STIWS

(Vesion 1.0)

用户手册

前　　言

感谢您选择使用 STIWS（Searching Targets and Identifying With mass Spectrometry）工具。作为一款专为中药成分鉴定设计的辅助工具，STIWS 致力于通过创新算法和智能化工作流程，为中药物质基础解析提供多方位的技术解决方案。本软件基于 Python 语言开发，集成了质谱数据处理、中药成分数据库检索和智能算法分析三大核心模块，主要提供以下四个功能：1.中药成分数据库查询；2.母离子精准匹配；3.未知化合物识别；4.已知化合物鉴定等四个功能。本手册将依次对软件的四个功能进行介绍，以便于用户使用。如在使用过程中存在相关问题，欢迎用户通过 https://github.com/67520/STIWS 与我们联系。在此特别感谢 TCMSP、ETCM、HERB、TM-MC 等中药成分数据库以及 GNPS、MoNA 等质谱数据库的宝贵贡献（排名不分先后）。

1. 中药成分数据库查询

STIWS 可基于药材的中文名称以及拼音大写查找相关的化合物信息（来源药材、化合物名、SMILES 及化合物类型）（图 1）。在查找时如未勾选“Accuracy Search”方框则视为模糊查找，用户将得到所有名称中含有查找目标的药材。若勾选则将进行精确查找，STIWS 将仅输出与查找目标名称完全相同的药材。当用户需要同时查询多个药材时，不同药材的名称可采用“；”或“，”等字符分隔。点击“Export To Excel”按钮即可将查询结果输出到指定文件夹。

