究显示，胎儿生长受限的风险与孕妇血清中烯啶虫胺的含量负相关，似乎烯啶虫胺对胎儿和儿童的危害并不是很大[72]。2023年河南的一项研究表明水果和蔬菜中的新烟碱类农药对人体不会有显著影响，但血浆中的烯啶虫胺含量与2型糖尿病发病率相关[73]。

### 机器学习、生物标志物在人类健康研究中的应用

机器学习（ML）带来了从具有丰富生物医学测量结果的队列中提取的新生物标志物的希望，一个好的生物标志物是能够可靠地检测相应疾病的生物标志物[74]。机器学习和人工智能(AI)的最新进展使得识别高度预测的疾病特异性生物标志物成为可能，这种生物标志物可用于诊断癌症患者，预测癌症预后，甚至预测治疗效果[75]。2019年的一项研究表示，机器学习可以很好得降维基于质谱的蛋白质组学数据，发现蛋白质生物标志物[76]。机器学习在骨关节炎的诊断中也有不少作用，2019年的一项研究显示，机器学习可以确定影像学发现、血清生物标志物和症状等变量的重要性来识别代表不同骨关节炎新表型[77]。2020年的研究概述了机器学习辅助神经影像学技术研究大脑结构/功能异常与神经精神疾病之间的相关性，实现对神经影像学数据的个体化预测，挖掘了潜在生物标志物[78]。2020年的另一项研究使用microRNA （miRNA）作为生物标志物建立机器学习模型，有效得区分了黑色素瘤和痣[79]。2020年的另一项研究使用电子鼻捕获人呼出的挥发性气体，结合机器学习模型来预测特应性哮喘[80]。使用机器学习（支持向量机、逻辑回归、随机森林和朴素贝叶斯）的方法，2021年的一项研究找到了一批关于阿尔兹海默症的政务标志物并用这些生物标志物建立了诊断模型，该模型具有高灵敏性和特异性[81]。2022年也有研究使用利用大脑中的生物标志物建立机器学习模型，很好的区分有症状的阿尔茨海默痴呆患者与轻度认知障碍和正常认知的患者[82]。2022年的另一项研究使用九个血液中的生物标志物，建立了支持向量机（SVM），决策树的极限梯度提升（XGB）和人工神经网络（ANN）三个机器学习模型，还算准确地预测了1642名痴呆患者和正常人[83]。2021年的另外一项研究开发了基于机器学习的糖尿病诊断算法，通过输入空腹血糖和血红蛋白A1c等生物标志物可以预测糖尿病[84]。其他研究使用机器学习诊断急性肾损伤的生物标志物，辅助实现精准用药[85, 86]。结合基因组学，人工智能和机器学习可以更好得实现精准医疗[87]。基于肝血浆蛋白组学数据，机器学习能够生成生物标志物面板来检测肝纤维化、炎症和脂肪变性[88]。2022年的一项研究表示，人工智能算法和统计方法在利用组学数据和揭示癌症免疫功能机制方面显示出巨大的潜力[89]。2023年的一项研究使用四种不同的机器学习方法（SVM、XGBoost、RF 和 kNN），将结肠组织中的五个基因（INHBA、FNBP1、PDE9A、HIST1H2BG、CADM3）作为结直肠癌 （CRC）的生物标志物，很好地预测了CRC患者[79]。然而，机器学习也有不好用的时候，比如它无法识别抑郁症的生物标志物[90]。

在农药的检测上，机器学习可以结合同步荧光传感方法快速定量红酒中噻苯达唑和呋喃咪唑[91]；使用近红外光谱结合逻辑回归（LR）可以以97%的准确度预测出葡萄中的农药残留水平[92]；结合近红外光谱法和机器学习，可以对白菜的农药残留进行无损检测[92]；荧光光谱结合宽度学习，可以准确测定白菜中吡虫啉的残留量[93]；结合一种新型电化学传感器和机器学习，可以快速检测出农药中残留的多菌灵[94]；荧光高光谱技术结合机器学习，准确测定出红茶中联苯菊酯、二芬硫脲、托芬吡喃、吡虫啉等农药的残留[95]；机器学习辅助荧光传感器阵列，可以100%准确分类以及鉴定果蔬上的四种不同拟除虫菊酯类农药（溴氰菊酯、氰戊菊酯、氯氟氰菊酯和芬丙病林）残留[96]；高光谱成像技术结合极端学习机（ELM）可以无损得检测出哈密瓜表面的不同农药残留[97]。机器学习还可以优化农药的喷洒，以避免农药的重复使用对环境带来不利影响[98]，预测植物种农药的半衰期[99]，评估杀虫剂暴露对于农民的耳部毒性[100]；监测农场农药滥用[101]。

生物标志物在农药对环境和人类影响中也早有应用。人体农药慢性暴露的生物标志物包括了化学物质（杀虫剂）和代谢产物[102]。2020年一篇综述描述了118个杀虫剂生物标志物，而只有只有 67 种被证实是不同于农药的生物标志物[103]。其中，乙酰胆碱酯酶和丁酰胆碱酯酶活性被认为是农民农药中毒的生物标志物[104]。2014年的研究提出氧化应激生物标志物作为农药毒性替代标志物在农药处理者（如生产农药的化学工业、喷洒农药的人和从事农业工作的人）中的重要性[105]。2015年的研究通过检测了农药对应生物标志物的在尿液、母乳、粪便中的含量得到当地居民持续长期暴露于这些化学物质的结论[106]。2020年的一项研究通过测定瑞典青少年的尿液中农药对应的生物标志物，确定了瑞典青少年暴露于低农药残留环境中[107]。2020年的另外一项研究通过评估四种生物标志物乙酰胆碱酯酶（AChE），过氧化氢酶（CAT），谷胱甘肽S-转移酶（GST）和碱性磷酸酶（ALP）发现了农药对于蜜蜂的冬季生存带来了问题[108]。2023年的研究通过测量血清中农药生物标志物得出农药暴露因性别、农药使用季节、酒精、SEP、居住纬度而异的结论[109]。巴西的一项研究发现是否农业劳动者（是否接触农药）只与氧化应激生物标志物中的一个：谷胱甘肽过氧化物酶有关[110]。通过检测血浆中的效果生物标志物（AChE和BChE活性）和暴露生物标志物（各种农药及其代谢物）验证了阿根廷的农药暴露指数[111]。2022年的一篇综述使用多种生物标志物（乙酰胆碱酯酶活性AChE、酸性磷酸酶ACP和碱性磷酸酶ALP、超氧化物歧化酶SOD、过氧化氢酶CAD、谷胱甘肽S-转移酶GST）评估了农药对蚯蚓的影响[112]。诸如蚤状钩虾Gammarus pulex中的细胞色素CYP1A1、CAT 和 GST的生物标志物可被用于跟踪特定农业土壤细菌（用于农药污染农田生物修复的接种物）的存活和效率[113]。

# 代谢组学方法研究农药环境行为

传统的评价农药的生态毒性的方法观测指标有限，评价效率低，灵敏度不足，不能完全满足当前的评价需求；作为一种新的研究方法，代谢组学从整体层面研究生物体的代谢活动和状态[114, 115]。近年来，代谢组学已被广泛应用于研究环境污染物（包括农药）对动植物的毒理机制，具有较高的灵敏度[116]。代谢组学与其他组学技术的结合将有助于探索环境污染物的毒性机制，并为环境污染物的毒理学评价提供新的研究思路[117]。多组学（基因组学、转录组学、蛋白质组学和代谢组学等组学）研究对于生成有关农药降解的基因和蛋白质、微生物农药降解产生的代谢物以及应对农药暴露引起的压力的细胞策略的相关信息至关重要[118]。代谢组学还是探索和利用微生物化学生态学的重要工具，微生物分泌多种分子来影响周围的生物和环境；可以通过调节微生物，在不严重依赖合成杀虫剂的情况下最大限度地减少病虫害造成的作物损失[119]。不过目前，环境毒理学中代谢组学的研究主要集中在单一污染物（重金属、有机污染物、抗生素、农药）上，还应考虑污染区的复杂情况[114]。

最先进的研究农药暴露的代谢组学工具有GC-MS、LC-MS/MS UHPLC、UPLC-IMS-QToF、GC/EI/MS、MALDI-TOF MS和H-1-HR-MAS NMR等；代谢组学可用于在暴露组测定和实验室研究中评估农药对农艺重要作物的毒理学影响[120]。2021年的研究通过代谢组学测定茶叶代谢物来评估茶叶的遮荫条件和农药处理方法[121]。通过空间代谢组学，评估了新烟碱类呋虫胺对蜜蜂的立体选择性毒性，得出S-（+）-呋虫胺毒性强于R-（-）-呋虫胺[122]。通过代谢组学，还可以评估农药如何通过饮食影响儿童健康[123]。2020年有研究采用气相色谱串联质谱（GC/MS）和Q-Exactive质谱仪（QE）检测母体血液中37种农药，发现混合农药暴露与出生体重总体呈负相关；乙型六氯环己烷和甲氨酸可破坏甲状腺激素代谢和甘油醛代谢，从而降低出生体重[124]。2022年的一项研究显示，葡萄用五种农药（己康唑、苯醚甲环唑、氟三酚、戊唑唑和丙环唑）的试验中，改变了葡萄酒中超过 86 种代谢物，这些代谢物大多是天然风味化合物（碳水化合物、氨基酸和短链脂肪酸及其衍生物），决定了葡萄酒的外观、香气、风味和口感[125]。2023年的一项研究通过液相色谱联质谱联用（UPLC-MS）对职业暴露和非暴露个体的血浆和尿液进行代谢组学分析，得到的17种代谢物可以很好得作为区分这两类人群的生物标志物[126]。2023年的另一项研究通过液相色谱联用高分辨率质谱（UPLC-HRMS）进行了非靶向土壤代谢组学，测量了草甘膦、2,4-D和苯醚甲环唑三种农药对这些土壤的吸附和解吸系数，然后建立模型通过农药的m/z和保留时间以预测农药在土壤中的吸附和解吸附系数，得到了不错的结果[127]。基于1HNMR的代谢组学发现大鼠接触农药二嗪农、乐果和氯氰菊酯可引起脂质和氨基酸代谢紊乱、氧化应激诱导、肝肾功能障碍等[128]。通过对英国的65对双胞胎的尿液靶向代谢组学分析。发现所有尿液样本均检出拟除虫菊酯和/或有机磷杀虫剂残留，53%的个体检出除草剂草甘膦[129]。基于GC-MS的非靶向代谢组学发现蚯蚓长期暴露于毒死蜱、氯氰菊酯、草甘膦后，多种代谢物发生显著上调或下调[130]。代谢组学在水稻害虫研究中也有不少应用,包括昆虫抗药性研究、昆虫-水稻相互作用研究及昆虫-共生菌相互作用研究等[131]。通过空间代谢组学发现鱼藤酮显著影响小菜蛾（*Plutella xylostella*）的嘌呤和氨基酸代谢以达到防治的效果[132]。代谢组学和质谱成像揭示了杀虫剂吲哚威对成年斑马鱼（*Danio rerio*）肝脏的慢性毒性[133]。

代谢组学也可以揭示一些物质对于农药暴露后果的预防。比如，槲皮素对四种有机磷农药混合物（乐果、乙酰甲胺磷、敌敌畏、甲拌磷）诱导的肾毒性的预防作用[134]。2022年的代谢组学分析显示，农药和叶面肥混合使用可显著提高黄瓜果实中有机酸含量和抗氧化水平，促进农药的消散[135]。

# 代谢组学的数据处理工具

通常，代谢组学的数据处理方法有无监督的主成分分析PCA，有监督的偏最小二乘分析PLS-DA、正交的偏最小二乘分析OPLS-DA，然后使用置换检验中的R2和Q2值作为质量指标检验多变量分析结果的好坏，然后计算VIP值以此寻找特征和异常值 [136, 137]。

|  |
| --- |
| Fig1- 4 Visualization of PLS-DA through R package Ropls  图1- 4使用Ropls做的PLS-DA可视化[3] |

常用的代谢组学数据处理工具比较早期的有R包Ropls[3]、SIMCA-P[138]，还有一直从1.0发展到5.0的MetaboAnalyst [139-142]。使用Ropls对代谢组学可视化的结果见图1-4，虽然很好得表达出了结果，但在美观性上有待提高[3]。SIMCA-P的弊端在于它是昂贵的收费软件。而MetaboAnalyst作为一款基于Web的代谢组学分析软件，功能比前两者都强大，不过有时候的网络卡顿导致实际使用体验不佳，而且无法手动选择图片的颜色。近年来不乏一些新兴代谢组学分析工具。2021年发表的MSCAT是基于python的、可以帮助搜索机器学习辅助代谢组学的软件工具，但无法直接处理代谢组数据[143]。PlantMetSuite是2023年新发表的一个专用于植物代谢组网络软件，无需编程基础和安装，但可视化效果比较普通，于Metaboanalyst等并无差别（图1-5）[1]。spectrum\_utils是在2023年发表的一个python组件，然而它只能实现质谱数据的可视化，而对随后的进一步数据处理没有帮助[144]。mwtab Python 库是一个新数据库，用于管理各种格式储存的代谢组数据，但并不用于数据可视化[145-147]。MESSES是另一个基于Python的模块，用于整理凌乱的代谢组学原始数据，使其能使用主流的代谢组学软件进一步处理或是存入在线的代谢组学数据库[148]。TidyMS，另一个基于Python的模块，用于预处理非靶向代谢组学工作流程的 LC-MS 数据[149]。Xconnector基于Python，可以解析来自人类代谢组数据库（HMDB）、牲畜代谢组数据库（LMDB）、酵母代谢组数据库（YMDB）、毒素和毒素靶标数据库（T3DB）、ReSpect植物化学物质数据库（ReSpectDB）、血液暴露组数据库、苯酚浏览器数据库、京都基因和基因组百科全书（KEGG）和小分子通路数据库（SMPDB）的信息，输出这些数据为csv表格[150]。DBDIpy是一个 Python 库，用于处理和鉴别来自实时等离子体电离质谱的非靶向数据集[151]。UmetaFlow是一个集预处理以及处理为一体的的python软件，当然在数据可视化上依旧略微欠缺[152]。Phenonaut是一个 Python 软件包，可以处理多组学（包括代谢组学数据）并且可视化，但可视化功能有限[153]。当然要考虑到能可视化代谢组学的Python软件或者模块非常之少。

|  |
| --- |
| Fig1- 5 PlantMetSuite to visualize  图1- 5 PlantMetSuite的可视化效果[1] |

基于R的软件包在可视化方面有一些优势，可能得益于ggplot2。2018年的MetaboDiff是用于差异代谢组学分析的 R 包，尽管可视化效果比较粗糙，但是功能较多[154]。MobilityTransformR是一个对基于毛细管区电泳质谱( CE-MS)的代谢组数据分析的R 包，虽然有一些特定功能，但它只能用于CE-MS[155]。Maplet是可用于模块化和可重现的代谢组学R 包，虽然可视化效果依旧一般，但它可以记录所有的数据处理步骤，使结果可重现[156]。MetaboAnalystR是MetaboAnalyst5.0的R包版，具有相同的功能，但它不像网页版那么卡[157]。不同于仅仅处理或预处理代谢组学数据，Lilikoi包的特点在于它基于个性化途径的方法，使用深度学习和代谢组学数据对疾病进行分类，值得一提的是它在2018年已经诞生，在2023年有了Lilikoi v.2.0版本[158, 159]。2024年的imputomics既有网页版也有R包版，它的长处在于插补代谢组数据中的缺失值，并且有着不错的可视化效果[160]。R中的代谢组学数据处理包还有注释的功能，比如基于 MetaboCoreUtils、MetaboAnnotation 和 CompoundDb 包，有着一套R包的生态系统，以MSP、MGF、mzML、mzXML、netCDF 以及 MassBank 文本文件和 SQL 为数据存储格式，注释质谱的MS1和MS2数据[161]。

