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1。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3。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4。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-Descriptor5，Mordred6，Dragon software7。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CDK8，基于R的rcdk包9，和基于python的rdkit10。而农药环境风险评估中常用的计算毒理学软件有EPI Suite31、QSAR Toolbox 32、PBT Profiler 33、PRZM-GW 34、China-PEARL 35、

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）2；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23；在2001年，预测了农药对蓝鳃太阳鱼Lepomis macrochirus的急性毒性27；其他一些研究使用QSAR模型预测不同杀虫剂对蜜蜂的毒性 20、28、30；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16；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22；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24；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25-26；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 19。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13。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14，辅助开发白纹伊蚊萜类驱虫剂15，指导开发新型肉桂酸衍生物作为杀菌剂17，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18。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57。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58。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59。

QSAR模型的建立如图3所示19，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11，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）12。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48、抗肿瘤药物的生物活性 47、预测吖啶酮衍生物与DNA的理化相互作用 49-50。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51，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 52。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36。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37。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 38。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 41；对芫荽和鼠尾草精油化合物的保留时间（RT）进行预测56。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 44。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45。

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荷 39；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 40。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44；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 54；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55。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46。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42-43。

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：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52。其他研究也有描述使用QSRR模型预测农药的保留时间（图2），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53。

# 使用机器学习的方法研究烯啶虫胺对小鼠毒性的生物标志物

## 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一60。新烟碱类杀虫剂作用于昆虫烟碱乙酰胆碱受体nAChRs61。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65。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66。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70；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71。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72。

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64。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63。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62，而尿液中烯啶虫胺含量高的儿童BMI z 评分较低67。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68。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69。

## 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85。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88。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82。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83。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89。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76。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81。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86。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80。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73、87。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77。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78。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79。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83。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96；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102；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97；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99；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98；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101。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93，预测植物种农药的半衰期94，评估杀虫剂暴露对于农民的耳部毒性95；监测农场农药滥用100。

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110。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113。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114。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103。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105。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104。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112。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106。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107。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108。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109。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111。

# 使用代谢组学研究农药环境行为

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116。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115。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124。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120。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126。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127。

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117。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118。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136。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134。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119。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123。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125。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130。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131。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132。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133。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135。

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128。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129。

# 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 137、138、139。

常用的代谢组学数据处理工具比较早期的有R包Ropls140、SIMCA-P141，还有一直从1.0发展到5.0的MetaboAnalyst 144、145、146、147。使用Ropls对代谢组学可视化的结果见图5，虽然很好得表达出了结果，但在美观性上有待提高140。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4）142。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148。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149、150、151。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152。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153。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154。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155。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157。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156。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。

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158。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-MS159。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160。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161。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162、164。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163。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165。

# 1.5主要研究目的、主要内容及技术路线