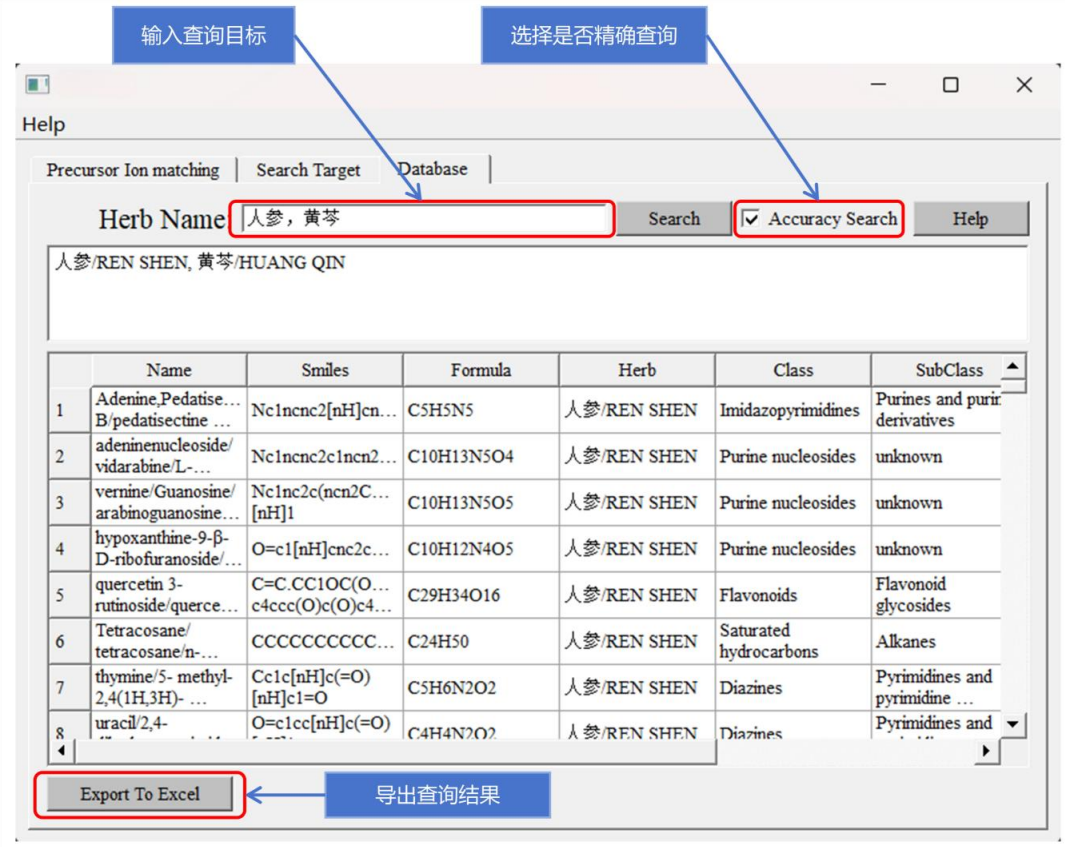

图 1 中药成分数据库查询界面

2. 母离子精准匹配

使用该功能时首先需导入 ms1 格式的一级质谱数据或 mgf 格式的二级质谱数据。STIWS 也支持同时导入一级和二级质谱数据，在导入一级质谱数据的情况下 STIWS 将根据一级质谱数据提取各个母离子的 EIC 图并对各母离子的所属峰进行注释。STIWS 的“Match ppm”既是提取 EIC 图时的误差也是后续母离子匹配的误差，而“Min Peak Intensity”则是 EIC 图中各个峰的最小峰高，所有峰强低于设定值的峰均划分为 0 号峰。进一步准备待匹配药材信息后，用户还可选择导入额外的化合物数据（Import Mdb），导入数据的格式可见“Help”。最后，选择需要考虑的加合状态后即可进行母离子匹配，STIWS 支持自定义导入加合状态，自定义输入格式可见“Help”。最后“All Herb”和“Mix Mode”分别对应全库匹配和混合匹配两种模式，在两者均不选择的情况下则将进行针对目标药材的匹配。三种匹配模式的区别如下：

① 目标药材匹配：基于该方法进行母离子匹配时将仅考虑大川芎方所含药材的化合物信息（川芎、天麻）以尽可能地减少噪音，得到少而精确的结果。

② 全库匹配：基于该方法进行母离子匹配时将考虑整个中药成分数据库以得到更加全面的匹配结果，但该方法也会带来更多的噪音。

③ 混合匹配：该方法结合①和②两种思路，其首先基于目标药材匹配进行分析，然后提取本次匹配中未匹配到候选化合物的母离子，基于全库匹配进一步分析。该方法结合①和②两种思路的优点，其可在保证噪音相对较小的情况下得到较为全面的结果。

最后，母离子匹配的结果将输出为 Excel，输出的表格中每行代表一张谱图与一个候选化合物的匹配结果。各列从左到右依次为该谱图的 PeaksIdx（EIC 图中的峰序号）、PeaksIntensity（EIC 图中的峰强度）、RTINSECONDS（保留时间）、PEPMASS（母离子 *m/z*）以及谱图与化合物匹配时的 ppm（匹配误差的绝对值）、*m/z*（仅在分析数据包含 mgf 格式的二级质谱数据时导出，该部分包括二级谱图各碎片离子的 *m/z* 及丰度）、Adduct（匹配化合物的加合状态）、SMILES（可复制到 ChemDraw 查看匹配化合物结构）、Formula（匹配化合物的分子式）、Theoretical PEPMASS（匹配化合物理论上的母离子 *m/z*）、Name（化合物名称）、Herb（化合物所属药材）、Class（化合物的主要类别）、SubClass

（化合物的次要类别）、DirectParent（化合物在分类上的直接父级）。需要注意的是，结果中的 Herb 在针对目标药材的匹配模式下将仅显示用户输入的药材，而在全库匹配模式的结果以及混合匹配模式的部分结果中 Herb 将显示该化合物所有的来源药材。