# 主要研究目的、主要内容及技术路线

### 主要研究目的

近年来，对农药环境和健康效应的研究非常之多，绝大部分研究集中于单个或几个农药的环境或健康效应，在这里，我们希望用大数据的处理方式，从宏观上研究农药对环境和健康影响的因素。同时，我们希望开发简单好用的农药代谢组学软件，方便数据的分析，直观地展现农药影响的代谢物和肠道微生物组数据。

### 主要内容

1. QSAR模型：以分子特征描述符为特征，筛选这些特征并且建立不同的机器学习模型预测农药的环境归趋、生态毒性和对人类健康的影响，通过ROC曲线下面积等标准判断模型的好坏；并且提取被选中的特征，以此从分子层面上分析农药对环境和健康影响的最重要的因素。
2. QSRR模型：以分子特征描述符为特征，筛选这些特征并且建立不同的机器学习模型预测农药可能影响的代谢物在HILIC色谱柱上的保留时间，以平均绝对偏差为标准选择最优模型；并且分析选中的特征。
3. 创建了R包meTool，创建这个包的初衷在于让刚刚接触R语言的编程者也能快速分析农药代谢组学和肠道菌群数据，并且输出的图片是美观易懂的。
4. 编写了软件neutralloss，该软件可以寻找质谱中性丢失的的离子对，方便农药的快速检测。

### 技术路线

|  |
| --- |
| Fig1- 6 Technical outline  图1- 6 技术路线 |

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[162-165]。

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[166]。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2-1）。分子描述符和它们对应的含义见表2-1。

|  |
| --- |
| Fig2- 1 The data used to establish the QSAR model is grouped again into High and Not High for warning levels High, Moderate, and Low. The data volume of these two groups is similar.  图2- 1建立QSAR模型用到的数据，将警告程度High、Moderate和Low再次分组为High和Not High。这样两个组别的数据量相近 |

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检验获得特征评分，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。

|  |  |
| --- | --- |
| 分子描述符（MDs）名称 | 含义 |
| MaxAbsEStateIndex | 分子中原子的电子态指数的最大绝对值 |
| MaxEStateIndex | 分子中原子的最大电子态指数 |
| MinAbsEStateIndex | 分子中原子的电子态指数的最小绝对值 |
| MinEStateIndex | 分子中原子的最小电子态指数 |
| qed | 用于评估分子的药物性质，值越高表示分子越可能具有药用价值 |
| SPS | 用于评估分子的溶解度，值越高表示分子在水中溶解度越高 |
| MolWt | 分子量 |
| HeavyAtomMolWt | 分子中重原子的分子量（除氢外） |
| ExactMolWt | 分子的精确分子量 |
| NumValenceElectrons | 分子中的总价电子数 |
| NumRadicalElectrons | 分子中的总自由基电子数 |
| MaxPartialCharge | 分子中原子的最大偏离电荷 |
| MinPartialCharge | 分子中原子的最小偏离电荷 |
| MaxAbsPartialCharge | 分子中原子的最大电荷偏离值的绝对值 |
| MinAbsPartialCharge | 分子中原子的最小电荷偏离值的绝对值 |
| FpDensityMorgan1 | Morgan环指纹的密度 |
| FpDensityMorgan2 | Morgan环指纹的密度 |
| FpDensityMorgan3 | Morgan环指纹的密度 |
| BCUT2D\_MWHI | 连接性指数：分子的最高的分子量对应的Burden eigenvalue |
| BCUT2D\_MWLOW | 连接性指数：分子的最低的分子量对应的Burden eigenvalue |
| BCUT2D\_CHGHI | 连接性指数：分子的最高的电荷数对应的Burden eigenvalue |
| BCUT2D\_CHGLO | 连接性指数：分子的最低的电荷数对应的Burden eigenvalue |
| BCUT2D\_LOGPHI | 连接性指数：分子的最高的脂水分配系数LogP的对数对应的Burden eigenvalue |
| BCUT2D\_LOGPLOW | 连接性指数：分子的最低的脂水分配系数LogP的对数对应的Burden eigenvalue |
| BCUT2D\_MRHI | 连接性指数：分子的最高的摩尔折射率对应的Burden eigenvalue |
| BCUT2D\_MRLOW | 连接性指数：分子的最低的摩尔折射率对应的Burden eigenvalue |
| AvgIpc | 分子中原子的的平均化学亲和力 |
| BalabanJ | 分子的芳香性程度 |
| BertzCT | 分子的连接性 |
| Chi0 | 0阶的Molecular connectivity indices，分子中所有原子间连接的数量 |
| Chi0n | Chi0除以分子的原子总数 |
| Chi0v | Chi0的变体，它考虑了分子中每个原子的价电子对数 |
| Chi1 | 一阶的Molecular connectivity indices |
| Chi1n | Chi1值除以分子的原子总数 |
| Chi1v | 类似于Chi0v，它考虑了分子中每个相邻原子对的价电子对数 |
| Chi2n | Chi2值除以分子的原子总数 |
| Chi2v | 类似于Chi0v，它考虑了分子中每个相邻原子对的价电子对数 |
| Chi3n | Chi3值除以分子的原子总数 |
| Chi3v | 类似于Chi0v，它考虑了分子中每个相邻原子对的价电子对数 |
| Chi4n | Chi4值除以分子的原子总数 |
| Chi4v | 类似于Chi0v，它考虑了分子中每个相邻原子对的价电子对数 |
| HallKierAlpha | Hall-Kier Alpha拓扑指数；描述分子的拓扑结构 |
| Ipc | 拓扑晶格指数；描述分子的拓扑结构 |
| Kappa1 | 分子表面曲率的高斯曲率的一半，表示了曲率的总和 |
| Kappa2 | 分子表面主曲率之间的差异的平方的平均值 |
| Kappa3 | 分子表面主曲率的乘积的平均值 |
| LabuteASA | 分子的表面积 |
| PEOE\_VSA1 | 分子表面处于1 Å范围内的正电子密度 |
| PEOE\_VSA10 | 分子表面处于10 Å范围内的正电子密度 |
| PEOE\_VSA11 | 分子表面处于11 Å范围内的正电子密度 |
| PEOE\_VSA12 | 分子表面处于12 Å范围内的正电子密度 |
| PEOE\_VSA13 | 分子表面处于13 Å范围内的正电子密度 |
| PEOE\_VSA14 | 分子表面处于14 Å范围内的正电子密度 |
| PEOE\_VSA2 | 分子表面处于2 Å范围内的正电子密度 |
| PEOE\_VSA3 | 分子表面处于3 Å范围内的正电子密度 |
| PEOE\_VSA4 | 分子表面处于4 Å范围内的正电子密度 |
| PEOE\_VSA5 | 分子表面处于5 Å范围内的正电子密度 |
| PEOE\_VSA6 | 分子表面处于6 Å范围内的正电子密度 |
| PEOE\_VSA7 | 分子表面处于7 Å范围内的正电子密度 |
| PEOE\_VSA8 | 分子表面处于8 Å范围内的正电子密度 |
| PEOE\_VSA9 | 分子表面处于9 Å范围内的正电子密度 |
| SMR\_VSA1 | 分子表面在1 Å范围内的溶剂可及表面积 |
| SMR\_VSA10 | 分子表面在10 Å范围内的溶剂可及表面积 |
| SMR\_VSA2 | 分子表面在2 Å范围内的溶剂可及表面积 |
| SMR\_VSA3 | 分子表面在3 Å范围内的溶剂可及表面积 |
| SMR\_VSA4 | 分子表面在4 Å范围内的溶剂可及表面积 |
| SMR\_VSA5 | 分子表面在5 Å范围内的溶剂可及表面积 |
| SMR\_VSA6 | 分子表面在6 Å范围内的溶剂可及表面积 |
| SMR\_VSA7 | 分子表面在7 Å范围内的溶剂可及表面积 |
| SMR\_VSA8 | 分子表面在8 Å范围内的溶剂可及表面积 |
| SMR\_VSA9 | 分子表面在9 Å范围内的溶剂可及表面积 |
| SlogP\_VSA1 | 分子表面上与脂水分配相关的区域1 |
| SlogP\_VSA10 | 分子表面上与脂水分配相关的区域10 |
| SlogP\_VSA11 | 分子表面上与脂水分配相关的区域11 |
| SlogP\_VSA12 | 分子表面上与脂水分配相关的区域12 |
| SlogP\_VSA2 | 分子表面上与脂水分配相关的区域2 |
| SlogP\_VSA3 | 分子表面上与脂水分配相关的区域3 |
| SlogP\_VSA4 | 分子表面上与脂水分配相关的区域4 |
| SlogP\_VSA5 | 分子表面上与脂水分配相关的区域5 |
| SlogP\_VSA6 | 分子表面上与脂水分配相关的区域6 |
| SlogP\_VSA7 | 分子表面上与脂水分配相关的区域7 |
| SlogP\_VSA8 | 分子表面上与脂水分配相关的区域8 |
| SlogP\_VSA9 | 分子表面上与脂水分配相关的区域9 |
| TPSA | 描述分子表面的极性表面积 |
| EState\_VSA1 | 分子表面上的电子状态相关的区域1 |
| EState\_VSA10 | 分子表面上的电子状态相关的区域10 |
| EState\_VSA11 | 分子表面上的电子状态相关的区域11 |
| EState\_VSA2 | 分子表面上的电子状态相关的区域2 |
| EState\_VSA3 | 分子表面上的电子状态相关的区域3 |
| EState\_VSA4 | 分子表面上的电子状态相关的区域4 |
| EState\_VSA5 | 分子表面上的电子状态相关的区域5 |
| EState\_VSA6 | 分子表面上的电子状态相关的区域6 |
| EState\_VSA7 | 分子表面上的电子状态相关的区域7 |
| EState\_VSA8 | 分子表面上的电子状态相关的区域8 |
| EState\_VSA9 | 分子表面上的电子状态相关的区域9 |
| VSA\_EState1 | 分子表面上的电子状态密度相关的区域1 |
| VSA\_EState10 | 分子表面上的电子状态密度相关的区域10 |
| VSA\_EState2 | 分子表面上的电子状态密度相关的区域2 |
| VSA\_EState3 | 分子表面上的电子状态密度相关的区域3 |
| VSA\_EState4 | 分子表面上的电子状态密度相关的区域4 |
| VSA\_EState5 | 分子表面上的电子状态密度相关的区域5 |
| VSA\_EState6 | 分子表面上的电子状态密度相关的区域6 |
| VSA\_EState7 | 分子表面上的电子状态密度相关的区域7 |
| VSA\_EState8 | 分子表面上的电子状态密度相关的区域8 |
| VSA\_EState9 | 分子表面上的电子状态密度相关的区域9 |
| FractionCSP3 | 分子中CSP3碳原子的比例 |
| HeavyAtomCount | 分子中的重原子数（除氢外） |
| NHOHCount | 分子中的羟基氮原子数 |
| NOCount | 分子中的氮氧原子数 |
| NumAliphaticCarbocycles | 分子中的脂肪环（由脂肪环烷基构成）数量。 |
| NumAliphaticHeterocycles | 分子中的脂肪杂环（由脂肪环烷基构成）数量。 |
| NumAliphaticRings | 分子中的脂肪环（由脂肪环烷基构成）总数。 |
| NumAromaticCarbocycles | 分子中的芳香环（由芳香环烷基构成）数量。 |
| NumAromaticHeterocycles | 分子中的芳香杂环（由芳香环烷基构成）数量。 |
| NumAromaticRings | 分子中的芳香环（由芳香环烷基构成）总数。 |
| NumHAcceptors | 分子中的氢键受体原子数 |
| NumHDonors | 分子中的氢键给体原子数 |
| NumHeteroatoms | 分子中的杂原子数 |
| NumRotatableBonds | 分子中的可转动键数 |
| NumSaturatedCarbocycles | 分子中的饱和碳环数 |
| NumSaturatedHeterocycles | 分子中的饱和杂环数 |
| NumSaturatedRings | 分子中的饱和环数 |
| RingCount | 分子中的环总数。 |
| MolLogP | 分子的分配系数的对数值；描述分子的亲水性/疏水性 |
| MolMR | 分子的相对分子质量 |
| fr\_Al\_COO | 含有α-羧基的脂肪醛的数量 |
| fr\_Al\_OH | 含有α-羟基的脂肪醛的数量 |
| fr\_Al\_OH\_noTert | 含有非三级碳的α-羟基的脂肪醛的数量 |
| fr\_ArN | 芳香胺基团的数量 |
| fr\_Ar\_COO | 含有芳香环的羧基的数量 |
| fr\_Ar\_N | 含有芳香环的氨基团的数量 |
| fr\_Ar\_NH | 含有芳香环的非脂肪醛胺的数量 |
| fr\_Ar\_OH | 含有芳香环的羟基的数量 |
| fr\_COO | 含有羧基的数量 |
| fr\_COO2 | 含有二元羧基的数量 |
| fr\_C\_O | 含有碳-氧键的数量 |
| fr\_C\_O\_noCOO | 不含有羧基的碳-氧键的数量 |
| fr\_C\_S | 含有碳-硫键的数量 |
| fr\_HOCCN | 含有HOCCN片段的数量 |
| fr\_Imine | 含有亚胺基团的数量 |
| fr\_NH0 | 含有0个氨基团的数量 |
| fr\_NH1 | 含有1个氨基团的数量 |
| fr\_NH2 | 含有2个氨基团的数量 |
| fr\_N\_O | 含有氮-氧键的数量 |
| fr\_Ndealkylation1 | 含有1个N去烷化基团的数量 |
| fr\_Ndealkylation2 | 含有2个N去烷化基团的数量 |
| fr\_Nhpyrrole | 含有N-吡咯基团的数量 |
| fr\_SH | 含有硫氢基团的数量 |
| fr\_aldehyde | 含有醛基团的数量 |
| fr\_alkyl\_carbamate | 含有烷基氨甲酸酯基团的数量 |
| fr\_alkyl\_halide | 含有烷基卤代基团的数量 |
| fr\_allylic\_oxid | 含有烯丙基氧基团的数量 |
| fr\_amide | 含有酰胺基团的数量 |
| fr\_amidine | 含有胍基团的数量 |
| fr\_aniline | 含有苯胺基团的数量 |
| fr\_aryl\_methyl | 含有芳香甲基基团的数量 |
| fr\_azide | 含有叠氮基团的数量 |
| fr\_azo | 含有偶氮基团的数量 |
| fr\_barbitur | 含有巴比妥基团的数量 |
| fr\_benzene | 含有苯基团的数量 |
| fr\_benzodiazepine | 含有苯二氮杂环基团的数量 |
| fr\_bicyclic | 含有双环基团的数量 |
| fr\_diazo | 含有重氮基团的数量 |
| fr\_dihydropyridine | 含有二氢吡啶基团的数量 |
| fr\_epoxide | 含有环氧基团的数量 |
| fr\_ester | 含有酯基团的数量 |
| fr\_ether | 含有醚基团的数量 |
| fr\_furan | 含有呋喃基团的数量 |
| fr\_guanido | 含有胍基团的数量 |
| fr\_halogen | 含有卤素基团的数量 |
| fr\_hdrzine | 含有二肼基团的数量 |
| fr\_hdrzone | 含有肼醇基团的数量 |
| fr\_imidazole | 含有咪唑基团的数量 |
| fr\_imide | 含有酰亚胺基团的数量 |
| fr\_isocyan | 含有异氰基团的数量 |
| fr\_isothiocyan | 含有异硫氰基团的数量 |
| fr\_ketone | 含有酮基团的数量 |
| fr\_ketone\_Topliss | 含有Topliss类酮基团的数量 |
| fr\_lactam | 含有内酰胺基团的数量 |
| fr\_lactone | 含有内酯基团的数量 |
| fr\_methoxy | 含有甲氧基团的数量 |
| fr\_morpholine | 含有吗啉基团的数量 |
| fr\_nitrile | 含有腈基团的数量 |
| fr\_nitro | 含有硝基团的数量 |
| fr\_nitro\_arom | 含有芳香环上的硝基团的数量 |
| fr\_nitro\_arom\_nonortho | 不邻位连接的芳香环上的硝基团的数量 |
| fr\_nitroso | 含有亚硝基团的数量 |
| fr\_oxazole | 含有噁唑基团的数量 |
| fr\_oxime | 含有肟基团的数量 |
| fr\_para\_hydroxylation | 含有对位羟基化基团的数量 |
| fr\_phenol | 含有酚基团的数量 |
| fr\_phenol\_noOrthoHbond | 不含有邻位氢键的酚基团的数量 |
| fr\_phos\_acid | 含有磷酸基团的数量 |
| fr\_phos\_ester | 含有磷酸酯基团的数量 |
| fr\_piperdine | 含有哌啶基团的数量 |
| fr\_piperzine | 含有哌嗪基团的数量 |
| fr\_priamide | 含有原酰胺基团的数量 |
| fr\_prisulfonamd | 含有对磺酰氨基基团的数量 |
| fr\_pyridine | 含有吡啶基团的数量 |
| fr\_quatN | 含有季铵基团的数量 |
| fr\_sulfide | 含有硫醚基团的数量 |
| fr\_sulfonamd | 含有磺酰氨基基团的数量 |
| fr\_sulfone | 含有砜基团的数量 |
| fr\_term\_acetylene | 含有末端乙炔基团的数量 |
| fr\_tetrazole | 含有四唑基团的数量 |
| fr\_thiazole | 含有噻唑基团的数量 |
| fr\_thiocyan | 含有硫氰基团的数量 |
| fr\_thiophene | 含有噻吩基团的数量 |
| fr\_unbrch\_alkane | 含有非支链烷基基团的数量 |
| fr\_urea | 含有尿素基团的数量 |

Table2- 1 Molecular descriptors and their meanings used in establishing QSRR and QSAR models

表2- 1建立QSRR和QSAR模型用到的分子描述符及其含义

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

图2-2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2-2（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。

|  |
| --- |
| Fig2- 2 The popular topics (A) and word cloud maps (B) of articles published within the research group from 2015 to 2023 on the impact of environmental pesticide exposure on model organisms through metabolomics methods over time  图2- 2 2015-2023年课题组内发表的通过代谢组学方法研究环境农药暴露对模式生物影响的文章的随时间变化的流行话题（A）和词云图（B） |

图2-3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上的文章已被引用79次[167]；在这些文章中，发表在Environmental Pollution上的是最多的，多达13篇[168-212]。

|  |
| --- |
| Fig2- 3 Citation count (A) and journal (B) of articles published within the research group from 2015 to 2023 on the impact of environmental pesticide exposure on model organisms through metabolomics methods  图2- 3 2015-2023年课题组内发表的通过代谢组学方法研究环境农药暴露对模式生物影响的文章的引用次数（A）及发表的期刊（B） |

### 2.3.3使用MDs建立不同QSAR模型的预测效果

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 模型编号 | 机器学习方法 | 预测对象 | 特征数 | 测试集上AUC | 训练集上AUC | 全集AUC | 测试集上的正确率 | 训练集上正确率 | 全集正确率 |
| 1 | 随机森林（RF） | Environmental Fate | 121 | 0.587 | 0.999 | 0.947 | 0.566 | 0.988 | 0.861 |
| 2 | 支持向量机（SVM） | Environmental Fate | 200 | 0.592 | 0.864 | 0.795 | 0.573 | 0.794 | 0.728 |
| 3 | 逻辑回归（LR） | Environmental Fate | 100 | 0.549 | 0.747 | 0.688 | 0.542 | 0.677 | 0.637 |
| 4 | 朴素贝叶斯（NB） | Environmental Fate | 60 | 0.551 | 0.652 | 0.623 | 0.528 | 0.598 | 0.577 |
| 5 | K近邻（KNN） | Environmental Fate | 22 | 0.544 | 0.750 | 0.685 | 0.552 | 0.681 | 0.641 |
| 6 | 梯度提升树（GB） | Environmental Fate | 118 | 0.567 | 0.991 | 0.891 | 0.559 | 0.945 | 0.829 |
| 7 | 随机森林（RF） | Ecotoxicity | 13 | 0.547 | 0.999 | 0.934 | 0.584 | 0.990 | 0.868 |
| 8 | 支持向量机（SVM） | Ecotoxicity | 19 | 0.574 | 0.779 | 0.720 | 0.605 | 0.700 | 0.672 |
| 9 | 逻辑回归（LR） | Ecotoxicity | 10 | 0.537 | 0.674 | 0.633 | 0.580 | 0.679 | 0.650 |
| 10 | 朴素贝叶斯（NB） | Ecotoxicity | 19 | 0.561 | 0.663 | 0.631 | 0.538 | 0.567 | 0.558 |
| 11 | K近邻（KNN） | Ecotoxicity | 16 | 0.536 | 0.812 | 0.731 | 0.598 | 0.751 | 0.705 |
| 12 | 梯度提升树（GB） | Ecotoxicity | 19 | 0.569 | 0.969 | 0.863 | 0.619 | 0.877 | 0.800 |
| 13 | 随机森林（RF） | Human Health | 4 | 0.461 | 0.588 | 0.550 | 0.395 | 0.556 | 0.508 |
| 14 | 支持向量机（SVM） | Human Health | 148 | 0.544 | 0.864 | 0.777 | 0.563 | 0.772 | 0.709 |
| 15 | 逻辑回归（LR） | Human Health | 138 | 0.547 | 0.795 | 0.720 | 0.535 | 0.712 | 0.659 |
| 16 | 朴素贝叶斯（NB） | Human Health | 98 | 0.540 | 0.683 | 0.641 | 0.577 | 0.574 | 0.575 |
| 17 | K近邻（KNN） | Human Health | 93 | 0.547 | 0.800 | 0.726 | 0.542 | 0.738 | 0.679 |
| 18 | 梯度提升树（GB） | Human Health | 4 | 0.460 | 0.588 | 0.550 | 0.395 | 0.556 | 0.508 |

Table2- 2 The effectiveness of different machine learning methods in establishing QSAR models for predicting environmental fate, ecotoxicity, and human health warning levels

表2- 2不同机器学习方法建立QSAR模型对环境归趋、生态毒性、人类健康警告程度预测的效果

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2-2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。

### 2.3.4不同QSRR模型对HILIC色谱柱上化学物质保留时间的预测

如表2-3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）。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 模型编号 | 特征选择方法 | 选择的特征数 | 建模方法 | 训练集上的MAE | 测试集上的MAE | 总MAE | 最大MAE | 最小MAE | 重复3次的5折交叉验证MAE |
| 1 | 相关性（F检验） | 31 | 多元线性回归 | 1.131 | 1.526 | 1.250 | 7.743 | 0.002 | 1.430 |
| 2 | 互信息 | 3 | 多元线性回归 | 1.647 | 1.649 | 1.647 | 4.730 | 0.011 | 1.751 |
| 3 | 相关性（F检验） | 2 | 支持向量回归 | 1.300 | 1.354 | 1.316 | 5.936 | 0.015 | 1.362 |
| 4 | 互信息 | 1 | 支持向量回归 | 1.402 | 1.447 | 1.416 | 6.284 | 0.009 | 1.523 |
| 5 | 相关性（F检验） | 56 | 偏最小二乘回归 | 1.318 | 1.535 | 1.383 | 5.318 | 0.013 | 1.460 |
| 6 | 互信息 | 105 | 偏最小二乘回归 | 1.442 | 1.565 | 1.479 | 5.096 | 0.012 | 1.520 |
| 7 | 相关性（F检验） | 49 | 随机森林回归 | 0.424 | 1.107 | 0.630 | 4.268 | 0.000 | 1.014 |
| 8 | 互信息 | 63 | 随机森林回归 | 0.423 | 1.063 | 0.616 | 5.186 | 0.001 | 1.009 |
| 9 | 相关性（F检验） | 114 | LASSO回归 | 1.425 | 1.622 | 1.484 | 5.689 | 0.000 | 1.559 |
| 10 | 互信息 | 104 | LASSO回归 | 1.351 | 1.593 | 1.424 | 5.947 | 0.017 | 1.500 |
| 11 | 相关性（F检验） | 79 | 岭回归 | 0.896 | 1.367 | 1.038 | 8.535 | 0.015 | 1.394 |
| 12 | 互信息 | 3 | 岭回归 | 1.647 | 1.649 | 1.648 | 4.724 | 0.011 | 1.755 |
| 13 | 相关性（F检验） | 172 | 梯度提升回归 | 0.222 | 1.037 | 0.468 | 5.740 | 0.001 | 0.956 |
| 14 | 互信息 | 132 | 梯度提升回归 | 0.198 | 1.039 | 0.451 | 5.669 | 0.001 | 0.958 |

Table2- 3 The effectiveness of QSRR prediction retention time model obtained through different feature selection and modeling methods.

表2- 3 不同特征选择和建模方式得到QSRR预测保留时间模型的效果。

GBR也有自身的缺陷，从表2-3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2-4），它仅需要的特征数量（49和63个，图2-4）不多，就可以很好的预测保留时间。

|  |
| --- |
| 图2- 4不同机器学习模型（从上到下依次是：多元线性回归、支持向量回归、偏最小二乘回归、随机森林回归、LASSO回归、岭回归和梯度上升回归）的特征选择数与平均绝对误差的关系。 |
| Fig2- 4 The relationship between the number of feature selections and the mean absolute error of different machine learning models (from top to bottom: multiple linear regression, support vector regression, partial least squares regression, random forest regression, LASSO regression, ridge regression, and gradient ascending regression) |

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2-4），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2-3的最后一列），但是需要的特征数量极少（31、3、2、1个，图2-4），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。

|  |
| --- |
| Fig2- 5 The optimal features selected by different feature selection methods and machine learning methods (from left to right: mutual information + multiple linear regression, correlation statistics + support vector regression, mutual information + support vector regression)  图2- 5不同特征选择方法和机器学习方法（从左往右：互信MI+多元线性回归MLR，相关性统计量CS+支持向量回归SVR，MI+SVR）挑选出的最优特征 |

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2-5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2-6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。

|  |
| --- |
| Fig2- 6 The optimal features selected by machine learning methods of different feature selection methods (correlation statistics CS, mutual information MI) and random forest RF.  图2- 6不同特征选择方法（相关性统计量CS、互信息MI）和随机森林RF的机器学习方法挑选出的最优特征。 |

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2-7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2-7所示。

|  |
| --- |
|  |
| 图2- 7相关性统计量CS+梯度提升回归GBR建立的模型预测的化合物的保留时间（深蓝）和实验所得的保留时间（浅蓝），按照二者MAE从小到大排列 |
| Fig2- 7 The correlation statistic CS+gradient boosting regression GBR model predicts the retention time of compounds (dark blue) and experimentally obtained retention time (light blue), arranged in ascending order of MAE for both |

|  |
| --- |
|  |
| 图2- 8互信息MI+随机森林回归RF建立的模型预测的化合物的保留时间（深绿）和实验所得的保留时间（浅绿），按照二者MAE从小到大排列 |
| Fig2- 8 The retention time of compounds predicted by the modesl established by mutual information MI+ random forest regression RF (dark green) and the retention time obtained from experiments (light green) are arranged in ascending order of their MAE |

## 2.4讨论

|  |
| --- |
|  |
| 图2- 9六种机器学习方法用分子描述符预测全集953种农药对环境归趋的警告程度的混淆矩阵（图A随机森林、B支持向量机、C逻辑回归、D朴素贝叶斯、EK近邻、F梯度提升决策树）和ROC曲线（G） |
| Fig2- 9 Six machine learning methods use molecular descriptors to predict the confusion matrix (Figure A Random Forest, B Support Vector Machine, C Logistic Regression, D Naive Bayes, EK Nearest Neighbors, F Gradient Boosting Decision Tree) and ROC curve (G) for the warning level of 953 pesticides on environmental fate in the complete set (Figure A Random Forest, B Support Vector Machine, C Logistic Regression, D Naive Bayes, EK Nearest Neighbors, F Gradient Boosting Decision Tree) |

|  |
| --- |
|  |
| 图2- 10六种机器学习方法用分子描述符预测全集953种农药对生态毒性的警告程度的混淆矩阵（图A随机森林、B支持向量机、C逻辑回归、D朴素贝叶斯、EK近邻、F梯度提升决策树）和ROC曲线（G） |
| Fig2- 10 Six machine learning methods use molecular descriptors to predict the confusion matrix (Figure A Random Forest, B Support Vector Machine, C Logistic Regression, D Naive Bayes, EK Nearest Neighbors, F Gradient Boosting Decision Tree) and ROC curve G for the warning level of ecological toxicity of 953 pesticides in the complete set (Figure A Random Forest, B Support Vector Machine, C Logistic Regression, D Naive Bayes, EK Nearest Neighbors, F Gradient Boosting Decision Tree) |

|  |
| --- |
|  |
| 图2- 11六种机器学习方法用分子描述符预测全集953种农药对人类健康的警告程度的混淆矩阵（图A随机森林、B支持向量机、C逻辑回归、D朴素贝叶斯、EK近邻、F梯度提升决策树）和ROC曲线（G） |
| Fig2- 11 Six machine learning methods use molecular descriptors to predict the confusion matrix (Figure A Random Forest, B Support Vector Machine, C Logistic Regression, D Naive Bayes, EK Nearest Neighbors, F Gradient Boosting Decision Tree) and ROC curve (G) of the warning level of 953 pesticides on human health in the complete set (Figure A Random Forest, B Support Vector Machine, C Logistic Regression, D Naive Bayes, EK Nearest Neighbors, F Gradient Boosting Decision Tree) |

虽然评估环境归趋时，结果看似不错：随机森林预测环境归趋全集的AUC是0.947（图2-9），然而模型有些过拟合：测试集上AUC仅仅只有0.587且需要121-200个分子描述符作为用于预测（表2-2）；再看预测人类健康（图2-11）：随机森林展现出了很差的准确率，全集正确个数为TP+TN=434个，错误个数为FN+FP=469个，甚至不如等于瞎猜，SVM表现稍好（AUC=0.770），正确率为TN+TP=676个，错误率FN+FP=277；不过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2-10（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。图2-13显示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[213]；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[214]。

|  |
| --- |
|  |
| 图2- 12随机森林建立QSAR模型预测不同农药对环境归趋的警告程度所用到的特征（按照ANOVA检验得分由高到低排列） |
| Fig2- 12 The features used to establish a QSAR model for predicting the warning level of different pesticides on environmental fate in random forests (ranked in descending order of ANOVA test scores) |

图2- 13随机森林建立QSAR模型预测不同农药对生态毒性的警告程度所用到的特征（按照ANOVA检验得分由高到低排列）

|  |
| --- |
|  |
|  |
| Fig2- 13 The features used to establish a QSAR model for predicting the warning level of different pesticides on ecological toxicity in random forests (arranged in descending order of ANOVA test scores) |

|  |
| --- |
| 图2- 14支持向量机建立QSAR模型预测不同农药对人类健康的警告程度所用到的特征（按照ANOVA检验得分由高到低排列）  Fig2- 14 The features used to establish a QSAR model using support vector machines to predict the warning level of different pesticides on human health (ranked in descending order of ANOVA test scores) |