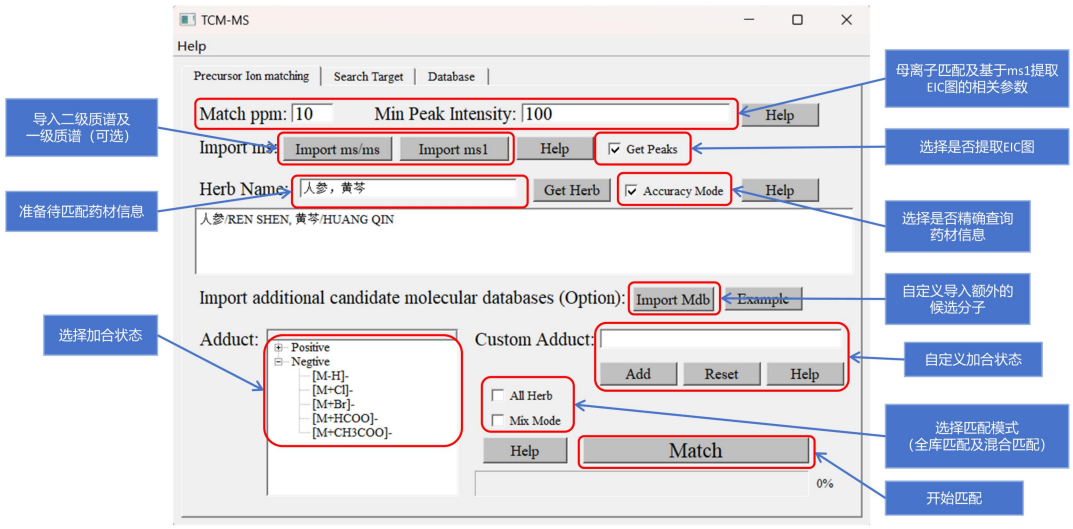

图 2 母离子匹配界面

3. 未知化合物识别

STIWS 可用于靶向识别二级谱图中特定类型的未知化合物，在使用时用户仅需准备待识别目标类型化合物所需具备的子结构即可（见图 3）。在准备时用户需完成一个 Excel 文件，该文件由目标类型的名称、子结构的 SMILES、子结构必须出现的取代基团以及子结构所有可能的其它取代基团（SMARTS 形式）等四部分组成。例如简单黄酮碳苷的子结构为简单黄酮母核，其必须出现的取代基团为碳糖，而其它可能的取代基团为羟基和甲氧基。在具体编写时，子结构的 SMILES 可直接通过 Pubchem（https://pubchem.ncbi.nlm.nih.gov/）搜索下载，如未搜索到相关 SMILES 则可基于 ChemDraw 绘制化合物结构并导出为 SMILES。在编写取代基团时，如对取代基团无要求则设置为 All 即可。相较 SMILES，采用 SMARTS 编写取代基团具有更高的灵活性，其可精确标记取代基与子结构的连接方式，而 SMILES 则无法做到。例如“[Cs:1]-[#8:2]-[H3;#6:3]”表示甲氧基，其中“Cs”代表子结构上的原子。当必须出现或可能出现的取代基团存在多个时，各取代基团用“/”进行分隔，例如“[Cs:1]-[#8:2]-[H3;#6:3]/[Cs:1]-[#8:2]-[#1:3]”表示甲氧基和羟基均可。SMARTS 的编写规范与语法说明可见表 1，在编写时环闭合用数字标记（如“C1CCCC1”表示环戊烷），分支结构用括号分隔（如“C(C)(C)O”表示叔丁醇）。需要注意的是，在表 1.1 中以“Fr”代替自由基以表示甲氧基脱甲基等裂解规则。

表 1 SMARTS 简要编写规范

SMARTS	说明
[+1]	带有一个正电荷的原子
[c]	小写原子符号表示芳香原子
[C]	大写原子符号表示非芳香原子
[#6]	原子序数为 6 的原子（c 或 C）
[H1]	与一个氢相连的原子
[#6:1]	代表编号为 1 的碳原子
*	任何原子
-	单键
=	双键
:	芳香键
#	三键
,	或
;	且