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2-12和图2-14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[215, 216]。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[217]。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2-7、图2-8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2-5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[218, 219]，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2-6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。

## 2.5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。

# 第三章 农药代谢物组学数据处理工具R包meTool和软件neutralloss

在上一章中介绍了如何宏观地预测有关农药引起的环境和健康问题；除了预测之外，必然需要做实验验证。在这一章里，介绍了我们自己编写的农药代谢组学联合肠道菌群工具R包meTool和寻找质谱中兴丢失问题的软件neutralloss，以期待对日后实验数据分析起到一定的帮助。在这章中，以闫森等人所著的文章： 妊娠期烯啶虫胺暴露诱导后代雌性小鼠肠道菌群和粪便代谢物失衡[195]为例，介绍了R包meTool如何使用。

## 3.1 R包meTool

|  |
| --- |
| Fig3- 1 Overview of the functions of R package meTool: principal component analysis, partial least squares analysis, volcano plot annotation, metabolite correlation analysis, gene metabolite co analysis, gut microbiota/metabolite relative abundance heatmap, gut microbiota metabolite co analysis, and some computational functions  图3- 1 R包meTool的功能总览：主成分分析、偏最小二乘分析、火山图注释、代谢物相关性分析、基因代谢物联合分析、肠道菌群/代谢物相对丰度热图、肠道菌群-代谢物联合分析以及一些计算功能 |

图3-1显示了meTool包的主要的可视化功能，当然meTool包同时有一些计算功能，后文会详细叙述。如果要下载安装且使用，只需在R中键入devtools::install\_github("YiFanYUE99/meTool")

### 3.1.1 R包meTool的输入文件格式

R包meTool的输入文件只用四种：（A）代谢物：第一列是样品名称，最后一列是组别，中间的列是不同代谢物的丰度（B）肠道菌群：第一列是样品名称，最后一列是组别，中间的列是不同菌群的丰度，菌群名称可以是界门纲目属一一叙述，但为了便于看清，通过命令输出的只会是它的最简格式，比如：输入*k\_\_Bacteria;p\_\_Firmicutes;c\_\_Bacilli;o\_\_Lactobacillales;f\_\_Streptococcaceae;g\_\_Lactococcus*会输出为*;g\_\_Lactococcus*；*k\_\_Bacteria;p\_\_Firmicutes*输出为*p\_\_Firmicutes*。（C）基因，这个需要注意输入文件没有组别，只有第一列列名和后面的列基因名称及其丰度（D）代谢通路：通过metadiff功能筛选差异代谢物，在Metaboanyalst.ca网站上找到的受到影响的代谢通路，注意：一定要保存xlsx格式，并且Match Status那一列以数值形式保存。注意：提前标准化或归一化数据不是必须的！

|  |
| --- |
| Fig3- 2 The input file format required for the meTool package is displayed as follows: (A) metabolites (B) gut microbiota (C) genes (D) metabolic pathways  图3- 2 meTool包需要的输入文件格式展示：（A）代谢物（B）肠道菌群（C）基因（D）代谢通路 |

### 3.1.2 1. PCA图：pcaplot

主成分分析PCA是无监督的机器学习模型，不能用于分类而只是用于降维，把多个特征（比如代谢物）降维到几个主成分。

通过闫森等人[195]的文章的原始数据，展示一下如何使用这个功能：

将目录设置到安装的meTool包下：

setwd(“meTool的安装地址”)

输入代谢物丰度表格所在位置及名称，为不同组别设置颜色，画出pca图：

a<-*pcaplot*("data/metabolitesg.csv",color= c("CK"="violet","L"="lightblue2","H"="lightgreen"))

设置要输出的目录和文件名：

output<-"… /pcaplot.png"

保存图片：

ggsave(output, plot = a, width = 20, height = 20,dpi = 300,units = "cm")

得到如下图所示的pca图：

不同于其他软件所得的PCA图，图3-3清晰的显示了主成分1和2对不同组的解释程度，主成分1对整个降维后的模型的解释程度为63.97%，主成分2对整个降维后的模型的解释程度为12.92%。

|  |
| --- |
| Fig3- 3 PCA diagram: Main component 1 on the horizontal axis and main component 2 on the vertical axis  图3- 3 PCA图：横坐标为主成分1，纵坐标为主成分2 |

### 3.1.3 PLS-DA图：plsdaplot

偏最小二乘判别分析（PLS-DA）是有监督的机器学习模型，可以用来聚类。示例：同将目录设置到安装的meTool包下然后设置要展示的组（CK/L）：

G=c("CK","L")

再设置各个组的颜色：

color= c("CK"="violet","L"="lightblue2")

画图：

b<-plsdaplot("data/metabolitesg.csv",G=G,color= color)

设置输出文件的路径和名称

output<-"…/ plsdaplot.png"

保存图片：

ggsave(output, plot = b, width = 20, height = 20,dpi = 300,units = "cm")

通常实验组总有多个，在闫森等人的文章中，还有实验组H：可以再做一张图c，然后用命令d<-b+c将两张图拼为一张图且输出：

|  |
| --- |
| 图3- 4 PLS-DA图：CK-L和CK-H  Fig3- 4 PLS-DA plot: CK-L and CK-H |

由图3-4可以看出PLS—DA模型可以很好得区分出CK组和L或H组，以CK-L模型为例，主成分1对模型的解释程度为73.88%，上方的密度图显示只通过主成分1已经很好得区分了这两组数据。

### 3.1.4 火山图：volcanoplot

示例：同将目录设置到安装的meTool包下，然后确认CK组和处理组TR的行数，这里处理组是H

d<-volcanoplot("data/metabolitesg.csv", G=c("CK","L"),titlename="CK H volcano plot")

然后，一样地，输入输出路径和名称：

output<-"…/volcano\_CK\_H.png"

接着保存：

ggsave(output, plot = d, width = 25, height = 25,dpi = 300,units = "cm")

一样的，可以使用3.1.3中的方法拼合两个实验组的图片并且保存：

|  |
| --- |
| Fig3- 5 Volcano plots of differential metabolites between CK and L groups, as well as CK and H groups  图3- 5 CK和L组以及CK和H组的差异代谢物火山图 |

以CK-L火山图为例，可以看出-log(FDR)值最大的差异代谢物为谷氨酸glutamate。在这张图上，颜色越红点越大表明-log(p)值越大即p值越小，横虚线的位置是FDR=0.05，仅仅注释了-log(FDR)>0.05且log2（FC）>0.5或<-0.5的代谢物。

### 3.1.5获得显著性getSig

根据p值获得其显著性

示例：getSig(0.04)

输出：\*

### 3.1.6计算VIP值：calc\_vip

变量在投影中的重要性（VIP, Variable Importance in Projection）得分是部分最小二乘回归（PLS）中用于评估模型中每个预测变量重要性的一种度量方法。VIP得分可以帮助我们了解哪些变量对模型的预测性能贡献较大[220]。

**各变量含义**

P是预测变量的数量：

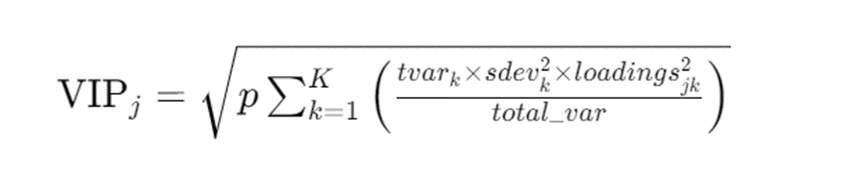
K是潜在因子的数量

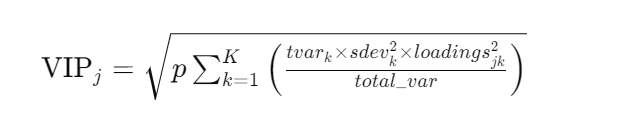
Sdev\_k：这是与每个潜在因子相关联的标准偏差，表明每个成分捕获的方差量。

Loadings\_jk： ：对载荷平方强调每个预测变量对潜在因子的贡献，同时正负效应都平等考虑。

Total\_var：这是成分得分的方差，表明每个潜在因子在预测变量中解释了方差。

**计算公式**

**VIP得分的总体计算公式如下：**



公式通过所有组成部分汇总每个预测变量的贡献。应用平方根是为了适当地缩放得分，确保它们是可解释和可比较的。公式中乘以预测变量数量 p 是确保重要性度量考虑了预测器集的大小，使VIP得分可以跨不同数量的预测变量的模型进行比较。

**评估标准**

VIP得分的缩放方式是，高于1的得分表示预测变量在模型中的重要性高于平均水平。小于1的VIP得分则表示贡献低于平均水平。

根据以上原理，编写了calc\_vip功能，结果与ropls中VipVn基本一致。

首先输入需要分析的文件路径和名称：

input<- "data/metabolitesg.csv"

然后输入要比较的组别：

comgr=c("CK","H")

计算建立主成分为2个的PLS-DA模型，计算VIP值：

vip<-calc\_vip(input,comgr)

就会获得vip值与其对应的特征名称

基于这个原理，继续开发了以下功能：

### 筛选差异代谢物：metadiff和metadiff\_all（依赖于calc\_vip）

代谢组学中筛选代谢物差异的方法主要是选择pls-da建模后VIP值大于1（且p值小于0.05的代谢物

示例：首先还是设置目录

设置要分析的文件所在位置及名称：

input<- "data/metabolitesg.csv"

设置要比较的组：

comgr=c("CK","H")

使用tables变量接受筛选出的代谢物和他们的p值、VIP值降序输出：

tables<-metadiff(input,comgr)

结果：得到一张差异代谢物的表，表中包含代谢物的VIP值、p值（并且这些化合物VIP值大一1，p值小于0.05）。

如果想要只计算VIP和p值而不删除不符合条件的代谢物：

tables2<-metadiff\_all(input,comgr)

### 3.1.8绘制差异代谢物VIP p值棒棒糖图：vipp\_plot

示例：继承上一步，计算VIP值和p值：

input<- "data/metabolitesg.csv"

选择CK组和H组建立pls-da模型：

comgr=c("CK","H")

选择CK组和L组建立pls-da模型：

comgr2=c("CK","L")

计算vip、p值并筛选代谢物：

tables<-*metadiff*(input,comgr)#筛选

只计算vip、p值：

tables2<-*metadiff\_all*(input,comgr2)#只计算值，不筛选

分别画出这两张图,并且设置柱子颜色：

vipp1<-*vipp\_plot*(tables,G=comgr, color=c("#86FF86","#E0FFE0"))

vipp2<-*vipp\_plot*(tables2,G=comgr2, ,color=c("#86FFFF","#E0FFFF")

拼合这两张图且保存：

vipp<-vipp1+vipp2

ggsave(plot = vipp,"…/… .png ",width = 20,height = 15,dpi = 300,units = "cm")

|  |
| --- |
| 图3- 6 VIP值棒棒糖图  Fig3- 6 VIP value Lollipop chart |

图3-6展示了差异代谢物的VIP值（柱子长度和颜色，越深越大）和p值（圆圈内数字与圆圈大小）左图展示的是VIP值大于1且p值小于0.05的差异代谢物（CK-H组），左图展示的是所有代谢物（CK-L组）。

### 3.1.9差异代谢物丰度热图： metabo\_heatmap（依赖于calc\_vip和metdiff）

设置要处理的文件

input<- "data/metabolitesg.csv"

设置要比较的组

comgr=c("CK","L")

画出热图

p<-*metabo\_heatmap*(input,comgr)

存储热图：

ggsave(plot = p,"…/metabolites\_CK\_L.png",width = 20,height = 15,dpi = 300,units = "cm")

自动在不同组处断开；需要注意一次只能同时画两组；图上展示的是标准化且中心化后的丰度。

|  |
| --- |
| Fig3- 7 Abundance Heatmap of Differential Metabolites  图3- 7 差异代谢物的丰度热图 |

### 3.1.10寻找差异肠道菌群：microdiff 和 microdiff\_all

结果只会显示菌群最简单的名字，而去除了其界门纲目分类。

第一个函数是计算两组间的t test p值且只保留p<0.05的代谢物：

a<-microdiff("data/microbiome.csv",comgr=c("CK","L"))

第二个函数是只计算两组间的t test p值，不删除不显著的代谢物：

b<-microdiff\_all("data/microbiome.csv",comgr=c("CK","L"))

### 3.1.11差异肠道菌群的丰度热图：micro\_heatmap

设置要分析的肠道菌群文件：

input<- "data/microbiome.csv"

需要比较的组：

comgr=c("CK","L")

画热图：

p<-micro\_heatmap(input,comgr)

保存该热图：

ggsave(plot = p,"… /microbiome\_CK\_L.png",width = 25,height = 15,dpi = 300,units = "cm")

|  |
| --- |
| 图3- 8 差异肠道菌群的丰度热图  Fig3- 8 Abundance Heatmap of Differential Microbiome |

一样的，一次只能比较两组，展示的是中心化和标准化后的差异菌群丰度。

### 3.1.12所有代谢物和菌群的热图：micmet

如果不单单只关注显著变化的代谢物和肠道菌群，而要显示所有的肠道菌群和代谢物，数据量会很大，这时候就需要依靠环形丰度热图了。

示例：输入肠道菌群丰度、代谢物丰度表名，要展示的组（有且只有两组），输出图片名称位置：

micmet(microbiome="data/microbiome.csv", metabolites="data/metabolitesg.csv",

comgr=c("CK","L"),

output="…/.png")

如图3-9所示：蓝色字体标注的菌群的相对丰度，黄色字体标注的是代谢物，只能同时展示两组，一组的名称用绿色标注，一组的名称用橘黄色标注。中间的聚类树对代谢物和菌群分别进行了聚类。该功能会对丰度进行中心化和标准化，因此和3.1.10，3.1.11一样，丰度的上限是1下限是-1。对照组和实验组中间由缝隙隔开，相对丰度越小颜色越红，越大越青。

|  |
| --- |
| 图3- 9 所有代谢物和菌群的相对丰度热图  Fig3- 9 The Abundance heatmap of all metabolites and microbiome |

### 3.1.13代谢物相关性图：cor\_met（依赖于getSign）

直接输入代谢物csv文件:

plot<-cor\_met("data/metabolitesg.csv")

存储该图片:

ggsave(plot=plot,"…/.png",width=20,height = 18,dpi=300,units="cm")

|  |
| --- |
| Fig3- 10 Metabolites Correlation map  图3- 10 代谢物相关性图 |

这张热图的不同之处在于，外侧表示Pearson相关性系数（绿色是负相关，红色是正相关），内侧是显著性p值，颜色越紫p值越小，最内侧的标签是显著性：p<0.001三星，极其显著；p<0.01二星，非常显著；p<=0.05，显著；p>0.05，无显著性。比如：甲酸盐formate和腺苷adenosine显著正相关。乳酸盐lactate和亮氨酸leucine显著负相关。

### 3.1.14肠道菌群相关性 cor\_micro（依赖于getSig）

示例：输入要处理的肠道菌群表格：

p<-cor\_micro("data/microbiome.csv")

保存图片：

ggsave(plot=p,"…/.png",width=60,height = 60,dpi=300,units="cm")

肠道菌群数据比较庞大，因此不再使用双层热图。点的大小代表r的绝对值大小，点上显示的\*代表显著性，蓝色表示负相关，橘色表示正相关，颜色越深相关性越强。比如，在图3-11中*c\_\_Clostridia*和*p\_\_Firmicutes*显著正相关。

|  |
| --- |
| 图3- 11肠道菌群相关性图  Fig3- 11Microbiome Correlation map |

### 3.1.15代谢物和肠道菌群相关性：cor\_metmicro

输入的仅有两个变量：代谢物的数据和肠道菌群的数据：

p<-*cor\_metmicro*(metabolites = "data/metabolitesg.csv", microbiome = "data/microbiome.csv")

存储图片：

ggsave(plot=p,"pic/cor\_met\_micro1.png",width=70,height = 40,dpi=300,units="cm")

|  |
| --- |
| Fig3- 12 Correlation map between Microbiome and Metabolites  图3- 12 代谢物和肠道菌群的相关性图 |

点的大小是Pearson r值的绝对值。黄色星号是显著性；x轴是肠道菌群名称，y轴是代谢物名称。为了和单独的代谢物相关性图区分，颜色进行了更改。r值在0以上越大，越红；0一下越小，越蓝；中间值为白色。比如图3-12所示：leucine和*c\_Bacilli*强负相关，lactate和*c\_Bacilli*强正相关。

### 3.1.16肠道菌群和代谢组mantel分析

肠道菌群过于复杂，以上普通的热图看起来会有些费力，因此还有mantel test热图用于在界门纲目属水平上展示菌群和代谢物的关联，虽然无法完全展示各个菌的名称，但是它更直观清晰。注意：这个功能额外以来linkET包，根据个人电脑设置，可能需要额外下载：devtools::install\_github("Hy4m/linkET", force = TRUE)

示例：首先设置肠道菌群和代谢物的文件路径及名称

microbiome<-"data/microbiome.csv"

metabolites<-"data/metabolitesg.csv"

运行此功能：

p<-mantelheatmap(metabolites, microbiome)

保存图片：

ggsave(plot = p,'…/.png',width = 10, height = 8, dpi = 300)

|  |
| --- |
| Fig3- 13 Mantel correlation map between microbiome and metabolites  图3- 13 肠道菌群和代谢物mantel检验相关性图 |

如图3-13所示，在纲（Class）水平上，肠道菌群和甲酸盐formate比较相关，在科Family水平上和谷氨酸glutamate、尿嘧啶uracil、谷氨酰胺gluatmine、萘乙酸NAA、糖核苷酸UDPG较为相关。

### 3.1.17菌群或代谢物丰度二维分布图 metmicro2dplot

示例：首先设置代谢物和肠道菌群丰度表的路径名称:

metabolites<-"data/metabolitesg.csv"

microbiome<-"data/microbiome.csv"

输入要展示的代谢物或菌群:

sub=c("glutamate","urea")

做图且保存:

p1<-*metmicro2dplot*(metabolites,microbiome,sub)

ggsave("…/.png", plot = p1, width = 20, height = 20,dpi = 300,units = "cm")

还可以做菌群和代谢物、菌群和菌群、代谢物和菌群2维图。当无法聚类时95%置信圈就不会出现，如图3-14Microbiome所示。

|  |
| --- |
| 图3- 14肠道菌群和代谢物二维丰度分布图  Fig3- 14 Two dimensional abundance distribution map of gut microbiota and metabolites |

### 3.1.18批量画代谢物丰度图：met\_pics

是否为了一张一张画代谢物图而感到烦恼？这个功能允许批量输出代谢物的丰度图

示例：调整路径，然后输入参数：

Input是文件地址和名称，G是需要显示的组别，color是各个组别对应的颜色

Filepath是这些图片需要存入的路径：met\_pics(input="data/metabolitesg.csv",G=…, color=…, filepath=…)

|  |
| --- |
| Fig3- 15 Batch output metabolite relative abundance plots  图3- 15批量输出代谢物相对丰度图 |

这张图上只是展示了一部分代谢物的相对丰度图（仅标准化，没有中心化），不过这个功能允许输出全部的代谢物，另外解释一下这里显著性的表示：具有相同字母的组之间无显著性：比如图3-15第一张图3-羟基丁酸酯（3-hydroxybutyrate）的CK组与L组有显著性差异，而与H组无显著性差异。这里的显著性由ANOVA检验计算。

### 3.1.19代谢通路图：pathway

Metadiff筛选化合物后放入metaboanlyst.ca得pathway table，注意Match status那一列要存成数字格式，必须是xlsx不能是csv

input<-"data/pathway\_CK\_H.xlsx"

p<-*pathway*(input)

保存图片：

ggsave(plot = p,'pic/pathway\_CK\_H.png',width = 7.5, height = 6, dpi = 300)

|  |
| --- |
| Fig3- 16 Affected metabolic pathway diagram  图3- 16 受到影响的代谢通路图 |

这张图上不仅显示了代谢通路受到的影响程度（横坐标与点的大小），还显示了通路变化的显著性（纵坐标）以及变化的代谢物与通路的匹配程度（点的颜色，越红匹配程度越高）。比如图3-16中，受到影响最大的是Phenylalanine, tyrosine and tryptophan biosynthesis。

### 3.1.20 基因—代谢物关联网络分析：cornet

这段代码的代谢物与基因数据集源于孙晓璇[212]等人的文章。

示例：同将目录设置到安装的meTool包下：

首先设置颜色绘图：genenode是基因节点的颜色，metabnode是代谢物节点的颜色，poscol正相关的颜色，negcol负相关的颜色

c<-cornet("data/gene.csv","data/metabolites.csv",genenode="#6495ED",metabnode="yellow",poscol="pink",negcol="green")

设置输出路径与名称：

output<-"… "

保存图片：

ggsave(output, plot = c, width = 25, height = 20,dpi = 300,units = "cm")

|  |
| --- |
| 图3- 17 代谢物-基因关联分析  Fig3- 17 Metabolites-Genes Correlation Anaylsis |

如图3-17：所有有连线的节点都是Spearman相关系数绝对值大于0.6的；深色连线表示绝对值大于0.9；在这张图上cholesterol和3-hydroxyisovalerate、3-hydroxyisovalerate和3-hydroxybutyra、alanine和inosine、inosine和glutamine、inosine和isoleucine、inosine和OAG强负相关（r<-0.9）；强正相关(r>0.9)的很多，比如malate和acetoacetate。

### 3.1.21 小结

以上就是meTool包现在的所有的功能了，有的时候图片和字的比例可能会失衡，比较好的办法就是在保存图片（ggsave）时，修改图片的宽和高（width、height）。这个包完全没有加密，因此，可以通过输入某个功能的名称（不带括号）来查看该功能具体代码。

## 3.2软件neutralloss

中性丢失是一种质谱的扫描方法：可以利用特征中性丢失和非靶向液相色谱-质谱联用鉴定烟草中的糖苷成分[221]，还可以气质联用加中性丢失的方法快速检出含氯化合物（DDT、六六六等）[222]等等。在这里，我们研发了软件neutralloss可以快速找到一定m/z的中性丢失的离子对。

软件下载地址：<https://github.com/YiFanYUE99/-A-software-to-find-neutral-loss->

下载软件后解压：

|  |
| --- |
| Fig3- 18 “Neutralloss”software’s Initial interface  图3- 18 软件neutralloss的初始界面 |

点击dist中的exe软件，等一小会至画面出现如图3-18

按提示输入各项内容

最后一行是输出框，完成后显示DONE和完成时间

注意要是csv格式输入

图3-19是示例的输入内容（比如丢失了一个葡萄糖基）；输入的两个表格是从仪器中导出的，第一列是列序号，第二列是mz，第三列是保留时间。

|  |
| --- |
| Fig3- 19 input files  图3- 19 输入软件的内容 |

注意：图上显示的只是示例；实际运行时不要保留中文和冒号；图3-20显示了查找完成时的状态；耗时17.069秒。

|  |
| --- |
|  |

随后就输出了两个表，表上显示了中兴丢失碎片对应的两个质谱峰（丢失前；丢失后，一对两行）

# 第四章 结论与展望

## 4.1结论

关于农药对环境和健康影响的研究已经有很多，在这里，我们使用QSAR模型在宏观层面上研究了对环境归趋、生态毒性、人类健康有高警告程度的农药的分子特征。同时，开发了一系列的方法从微观角度研究单个或几个农药对模式生物的影响：使用QSRR模型预测农药代谢物的保留时间；制作R包meTool以分析受农药影响的代谢物组学和肠道菌群的数据、制作软件寻找质谱的中性丢失以便在实验样本中检出农药。

### 4.1.1农药中叠氮基团的数量与农药的环境归趋和人类健康密不可分

通过QSAR模型发现，虽然无法很好地通过分子描述符精确预测某个农药的环境归趋和人类健康警告程度，农药分子中的叠氮基团数目在预测环境归趋和人类健康模型的F检验得分远远高于其它特征，是一个不可忽视的重要因素。叠氮基团会抑制细胞色素氧化酶活性，刺激呼吸，大剂量会升高血压。

### 4.1.2农药的生态毒性取决于其亲水性/亲脂性

这可能是由于生态毒性的测量是根据水生生物斑马鱼或藻类测定的。并且由随机森林建立的QSAR模型预测环境归趋和生态毒性的效果最好。而支持向量机更好地预测农药对人类健康的影响。

### 4.1.3代谢物在HILIC色谱柱上的保留时间

代谢物在HILIC色谱柱上的保留时间和代谢物的分子量、分子中氮氧原子数量、分子表面上的电子状态等因素密切相关。互信息的特征选择方式和随机森林的机器学习方法建立的QSRR模型最好得预测了保留时间，平均绝对误差仅仅为0.616分钟。

### 4.1.4 R包meTool

如前所述，meTool功能众多，需要查看详细内置数据集和源代码，访问<https://github.com/YiFanYUE99/meTool>

### 4.1.5 软件neutralloss

这款软件可以快速、方便地查找出质谱中性丢失的离子对。只需十几秒！

## 4.2展望

农药对环境和人类的健康影响一直受到热点关注，研究的方法也一直在更新，以下两点是可以在此研究上更进一步的地方。

### 4.2.1优化农药的QSAR模型

当前已经有不少利用QSAR模型预测农药对环境的影响，这篇文章里直接利用了PPDB数据库的分类。日后的研究可以根据不同农药对不同生物的LD50建立。并且使用一些更新的机器学习或深度学习模型，以期结果会更准确。

### 4.2.2对农药代谢组学工具进一步研发

尽管meTool已经涵盖了很多的功能，并且和肠道菌群数据、基因联合分析是一大亮点，无法直接搜索变化的代谢通路依旧是一个弱点，未来可以继续研究改进，做到直接在这个包中检索KEGG数据库找到变化的代谢通路。

# 参考文献

[1] LIU Y, LIU H Z, CHEN D K, et al. PlantMetSuite: A User-Friendly Web-Based Tool for Metabolomics Analysis and Visualisation [J]. Plants-Basel, 2023, 12(15): 11.

[2] PARINET J. Prediction of pesticide retention time in reversed-phase liquid chromatography using quantitative-structure retention relationship models: A comparative study of seven molecular descriptors datasets [J]. Chemosphere, 2021, 275: 10.

[3] THéVENOT E A, ROUX A, XU Y, et al. Analysis of the Human Adult Urinary Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive Workflow for Univariate and OPLS Statistical Analyses [J]. Journal of Proteome Research, 2015, 14(8): 3322-35.

[4] PANDEY V. Predictionof Environmental FateandToxicityofInsecticidesUsing Multi-Target QSAR Approach [J]. Chem Biodivers, 2024, 21(1): 10.

[5] PARINET J. Predicting reversed-phase liquid chromatographic retention times of pesticides by deep neural networks [J]. Heliyon, 2021, 7(12): 7.

[6] LEWIS K A, TZILIVAKIS J, WARNER D J, et al. An international database for pesticide risk assessments and management [J]. Hum Ecol Risk Assess, 2016, 22(4): 1050-64.

[7] BURDEN N, MAYNARD S K, WELTJE L, et al. The utility of QSARs in predicting acute fish toxicity of pesticide metabolites: A retrospective validation approach [J]. Regul Toxicol Pharmacol, 2016, 80: 241-6.

[8] QIAO L S, CAI Y L, HE Y S, et al. Trend of Multi-Scale QSAR in Drug Design [J]. Asian J Chem, 2014, 26(18): 5917-22.

[9] TSENG Y J, HOPFINGER A J, ESPOSITO E X. The great descriptor melting pot: mixing descriptors for the common good of QSAR models [J]. J Comput-Aided Mol Des, 2012, 26(1): 39-43.

[10] YAP C W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints [J]. J Comput Chem, 2011, 32(7): 1466-74.

[11] MORIWAKI H, TIAN Y S, KAWASHITA N, et al. Mordred: a molecular descriptor calculator [J]. J Cheminformatics, 2018, 10: 14.

[12] MAURI A, CONSONNI V, PAVAN M, et al. Dragon software: An easy approach to molecular descriptor calculations [J]. Match-Commun Math Cmput Chem, 2006, 56(2): 237-48.

[13] RUIZ I L, GóMEZ-NIETO M A. A Java library for the calculation of molecular descriptors; proceedings of the International Conference on Computational Methods in Science and Engineering, Corfu, GREECE, F Sep 25-30, 2007 [C]. Amer Inst Physics: MELVILLE, 2007.

[14] GUHA R. Chemical Informatics functionality in R [J]. J Stat Softw, 2007, 18(5): 16.

[15] LOVRIC M, MOLERO J M, KERN R. PySpark and RDKit: Moving towards Big Data in Cheminformatics [J]. Mol Inf, 2019, 38(6): 4.

[16] WONG E, CITRA M, KAWA M, et al. Use of EPI Suite™ fugacity model in assessing environmental fate [J]. Abstr Pap Am Chem Soc, 2019, 258: 1.

[17] BENIGNI R, DE KNECHT J. The future of the QSAR Toolbox: Moving to less uncertainty in predictive toxicology [J]. Toxicol Lett, 2014, 229: S7-S.

[18] PIZZO F, LOMBARDO A, MANGANARO A, et al. Integrated <i>in silico</i> strategy for PBT assessment and prioritization under REACH [J]. Environ Res, 2016, 151: 478-92.

[19] LIN J. Comparison of two approaches to modeling ground water exposure with EPA's PRZM-GW model [J]. Abstr Pap Am Chem Soc, 2015, 250: 1.

[20] WERNER D J, WORTHINGTON S M, SNELLING L K. 'China Pearl' peach [J]. J Amer Pomolog Soc, 2002, 56(2): 69-71.

[21] DEVILLERS J, FLATIN J. A general QSAR model for predicting the acute toxicity of pesticides to <i>Oncorhynchus mykiss</i> [J]. SAR QSAR Environ Res, 2000, 11(1): 25-43.

[22] JIA Q Z, LIU T, YAN F Y, et al. Norm Index-Based QSAR Model for Acute Toxicity of Pesticides Toward Rainbow Trout [J]. Environ Toxicol Chem, 2020, 39(2): 352-8.

[23] DEVILLERS J. A general QSAR model for predicting the acute toxicity of pesticides to <i>Lepomis macrochirus</i> [J]. SAR QSAR Environ Res, 2001, 11(5-6): 397-417.

[24] DEVILLERS J. QSAR Modeling of Pesticide Toxicity to Bees [M]. Boca Raton: Crc Press-Taylor & Francis Group, 2014.

[25] DEVILLERS J, PHAM-DELèGUE M H, DECOURTYE A, et al. Structure-toxicity modeling of pesticides to honey bees [J]. SAR QSAR Environ Res, 2002, 13(7-8): 641-8.

[26] GIRIREDDY M, SAIAKHOV R. QSAR Models for Identifying Pesticides Exhibiting High, Moderate, and Low Toxicity in Honey Bees [J]. Int J Toxicol, 2020, 39(1): 59-.