点击“Import Target Smiles”导入子结构文件后需进一步选择正负离子模式并设置召回率相对于精确率的权重。较高的权重值意味着相对于识别结果的精确性，模型更期望尽可能地挖掘二级谱图中所有目标类型的未知化合物。设置完毕后仅需点击“Train Model”即可，STIWS 将自动完成数据集的划分、二级质谱的向量化表征、模型的训练与测试等。所有步骤执行完毕后，模型的测试结果将展示在“Training Results”框中。已训练完毕的模型将自动保存在 STIWS 根目录下的“models”文件夹中，用户可自行修改模型的名称。在已有模型的情况下，用户可直接通过“Choose Model”导入已训练的模型。同时，用户也可基于“Get Key Ion”功能提取模型在识别目标类型化合物时所考察的特征离子和特征中性丢失。STIWS 采用的提取方法同样采用的是 SHAP 值分析，相应计算结果将可视化保存在 STIWS 根目录下的 KeyIon 文件夹中。最后，用户点击“Import ms/ms”导入 mgf 格式的二级质谱数据后，再点击“Search Target”选择结果的输出路径即可基于已训练的模型进行识别分析。如已有的二级质谱数据非 mgf 格式，则用户可使用 ProteoWizard（https://proteowizard.sourceforge.io/download.html）的 MSConvert 工具将二级质谱数据转换为 mgf 格式后再进行分析。STIWS 还支持

用户导入 ms1 格式的一级质谱数据（“Import ms1”）从而进行更全面的分析。当用户导入一级质谱数据时，STIWS 将自动提取各个母离子的 EIC 图。在提取时用户需要输入两个参数，即提取 EIC 图时的 ppm 以及最小峰高。然后 STIWS 将根据各二级谱图在 EIC 图上的保留时间确定其所属峰的序号，所有峰值低于最小峰高的峰的序号均设为 0。用户同样可采用 MSConvert 工具将非 ms1 格式的质谱数据转换为 ms1 格式，需要注意的是二级质谱数据同样也可提取 ms1 格式的一级质谱信息以供 STIWS 分析。

分析完毕后 STIWS 将自动输出两个 Excel 文件（见图 4）。第一个文件为完整的分析结果，其包含 PeaksIdx（EIC 图中的峰序号）、PeaksIntensity（EIC 图中的峰强度）、RTINSECONDS（保留时间）、PEPMASS（母离子 *m/z*）、*m/z*（碎片离子 *m/z* 及对应丰度）以及 Identify Label（模型识别的化合物类型）。而另一个文件则是精简后的分析结果，该结果仅在用户导入一级质谱数据后才会生成。与第一个文件相比，其在各母离子的每一个 EIC 峰下仅保留碎片离子最多的一张二级谱图。然而，由于碎片离子多并不代表该二级谱图的识别结果可靠性高，因此 STIWS 还会分别将各 EIC 峰下其它二级谱图的识别结果整合。具体而言，STIWS 将该 EIC 峰下除“Others”以外的所有结果均列举在碎片离子最多的二级谱图的识别结果中，例如“flavoneO or (flavone)”中“flavoneO”是 STIWS 在精简过程中所保留的 EIC 峰中碎片离子最多的二级谱图的识别结果，而“flavone”则是该 EIC 峰下其它二级谱图的识别结果。STIWS 的各部分均设有“Help”按钮，用户点击“Help”后 STIWS 将弹出窗口对该部分内容进行详细解释。