[27] GALLAGHER A, KAR S. Unveiling first report on in silico modeling of aquatic toxicity of organic chemicals to Labeo rohita (Rohu) employing QSAR and q-RASAR [J]. Chemosphere, 2024, 349: 140810.

[28] BORA A, CRISAN L, BOROTA A, et al. Ecotoxicological QSAR Modeling of Organophosphorus and Neonicotinoid Pesticides [M]//ROY K. Ecotoxicological Qsars. Totowa; Humana Press Inc. 2020: 513-44.

[29] HAMADACHE M, BENKORTBI O, HANINI S, et al. A Quantitative Structure Activity Relationship for acute oral toxicity of pesticides on rats: Validation, domain of application and prediction [J]. J Hazard Mater, 2016, 303: 28-40.

[30] SPECK-PLANCHE A. Multi-scale QSAR Approach for Simultaneous Modeling of Ecotoxic Effects of Pesticides [M]//ROY K. Ecotoxicological Qsars. Totowa; Humana Press Inc. 2020: 639-60.

[31] HAMADACHE M, AMRANE A, BENKORTBI O, et al. Environmental Toxicity of Pesticides, and Its Modeling by QSAR Approaches [M]//ROY K. Advances in Qsar Modeling: Applications in Pharmaceutical, Chemical, Food, Agricultural and Environmental Sciences. Dordrecht; Springer. 2017: 471-501.

[32] DAI P, WANG Q Q, TENG P, et al. Design, Synthesis, Antifungal Activity, and 3D-QASR of Novel Oxime Ether-Containing Coumarin Derivatives as Potential Fungicides [J]. J Agric Food Chem, 2024, 72(11): 5983-92.

[33] XU B, CUI W, TAO L, et al. Risk mitigation strategy and mechanism analysis of neonicotinoid pesticides on earthworms [J]. Environmental pollution (Barking, Essex : 1987), 2024, 347: 123719.

[34] WANG J, FENG X, YUAN W, et al. Development of terpenoid repellents against <i>Aedes albopictus</i>: a combined study of biological activity evaluation and computational modelling [J]. SAR QSAR Environ Res, 2024, 35(2): 71-89.

[35] LIU H R, CAI C L, ZHANG X J, et al. Discovery of Novel Cinnamic Acid Derivatives as Fungicide Candidates [J]. J Agric Food Chem, 2024, 72(5): 2492-500.

[36] GALVEZ-LLOMPART M, ZANNI R, VELA-CORCíA D, et al. Rational Design of a Potential New Nematicide Targeting Chitin Deacetylase [J]. J Agric Food Chem, 2024, 72(5): 2482-91.

[37] KARADUMAN G, KELLECI CELIK F. Towards safer pesticide management: A quantitative structure-activity relationship based hazard prediction model [J]. The Science of the total environment, 2024, 916: 170173.

[38] KUMAR A, OJHA P K, ROY K. First report on pesticide sub-chronic and chronic toxicities against dogs using QSAR and chemical read-across [J]. SAR QSAR Environ Res, 2024, 35(3): 241-63.

[39] VIGANò E L, BALLABIO D, RONCAGLIONI A. Artificial Intelligence and Machine Learning Methods to Evaluate Cardiotoxicity following the Adverse Outcome Pathway Frameworks [J]. Toxics, 2024, 12(1): 22.

[40] KALISZAN R. QSRR: Quantitative Structure-(Chromatographic) retention relationships [J]. Chem Rev, 2007, 107(7): 3212-46.

[41] HANCOCK T, PUT R, COOMANS D, et al. A performance comparison of modem statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies [J]. Chemometrics Intell Lab Syst, 2005, 76(2): 185-96.

[42] PERISIC-JANJIC N U, PODUNAVAC-KUZMANOVIC S O. RPTLC study of QSRR and QSAR for some benzimidazole derivatives [J]. JPC-J Planar Chromatogr-Mod TLC, 2008, 21(2): 135-41.

[43] SZATKOWSKA-WANDAS P, KOBA M, SMOLINSKI G, et al. QSRR and QSAR Studies of Antitumor Drugs in View of their Biological Activity Prediction [J]. Med Chem, 2016, 12(6): 592-600.

[44] SZATKOWSKA-WANDAS P, KOBA M, KUCHCICKA A, et al. The Application of Connected QSRR and QSAR Strategies to Predict the Physicochemical Interaction of Acridinone Derivatives with DNA [J]. Comb Chem High Throughput Screen, 2014, 17(10): 820-6.

[45] SZATKOWSKA-WANDAS P, KOBA M. Prediction of Acridinones' Ability to Interstrand DNA Crosslinks Formation Using Connected QSRR and QSAR Analysis [J]. Lett Drug Des Discov, 2016, 13(5): 387-94.

[46] BUSZEWSKI B, MICHEL M. Quantitative Structure-Retention Relationship Studies as an Analytical Tool in the Determination and Modeling of Pesticide Residues in Plant Organisms [J]. J AOAC Int, 2010, 93(6): 1703-14.

[47] OBRADOVIC D, STAVRIANIDI A, FEDOROVA E, et al. A comparative study of the predictive performance of different descriptor calculation tools: Molecular-based elution order modeling and interpretation of retention mechanism for isomeric compounds from METLIN database [J]. Journal of chromatography A, 2024, 1719: 464731.

[48] FINE J, MANN A K P, AGGARWAL P. Structure Based Machine Learning Prediction of Retention Times for LC Method Development of Pharmaceuticals [J]. Pharm Res, 2024: 10.

[49] SUN S Y, CUI B Y, KONG F Y, et al. Construction and application of a QSRR approach for identifying flavonoids [J]. J Pharm Biomed Anal, 2024, 240: 10.

[50] RAJPUT K, DHIMAN S, VENI N K, et al. Support Vector Models-Based Quantitative Structure-Retention Relationship (QSRR) in the Development and Validation of RP-HPLC Method for Multi-component Analysis of Anti-diabetic Drugs [J]. Chromatographia, 2024, 87(1): 3-16.

[51] MICIC D, OSTOJIC S, PEZO L, et al. Essential oils of coriander and sage: Investigation of chemical profile, thermal properties and QSRR analysis [J]. Ind Crop Prod, 2019, 138: 8.

[52] SUN M X, LI X H, JIANG M T, et al. A practical strategy enabling more reliable identification of ginsenosides from<i> Panax</i><i> quinquefolius</i> flower by dimension-enhanced liquid chromatography/mass spectrometry and quantitative structure-retention relationship-based retention behavior prediction [J]. J Chromatogr A, 2023, 1706: 12.

[53] ZENG W J, QIU Y Q, HUANG Y T, et al. Quantitative structure-retention relationship by databases of illegal additives [J]. J Food Compos Anal, 2023, 122: 7.

[54] CIURA K. Modeling of small molecule's affinity to phospholipids using IAM-HPLC and QSRR approach enhanced by similarity-based machine algorithms [J]. J Chromatogr A, 2024, 1714: 8.

[55] SADEGHI M, MOHAMMADINASAB E, ISFAHANI T M. QSPR models for predicting the Kovats retention indices of synthetic ester derivatives based on pyrethrin essential oil [J]. J Essent Oil Res, 2023, 35(6): 542-62.

[56] CIURA K, FEDOROWICZ J, ZUVELA P, et al. Affinity of Antifungal Isoxazolo 3,4-<i>b</i> pyridine-3(1<i>H</i>)-Ones to Phospholipids in Immobilized Artificial Membrane (IAM) Chromatography [J]. Molecules, 2020, 25(20): 9.

[57] MALJURIC N, OTASEVIC B, MALENOVIC A, et al. Quantitative structure retention relationship modeling as potential tool in chromatographic determination of stability constants and thermodynamic parameters of β-cyclodextrin complexation process [J]. J Chromatogr A, 2020, 1619: 11.

[58] WEN S-S, LI P, GAO W. Characterization and identification of alkaloids in Phellodendri Chinensis Cortex and Phellodendri Amurensis Cortex based on UHPLC-IM-Q-TOF-MS [J]. Zhongguo Zhong yao za zhi = Zhongguo zhongyao zazhi = China journal of Chinese materia medica, 2023, 48(12): 3294-307.

[59] PARINET J, MAKNI Y, DIALLO T, et al. Liquid chromatographic retention time prediction models to secure and improve the feature annotation process in high-resolution mass spectrometry [J]. Talanta, 2024, 267: 5.

[60] KUMARI P, VAN LAETHEM T, DUROUX D, et al. A multi-target QSRR approach to model retention times of small molecules in RPLC [J]. J Pharm Biomed Anal, 2023, 236: 8.

[61] ELBERT A, HAAS M, SPRINGER B, et al. Applied aspects of neonicotinoid uses in crop protection [J]. Pest Manag Sci, 2008, 64(11): 1099-105.

[62] JESCHKE P, NAUEN R, SCHINDLER M, et al. Overview of the Status and Global Strategy for Neonicotinoids [J]. J Agric Food Chem, 2011, 59(7): 2897-908.

[63] LIU B-L, GUO L-L, WANG J-W, et al. Toxicity and field control efficacy of bioinsecticides and neonicotinoid insecticides againstAmrasca biguttula(Hemiptera: Cicadellidae) in north Hunan cotton area [J]. Acta Entomologica Sinica, 2024, 66(12): 1667-74.

[64] DU Q K, SHAN Y P, HU H Y, et al. Fitness effect and transcription profile reveal sublethal effect of nitenpyram on the predator <i>Chrysopa pallens</i> (Neuroptera: Chrysopidae) [J]. Arch Insect Biochem Physiol, 2024, 115(1): 15.

[65] REN Z J, CAI T W, WAN Y, et al. Unintended consequences: Disrupting microbial communities of Nilaparvata lugens with non-target pesticides [J]. Pest Biochem Physiol, 2023, 194: 9.

[66] ZHU X H, WEI Q, WAN P J, et al. Effect of Paclobutrazol Application on Enhancing the Efficacy of Nitenpyram against the Brown Planthopper, <i>Nilaparvata lugens</i> [J]. Int J Mol Sci, 2023, 24(13): 14.

[67] WANG X L, SU H, WANG J, et al. Monitoring of insecticide resistance for Apolygus lucorum populations in the apple orchard in China [J]. Crop Prot, 2023, 170: 8.

[68] YU W F, WU R X, ZHANG L, et al. Identification of key factors affecting neonicotinoid residues in crops and risk of dietary exposure [J]. Environ Pollut, 2024, 345: 9.

[69] BENCHIKH I, ZIANI K, MATEOS A G, et al. Non-acute exposure of neonicotinoids, health risk assessment, and evidence integration: a systematic review [J]. Crit Rev Toxicol, 2024: 20.

[70] WU X, LIU Q, LI Y, et al. Urinary neonicotinoid concentrations and obesity: A cross-sectional study among Chinese adolescents [J]. Environ Pollut, 2024, 345: 11.

[71] YANG Z C, WANG Y P, TANG C X, et al. Urinary neonicotinoids and metabolites are associated with obesity risk in Chinese school children [J]. Environ Int, 2024, 183: 12.

[72] PAN D, LIN M, MU C, et al. Maternal exposure to neonicotinoid insecticides and fetal growth restriction: A nested case-control study in the guangxi Zhuang birth cohort [J]. Chemosphere, 2023, 336: 139217.

[73] WANG L L, MA C C, WEI D D, et al. Health risks of neonicotinoids chronic exposure and its association with glucose metabolism: A case-control study in rural China\* [J]. Environ Pollut, 2023, 334: 10.

[74] DOCKèS J, VAROQUAUX G, POLINE J B. Preventing dataset shift from breaking machine-learning biomarkers [J]. GigaScience, 2021, 10(9): 11.

[75] HAJJO R, SABBAH D A, BARDAWEEL S K, et al. Identification of Tumor-Specific MRI Biomarkers Using Machine Learning (ML) [J]. Diagnostics, 2021, 11(5): 27.

[76] XU K, HAN M, HUANG C, et al. Research progress of feature selection and machine learning methods for mass spectrometry-based protein biomarker discovery [J]. Sheng wu gong cheng xue bao = Chinese journal of biotechnology, 2019, 35(9): 1619-32.

[77] KLUZEK S, MATTEIZ T A. Machine-learning for osteoarthritis research [J]. Osteoarthritis Cartilage, 2019, 27(7): 977-8.

[78] LEI B, PAN J, WU F, et al. Advances in auxiliary diagnosis of neuropsychiatric diseases based on machine learning [J]. Chinese Journal of Medical Physics, 2020, 37(2): 257-64.

[79] SHELLMAN M H, SHELLMAN Y G. Human against Machine? Machine Learning Identifies MicroRNA Ratios as Biomarkers for Melanoma [J]. J Invest Dermatol, 2020, 140(1): 18-20.

[80] WU W. Predicting atopic asthma by using eNose breath profiles with machine learning [J]. J Allergy Clin Immunol, 2020, 146(5): 1010-2.

[81] CHANG C H, LIN C H, LANE H Y. Machine Learning and Novel Biomarkers for the Diagnosis of Alzheimer's Disease [J]. Int J Mol Sci, 2021, 22(5): 12.

[82] RICHARDSON A, ROBBINS C B, WISELY C E, et al. Artificial intelligence in dementia [J]. Curr Opin Ophthalmol, 2022, 33(5): 425-31.

[83] LIN H H, HIMALI J J, SATIZABAL C L, et al. Identifying Blood Biomarkers for Dementia Using Machine Learning Methods in the Framingham Heart Study [J]. Cells, 2022, 11(9): 13.

[84] CUI R, DASKALAKI E, HOSSAIN M Z, et al. A Significance Assessment of Diabetes Diagnostic Biomarkers Using Machine Learning [J]. Studies in health technology and informatics, 2021, 284: 36-8.

[85] ZHANG Z H. Machine learning method for the management of acute kidney injury: more than just treating biomarkers individually [J]. Biomark Med, 2019, 13(15): 1251-3.

[86] DING Q, LI W. How to predict acute kidney injury [J]. Chinese Journal of Practical Internal Medicine, 2023, 43(12): 982-5.

[87] FARNOUD A, OHNMACHT A J, MEINEL M, et al. Can artificial intelligence accelerate preclinical drug discovery and precision medicine? [J]. Expert Opin Drug Discov, 2022, 17(7): 661-5.

[88] RAY K. Proteomics and machine-learning models for alcohol-related liver disease biomarkers [J]. Nat Rev Gastroenterol Hepatol, 2022, 19(8): 488-.

[89] CHENG L, ZHANG X, LI C X, et al. Editorial: Identification of immune-related biomarkers for cancer diagnosis based on multi-omics data [J]. Front Oncol, 2022, 12: 3.

[90] HARRIS E. Machine Learning Algorithms Failed to Find Depression Biomarker [J]. JAMA-J Am Med Assoc, 2024: 1.

[91] HE J R, WEI J W, CHEN S Y, et al. Machine Learning-Assisted Synchronous Fluorescence Sensing Approach for Rapid and Simultaneous Quantification of Thiabendazole and Fuberidazole in Red Wine [J]. Sensors, 2022, 22(24): 13.

[92] YE W X, YAN T Y, ZHANG C, et al. Detection of Pesticide Residue Level in Grape Using Hyperspectral Imaging with Machine Learning [J]. Foods, 2022, 11(11): 16.

[93] LIU C, LI J, SUN X, et al. Detection of Pesticide Residues in Cabbage Based on Fluorescence Spectroscopy Combined with Broad Learning [J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(10): 198-204.

[94] TANG M, GUO J Y, SHEN Z. Rapid detection of carbendazim residue in tea by machine learning assisted electrochemical sensor [J]. J Food Meas Charact, 2023, 17(6): 6363-9.

[95] SUN J, HU Y, ZOU Y L, et al. Identification of pesticide residues on black tea by fluorescence hyperspectral technology combined with machine learning [J]. Food Sci Technol, 2022, 42: 9.

[96] LI M, PAN Q L, WANG J, et al. Machine learning-assisted fluorescence sensor array for qualitative and quantitative analysis of pyrethroid pesticides [J]. Food Chem, 2024, 433: 9.

[97] HU Y T, MA B X, WANG H T, et al. Non-Destructive Detection of Different Pesticide Residues on the Surface of Hami Melon Classification Based on tHBA-ELM Algorithm and SWIR Hyperspectral Imaging [J]. Foods, 2023, 12(9): 18.

[98] INDU, BAGHEL A S, BHARDWAJ A, et al. Optimization of Pesticides Spray on Crops in Agriculture using Machine Learning [J]. Comput Intell Neurosci, 2022, 2022: 10.

[99] SHEN Y K, ZHAO E R, ZHANG W, et al. Predicting pesticide dissipation half-life intervals in plants with machine learning models [J]. J Hazard Mater, 2022, 436: 8.

[100] TOMIAZZI J S, PEREIRA D R, JUDAI M A, et al. Performance of machine-learning algorithms to pattern recognition and classification of hearing impairment in Brazilian farmers exposed to pesticide and/or cigarette smoke [J]. Environ Sci Pollut Res, 2019, 26(7): 6481-91.

[101] THAO L, THIEN N D, BACH N C, et al. PesViT: a deep learning approach for detecting misuse of pesticides on farm [J]. J Supercomput, 2023, 79(14): 15790-813.

[102] DENG Y, HE W, TIAN J. Progress in researches on biological monitoring of chronic low dose pesticide exposure [J]. China Journal of Public Health, 2015, 31(7): 900-3.

[103] DALMOLIN S P, DREON D B, THIESEN F V, et al. Biomarkers of occupational exposure to pesticides: Systematic review of insecticides [J]. Environ Toxicol Pharmacol, 2020, 75: 11.

[104] MARTIN-REINA J, CASANOVA A G, DAHIRI B, et al. Adverse Health Effects in Women Farmers Indirectly Exposed to Pesticides [J]. Int J Environ Res Public Health, 2021, 18(11): 17.

[105] PRABHA MOIRANGTHEM P M, DIPANKER DEY D D, SINGH H P, et al. Surrogate biomarkers of pesticide toxicity among pesticide Handlers [J]. Indo Global Journal of Pharmaceutical Sciences, 2014, 4(3): 163.

[106] YUSA V, MILLET M, COSCOLLA C, et al. Occurrence of biomarkers of pesticide exposure in non-invasive human specimens [J]. Chemosphere, 2015, 139: 91-108.

[107] NORéN E, LINDH C, RYLANDER L, et al. Concentrations and temporal trends in pesticide biomarkers in urine of Swedish adolescents, 2000-2017 [J]. J Expo Sci Environ Epidemiol, 2020, 30(4): 756-67.

[108] LUPI D, TREMOLADA P, COLOMBO M, et al. Effects of Pesticides and Electromagnetic Fields on Honeybees: A Field Study Using Biomarkers [J]. Int J Environ Res, 2020, 14(1): 107-22.

[109] PALANISWAMY S, ABASS K, RYSä J, et al. Investigating the relationship between non-occupational pesticide exposure and metabolomic biomarkers [J]. Front Public Health, 2023, 11: 13.

[110] SANTOS A D E, PARKS C G, SENNA M M, et al. Exposure to pesticides and oxidative stress in Brazilian agricultural communities [J]. Biomarkers, 2021, 26(6): 539-47.

[111] FILIPPI I, LUCERO P, BONANSEA R I, et al. Validation of exposure indexes to pesticides through the analysis of exposure and effect biomarkers in ground pesticide applicators from Argentina [J]. Heliyon, 2021, 7(9): 10.

[112] YATOO A M, ALI M N, ZAHEEN Z, et al. Assessment of pesticide toxicity on earthworms using multiple biomarkers: a review [J]. Environ Chem Lett, 2022, 20(4): 2573-96.

[113] ERGUVEN G O, YILDIRIM N, YILDIRIM N C. Use of oxidative biomarkers in the evaluation of bioremediation efficiency [M]. Edirne, Turkey: Trakya University, 2021.

[114] LIN J-Y, XIE J, ZHU Y-P, et al. Metabolomics and its application in environmental toxicology [J]. Current Biotechnology, 2022, 12(5): 683-9.

[115] YAN S, MENG Z-Y, ZHU W-T, et al. Application of metabolomics in pesticide environmental toxicology [J]. Chinese Journal of Pesticide Science, 2019, 21(Z1): 815-22.

[116] DONG B Z. New Toxicological Endpoints of Pesticides: Perspectives on Metabolomics [J]. Curr Anal Chem, 2023, 19(7): 509-12.

[117] LIU L, WU Q C, MIAO X Y, et al. Study on toxicity effects of environmental pollutants based on metabolomics: A review [J]. Chemosphere, 2022, 286: 12.

[118] RODRíGUEZ A, CASTREJóN-GODíNEZ M L, SALAZAR-BUSTAMANTE E, et al. Omics Approaches to Pesticide Biodegradation [J]. Curr Microbiol, 2020, 77(4): 545-63.

[119] KELLOGG J, KANG S. Metabolomics, an Essential Tool in Exploring and Harnessing Microbial Chemical Ecology [J]. Phytobiomes J, 2020, 4(3): 195-210.

[120] SHAHID M, SINGH U B, KHAN M S. Metabolomics-Based Mechanistic Insights into Revealing the Adverse Effects of Pesticides on Plants: An Interactive Review [J]. Metabolites, 2023, 13(2): 25.

[121] RAY S, CHATTERJEE J, GHOSH A, et al. Assessment of shade-unshade condition and subsequently pesticide treatment on first flush tea leaf metabolites through GC/MS based metabolomics approach [J]. Cogent Food Agr, 2021, 7(1): 26.

[122] ZHANG Y, CHEN D, XU Y Z, et al. Stereoselective toxicity mechanism of neonicotinoid dinotefuran in honeybees: New perspective from a spatial metabolomics study [J]. Sci Total Environ, 2022, 809: 12.

[123] CHENG Q, LIU Q Q, LI K Y, et al. Assessing Dietary Pesticide Intake and Potential Health Effects: The Application of Global Metabolomics Analysis [J]. J Agric Food Chem, 2022, 70(13): 4086-91.

[124] YANG X, ZHANG M Z, LU T, et al. Metabolomics study and meta-analysis on the association between maternal pesticide exposome and birth outcomes [J]. Environ Res, 2020, 182: 10.

[125] SONG B B, ZHOU Y Y, ZHAN R, et al. Effects of Different Pesticides on the Brewing of Wine Investigated by GC-MS-Based Metabolomics [J]. Metabolites, 2022, 12(6): 15.

[126] NOLASCO D M, MENDES M P R, MARCIANO L P D, et al. An Exploratory Study of the Metabolite Profiling from Pesticides Exposed Workers [J]. Metabolites, 2023, 13(5): 25.

[127] DOLLINGER J, PETRIACQ P, FLANDIN A, et al. Soil metabolomics: A powerful tool for predicting and specifying pesticide sorption [J]. Chemosphere, 2023, 337: 139302.

[128] LIANG Y J, LONG D X, WANG S S, et al. Metabolomic analysis of the serum and urine of rats exposed to diazinon, dimethoate, and cypermethrin alone or in combination [J]. BMC Pharmacol Toxicol, 2024, 25(1): 10.

[129] MESNAGE R, BOWYER R C E, EL BALKHI S, et al. Impacts of dietary exposure to pesticides on faecal microbiome metabolism in adult twins [J]. Environ Health, 2022, 21(1): 14.

[130] MALLA M A, DUBEY A, KORI R K, et al. GC-MS based untargeted metabolomics reveals the metabolic response of earthworm (<i>Eudrilus eugeniae</i>) after chronic combinatorial exposure to three different pesticides [J]. Sci Rep, 2023, 13(1): 18.

[131] CUI Y-L, YU X-P, LI D-T. Current progress and prospects for future research on the rice pest metabolomes [J]. Chinese Journal of Applied Entomology, 2022, 59(5): 950-9.

[132] LI P, TIAN Y Q, DU M Y, et al. Mechanism of Rotenone Toxicity against<i> Plutella</i><i> xylostella:</i> New Perspective from a Spatial Metabolomics and Lipidomics Study [J]. J Agric Food Chem, 2022: 12.

[133] MA L L, YIN Z B, XIE Q R, et al. Metabolomics and mass spectrometry imaging reveal the chronic toxicity of indoxacarb to adult zebrafish<i> (Danio</i><i> rerio)</i> livers [J]. J Hazard Mater, 2023, 453: 12.

[134] HOU Y L, DING T T, GUAN Z Y, et al. Untargeted metabolomics reveals the preventive effect of quercetin on nephrotoxicity induced by four organophosphorus pesticide mixtures [J]. Food Chem Toxicol, 2023, 175: 10.

[135] PAN L X, ZHOU C G, JING J, et al. Metabolomics analysis of cucumber fruit in response to foliar fertilizer and pesticides using UHPLC-Q-Orbitrap-HRMS [J]. Food Chem, 2022, 369: 9.

[136] TRYGG J, WOLD S. Orthogonal projections to latent structures (O-PLS) [J]. Journal of Chemometrics, 2002, 16(3): 119-28.

[137] SZYMAŃSKA E, SACCENTI E, SMILDE A K, et al. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies [J]. Metabolomics, 2012, 8(1): 3-16.

[138] POLSKY B R. SIMCA-P, version 7.0 [J]. Journal of Chemometrics, 1999, 13(5): 539-40.

[139] HOWELL A, YAROS C. Downloading and Analysis of Metabolomic and Lipidomic Data from Metabolomics Workbench Using MetaboAnalyst 5.0 [J]. Methods in molecular biology (Clifton, NJ), 2023, 2625: 313-21.

[140] XIA J G, MANDAL R, SINELNIKOV I V, et al. MetaboAnalyst 2.0-a comprehensive server for metabolomic data analysis [J]. Nucleic Acids Res, 2012, 40(W1): W127-W33.

[141] CHONG J, SOUFAN O, LI C, et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis [J]. Nucleic Acids Res, 2018, 46(W1): W486-W94.

[142] XIA J, WISHART D S. Using MetaboAnalyst 3.0 for Comprehensive Metabolomics Data Analysis [J]. Current protocols in bioinformatics, 2016, 55: 14.0.1-.0.91.

[143] DEKERMANJIAN J, LABEIKOVSKY W, GHOSH D, et al. MSCAT: A Machine Learning Assisted Catalog of Metabolomics Software Tools [J]. Metabolites, 2021, 11(10): 11.

[144] BITTREMIEUX W, LEVITSKY L, PILZ M, et al. Unified and Standardized Mass Spectrometry Data Processing in Python Using spectrum\_utils [J]. Journal of Proteome Research, 2023: 7.

[145] POWELL C D, MOSELEY H N B. The mwtab Python Library for RESTful Access and Enhanced Quality Control, Deposition, and Curation of the Metabolomics Workbench Data Repository [J]. Metabolites, 2021, 11(3): 16.

[146] SMELTER A, MOSELEY H N B. A Python library for FAIRer access and deposition to the Metabolomics Workbench Data Repository [J]. Metabolomics, 2018, 14(5): 8.

[147] POWELL C D, MOSELEY H N B. The metabolomics workbench file status website: a metadata repository promoting FAIR principles of metabolomics data [J]. BMC Bioinformatics, 2023, 24(1): 10.

[148] THOMPSON P T, MOSELEY H N B. MESSES: Software for Transforming Messy Research Datasets into Clean Submissions to Metabolomics Workbench for Public Sharing [J]. Metabolites, 2023, 13(7): 25.

[149] RIQUELME G, ZABALEGUI N, MARCHI P, et al. A Python-Based Pipeline for Preprocessing LC-MS Data for Untargeted Metabolomics Workflows [J]. Metabolites, 2020, 10(10): 14.

[150] ANWAR A M, AHMED E A, SOUDY M, et al. Xconnector: Retrieving and visualizing metabolites and pathways information from various database resources [J]. J Proteomics, 2021, 245: 6.

[151] WEIDNER L, HEMMLER D, RYCHLIK M, et al. DBDIpy: a Python library for processing of untargeted datasets from real-time plasma ionization mass spectrometry [J]. Bioinformatics, 2023, 39(2): 2.

[152] KONTOU E E, WALTER A, ALKA O, et al. UmetaFlow: an untargeted metabolomics workflow for high-throughput data processing and analysis [J]. J Cheminformatics, 2023, 15(1): 12.

[153] SHAVE S, DAWSON J C, ATHAR A M, et al. Phenonaut: multiomics data integration for phenotypic space exploration [J]. Bioinformatics, 2023, 39(4): 2.

[154] MOCK A, WARTA R, DETTLING S, et al. MetaboDiff: an R package for differential metabolomic analysis [J]. Bioinformatics, 2018, 34(19): 3417-8.

[155] SALZER L, WITTING M, SCHMITT-KOPPLIN P. MobilityTransformR: an R package for effective mobility transformation of CE-MS data [J]. Bioinformatics, 2022, 38(16): 4044-5.

[156] CHETNIK K, BENEDETTI E, GOMARI D P, et al. maplet: an extensible R toolbox for modular and reproducible metabolomics pipelines [J]. Bioinformatics, 2022, 38(4): 1168-70.

[157] CHONG J, XIA J G. MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data [J]. Bioinformatics, 2018, 34(24): 4313-4.

[158] FANG X Y, LIU Y, REN Z J, et al. Lilikoi V2.0: a deep learning-enabled, personalized pathway-based R package for diagnosis and prognosis predictions using metabolomics data [J]. GigaScience, 2021, 10(1): 11.

[159] AL-AKWAA F M, YUNITS B, HUANG S J, et al. Lilikoi: an R package for personalized pathway-based classification modeling using metabolomics data [J]. GigaScience, 2018, 7(12): 9.

[160] CHILIMONIUK J, GRZESIAK K, KALA J, et al. imputomics: web server and R package for missing values imputation in metabolomics data [J]. Bioinformatics, 2024, 40(3): 4.

[161] RAINER J, VICINI A, SALZER L, et al. A Modular and Expandable Ecosystem for Metabolomics Data Annotation in R [J]. Metabolites, 2022, 12(2): 13.

[162] KOTLI M, PIIR G, MARAN U. Pesticide effect on earthworm lethality via interpretable machine learning [J]. J Hazard Mater, 2024, 461: 11.

[163] ZHANG J, LIU T, WU G, et al. DFT study on substituted aromatic compounds' caused acute toxicities to on Daphnia Magna Straus [J]. Computers and Applied Chemistry, 2010, 27(6): 839-43.

[164] ZIQIANG T, CHANGJUN F. Acute Toxicity of Substituted Phenols to Daphnia Magna Strausat Different pH Values by Kier's Shape IndeX [J]. Journal of Wuhan University Natural Science Edition, 2006, 52(6): 685-9.

[165] YANG Y, ZHANG S, ZHAO D. DFT study on the structure of aromatic derivatives and its toxicity to fathead minnows [J]. Journal of Shaanxi Normal University Natural Science Edition, 2016, 44(6): 43-7.

[166] DOBRICIC V, TURKOVIC N, IVKOVIC B, et al. Evaluation of the lipophilicity of chalcones by RP-TLC and computational methods [J]. JPC-J Planar Chromatogr-Mod TLC, 2020, 33(3): 245-53.

[167] TENG M M, QI S Z, ZHU W T, et al. Sex-specific effects of difenoconazole on the growth hormone endocrine axis in adult zebrafish (<i>Danio rerio</i>) [J]. Ecotox Environ Safe, 2017, 144: 402-8.

[168] WANG X R, QIU J, XU P, et al. Rapid Metabolite Discovery, Identification, and Accurate Comparison of the Stereoselective Metabolism of Metalaxyl in Rat Hepatic Microsomes [J]. J Agric Food Chem, 2015, 63(3): 754-60.

[169] WANG Y, QIU J, ZHU W T, et al. Enantioselective Metabolism and Interference on Tryptophan Metabolism of Myclobutanil in Rat Hepatocytes [J]. Chirality, 2015, 27(9): 643-9.

[170] WANG Y, ZHU W T, QIU J, et al. Monitoring tryptophan metabolism after exposure to hexaconazole and the enantioselective metabolism of hexaconazole in rat hepatocytes <i>in vitro</i> [J]. J Hazard Mater, 2015, 295: 9-16.

[171] WANG Y, XU L, LI D Z, et al. Enantioselective bioaccumulation of hexaconazole and its toxic effects in adult zebrafish (<i>Danio rerio</i>) [J]. Chemosphere, 2015, 138: 798-805.

[172] WANG D Z, QIU J, ZHU W T, et al. Evaluating the enantioselective distribution, degradation and excretion of epoxiconazole in mice following a single oral gavage [J]. Xenobiotica, 2015, 45(11): 1009-15.

[173] YAN J, ZHANG P, WANG X R, et al. Stereoselective Degradation of alpha-Cypermethrin and Its Enantiomers in Rat Liver Microsomes [J]. Chirality, 2016, 28(1): 58-64.

[174] WANG X R, WANG D Z, WANG Y, et al. A combined non-targeted and targeted metabolomics approach to study the stereoselective metabolism of benalaxyl enantiomers in mouse hepatic microsomes [J]. Environ Pollut, 2016, 212: 358-65.

[175] WANG X R, ZHU W T, QIU J, et al. Enantioselective metabolism and toxic effects of metalaxyl on primary hepatocytes from rat [J]. Environ Sci Pollut Res, 2016, 23(18): 18649-56.

[176] ZHANG P, ZHU W T, WANG D Z, et al. Enantioselective Effects of Metalaxyl Enantiomers on Breast Cancer Cells Metabolic Profiling Using HPLC-QTOF-Based Metabolomics [J]. Int J Mol Sci, 2017, 18(1): 16.

[177] WANG Y, TENG M M, WANG D Z, et al. Enantioselective bioaccumulation following exposure of adult zebrafish (<i>Danio rerio</i>) to epoxiconazole and its effects on metabolomic profile as well as genes expression [J]. Environ Pollut, 2017, 229: 264-71.

[178] WANG D Z, WANG X R, ZHANG P, et al. The fate of technical-grade chlordane in mice fed a high-fat diet and its roles as a candidate obesogen [J]. Environ Pollut, 2017, 222: 532-42.

[179] WANG D Z, ZHU W T, WANG Y, et al. Metabolomics Approach to Investigate Estrogen Receptor-Dependent and Independent Effects of o,p′-DDT in the Uterus and Brain of Immature Mice [J]. J Agric Food Chem, 2017, 65(18): 3609-16.

[180] MIAO J Y, WANG D Z, YAN J, et al. Comparison of subacute effects of two types of pyrethroid insecticides using metabolomics methods [J]. Pest Biochem Physiol, 2017, 143: 161-7.

[181] WANG X R, WANG D Z, ZHOU Z Q, et al. Subacute oral toxicity assessment of benalaxyl in mice based on metabolomics methods [J]. Chemosphere, 2018, 191: 373-80.

[182] TENG M M, ZHU W T, WANG D Z, et al. Metabolomics and transcriptomics reveal the toxicity of difenoconazole to the early life stages of zebrafish (<i>Danio rerio</i>) [J]. Aquat Toxicol, 2018, 194: 112-20.

[183] TENG M M, QI S Z, ZHU W T, et al. Effects of the bioconcentration and parental transfer of environmentally relevant concentrations of difenoconazole on endocrine disruption in zebrafish (<i>Danio rerio</i>) [J]. Environ Pollut, 2018, 233: 208-17.

[184] YAN J, WANG D Z, MIAO J Y, et al. Discrepant effects of α-endosulfan, β-endosulfan, and endosulfan sulfate on oxidative stress and energy metabolism in the livers and kidneys of mice [J]. Chemosphere, 2018, 205: 223-33.

[185] TENG M M, ZHU W T, WANG D Z, et al. Acute exposure of zebrafish embryo (<i>Danio rerio</i>) to flutolanil reveals its developmental mechanism of toxicity via disrupting the thyroid system and metabolism [J]. Environ Pollut, 2018, 242: 1157-65.

[186] JIA M, WANG Y, WANG D Z, et al. The effects of hexaconazole and epoxiconazole enantiomers on metabolic profile following exposure to zebrafish (<i>Danio rerio</i>) as well as the histopathological changes [J]. Chemosphere, 2019, 226: 520-33.

[187] MENG Z Y, LIU L, JIA M, et al. Impacts of Penconazole and Its Enantiomers Exposure on Gut Microbiota and Metabolic Profiles in Mice [J]. J Agric Food Chem, 2019, 67(30): 8303-11.

[188] ZHANG P, WANG S, HE Y H, et al. Identifying Metabolic Perturbations and Toxic Effects of <i>Rac</i>-Metalaxyl and Metalaxyl-M in Mice Using Integrative NMR and UPLC-MS/MS Based Metabolomics [J]. Int J Mol Sci, 2019, 20(21): 17.

[189] TIAN S, TENG M M, MENG Z Y, et al. Toxicity effects in zebrafish embryos (<i>Danio rerio</i>) induced by prothioconazole [J]. Environ Pollut, 2019, 255: 8.

[190] MENG Z Y, LIU L, XI Y X, et al. Different effects of exposure to penconazole and its enantiomers on hepatic glycolipid metabolism of male mice [J]. Environ Pollut, 2020, 257: 9.

[191] YAN S, MENG Z Y, TIAN S N, et al. Neonicotinoid insecticides exposure cause amino acid metabolism disorders, lipid accumulation and oxidative stress in ICR mice [J]. Chemosphere, 2020, 246: 9.

[192] ZHANG R K, ZHOU Z Q, ZHU W T. Evaluating the effects of the tebuconazole on the earthworm, <i>Eisenia fetida</i> by H-1 NMR-Based untargeted metabolomics and mRNA assay [J]. Ecotox Environ Safe, 2020, 194: 9.

[193] YAO C Y, SHENG J, YAN S, et al. Enantioselectivity effects of imazethapyr enantiomers to metabolic responses in mice [J]. Pest Biochem Physiol, 2020, 168: 8.

[194] LI R S, MENG Z Y, SUN W, et al. Bioaccumulation and toxic effects of penconazole in earthworms (Eisenia fetida) following soil exposure [J]. Environ Sci Pollut Res, 2020, 27(30): 38056-63.

[195] YAN S, TIAN S N, MENG Z Y, et al. Imbalance of gut microbiota and fecal metabolites in offspring female mice induced by nitenpyram exposure during pregnancy [J]. Chemosphere, 2020, 260: 10.

[196] JIA M, TENG M M, TIAN S N, et al. Developmental toxicity and neurotoxicity of penconazole enantiomers exposure on zebrafish (<i>Danio rerio</i>) [J]. Environ Pollut, 2020, 267: 8.

[197] SUN W, MENG Z Y, LI R S, et al. Joint effects of microplastic and dufulin on bioaccumulation, oxidative stress and metabolic profile of the earthworm (<i>Eisenia fetida</i>) [J]. Chemosphere, 2021, 263: 8.

[198] YAN J, WANG D Z, MENG Z Y, et al. Effects of incremental endosulfan sulfate exposure and high fat diet on lipid metabolism, glucose homeostasis and gut microbiota in mice [J]. Environ Pollut, 2021, 268: 12.

[199] MENG Z Y, TIAN S N, SUN W, et al. Effects of exposure to prothioconazole and its metabolite prothioconazole-desthio on oxidative stress and metabolic profiles of liver and kidney tissues in male mice [J]. Environ Pollut, 2021, 269: 11.

[200] YAN S, TIAN S N, MENG Z Y, et al. Exposure to nitenpyram during pregnancy causes colonic mucosal damage and non-alcoholic steatohepatitis in mouse offspring: The role of gut microbiota [J]. Environ Pollut, 2021, 271: 10.

[201] JIA M, TENG M M, TIAN S U, et al. Effects of penconazole enantiomers exposure on hormonal disruption in zebrafish Danio rerio (Hamilton, 1822) [J]. Environ Sci Pollut Res, 2021, 28(32): 43476-82.

[202] MENG Z Y, HUANG S R, SUN W, et al. A Typical Fungicide and Its Main Metabolite Promote Liver Damage in Mice through Impacting Gut Microbiota and Intestinal Barrier Function [J]. J Agric Food Chem, 2021, 69(45): 13436-47.

[203] MENG Z Y, SUN W, LIU W, et al. A common fungicide tebuconazole promotes colitis in mice via regulating gut microbiota [J]. Environ Pollut, 2022, 292: 13.

[204] TIAN S N, YAN H, MENG Z Y, et al. Prothioconazole and prothioconazole-desthio induced different hepatotoxicities via interfering with glycolipid metabolism in mice [J]. Pest Biochem Physiol, 2022, 180: 9.

[205] YAN S, TIAN S N, MENG Z Y, et al. Synergistic effect of ZnO NPs and imidacloprid on liver injury in male ICR mice: Increase the bioavailability of IMI by targeting the gut microbiota\* [J]. Environ Pollut, 2022, 294: 10.

[206] SUN W, YAN S, MENG Z Y, et al. Combined ingestion of polystyrene microplastics and epoxiconazole increases health risk to mice: Based on their synergistic bioaccumulation in vivo [J]. Environ Int, 2022, 166: 12.

[207] LI L, LIANG H W, ZHAO T T, et al. Differential effects of thiamethoxam and clothianidin exposure on their tissue distribution and chronic toxicity in mice [J]. Chem-Biol Interact, 2022, 366: 9.

[208] TIAN S N, YAN S, MENG Z Y, et al. Widening the Lens on Prothioconazole and Its Metabolite Prothioconazole-Desthio: Aryl Hydrocarbon Receptor-Mediated Reproductive Disorders through <i>in Vivo</i>, <i>in Vitro</i>, and <i>in Silico</i> Studies [J]. Environ Sci Technol, 2022, 56(24): 17890-901.

[209] MENG Z Y, YAN Z X, SUN W, et al. Azoxystrobin Disrupts Colonic Barrier Function in Mice<i> via</i> Metabolic Disorders Mediated by Gut Microbiota [J]. J Agric Food Chem, 2023, 71(1): 789-801.

[210] TIAN S N, SUN W, SUN X X, et al. Intergenerational reproductive toxicity of parental exposure to prothioconazole and its metabolite on offspring and epigenetic regulation associated with DNA methylation in zebrafish [J]. Environ Int, 2023, 173: 10.

[211] MENG Z Y, YAN S, SUN W, et al. Chlorothalonil induces obesity in mice by regulating host gut microbiota and bile acids metabolism via FXR pathways [J]. J Hazard Mater, 2023, 452: 16.

[212] SUN X X, TIAN S N, YAN S, et al. <i>Bifidobacterium</i> mediate gut microbiota-remedied intestinal barrier damage caused by cyproconazole in zebrafish<i> (Danio</i><i> rerio)</i> [J]. Sci Total Environ, 2024, 912: 14.

[213] KUMARI B, KUMAR V, SINHA A K, et al. Toxicology of arsenic in fish and aquatic systems [J]. Environ Chem Lett, 2017, 15(1): 43-64.

[214] IPEK Z Z, MINAZ M, KAYIS S. Determination of the use of <i>Ligula intestinalis</i> as a bioindicator in malathion residues [J]. Environ Sci Pollut Res, 2023, 30(33): 80732-40.

[215] PRAEKUNATHAM H. Mechanisms of Cyanide and Azide Binding to Cobalt Complexes Relevant to Their Antidotal Action [M]. 2018.

[216] SUZUKI Y, TAGUCHI K, HANYU S, et al. Oxidized liposomal artificial red blood cells rescue azide-poisoned mice from lethal toxidrome by recovering cytochrome c oxidase activity [J]. J Drug Deliv Sci Technol, 2022, 71: 9.

[217] PRAEKUNATHAM H, GARRETT K K, BAE Y, et al. A Cobalt Schiff-Base Complex as a Putative Therapeutic for Azide Poisoning [J]. Chem Res Toxicol, 2020, 33(2): 333-42.

[218] YANG Q, JI H C, FAN X Q, et al. Retention time prediction in hydrophilic interaction liquid chromatography with graph neural network and transfer learning [J]. J Chromatogr A, 2021, 1656: 9.

[219] TORIGOE T, TAKAHASHI M, HERAVIZADEH O, et al. Predicting Retention Time in Unified-Hydrophilic-Interaction/Anion-Exchange Liquid Chromatography High-Resolution Tandem Mass Spectrometry (Unified-HILIC/AEX/HRMS/MS) for Comprehensive Structural Annotation of Polar Metabolome [J]. Anal Chem, 2024, 96(3): 1275-83.

[220] GALINDO-PRIETO B, ERIKSSON L, TRYGG J. Variable influence on projection (VIP) for orthogonal projections to latent structures (OPLS) [J]. Journal of Chemometrics, 2014, 28(8): 623-32.

[221] 吴琼, 汪阳忠, 安彤, et al. 利用特征中性丢失和非靶向液相色谱-质谱联用技术鉴定烟草中的糖苷成分 [J]. 分析测试学报, 2023, 42(01): 68-74.

[222] 李燕莹, 易云婷, 陈贤驰, et al. 气质联用中性丢失扫描定性含氯化合物的方法 [J]. 食品工业, 2021, 42(05): 452-4.

# 致 谢

乐衣凡

2024年4月

# 作者简介

乐衣凡，女，1999年出生于上海市。2021年毕业于中国农业大学理学院应用化学系，获理学学士学位。2021年继续在中国农业大学攻读学术硕士学位，专业农药学。研究方向涉及农药代谢组学数据处理，师从朱文涛副教授。硕士期间以一作发表SCI论文一篇，参与发表SCI论文3篇。成果如下：

1. **Yue, YF**, Sun, XX, Tian, SN, Yan, S, Sun, W, Miao, JY, Huang, SR, Diao, JL, Zhou, ZQ, Zhu, WT. (2024). Multi-omics and gut microbiome: Unveiling the pathogenic mechanisms of early-life pesticide exposure. PESTICIDE BIOCHEMISTRY AND PHYSIOLOGY 2024, 199, 105770

2. Sun, XX, Tian, SN, Yan, S, Sun, W, Miao, JY, **Yue, YF,** Han, SH, Huang, SR, Xu, N, Diao, JL, Zhou, ZQ, Zhu, WT. (2024). *Bifidobacterium* mediate gut microbiota-remedied intestinal barrier damage caused by cyproconazole in zebrafish (*Danio rerio*). SCIENCE OF THE TOTAL ENVIRONMENT 2024, 912, 169556

3. Tian, SN, Sun, W, Sun, XX, **Yue, YF,** Jia, M, Huang, SR, Zhou, ZQ, Li, L, Diao, JL, Yan, S, Zhu, WT. (2023). Intergenerational reproductive toxicity of parental exposure to prothioconazole and its metabolite on offspring and epigenetic regulation associated with DNA methylation in zebrafish. ENVIRONMENT INTERNATIONAL 2023, 173, 107830