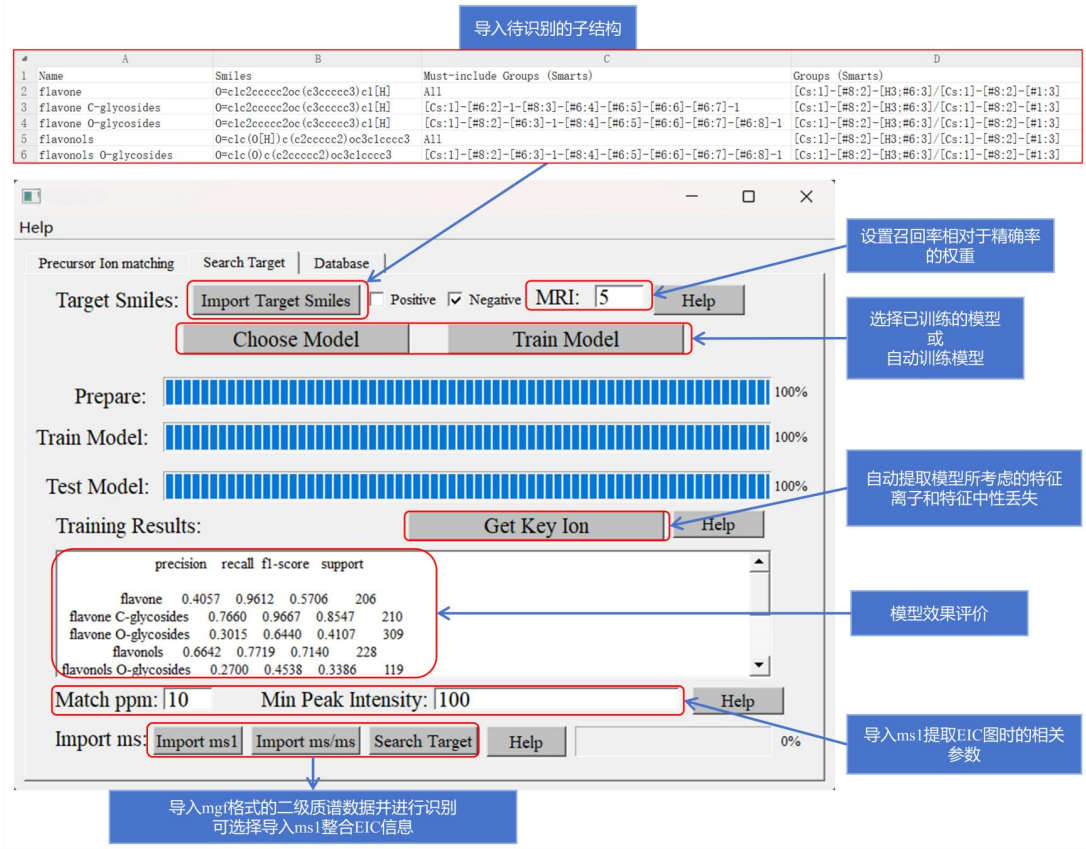

图 3 未知化合物识别界面

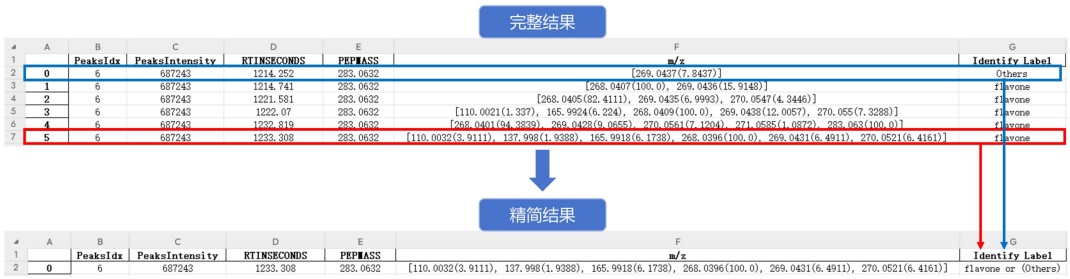

图 4 输出完整结果以及精简结果

4. 已知化合物鉴定

用户在使用时需要导入两个 Excel 表格，第一个表格记录了待鉴定的化合物类型，其具体格式可见未知化合物识别部分。第二个表格则记录了需要考虑的裂解规则，裂解规则采用 SMARTS 编写，例如黄酮类化合物 1,3 键断裂的 RDA 裂解可表示为：

“[c:1]1[c:2]2[#8:3][#6:4](=,:[#6:5][#6:6](=[O:7])[c:8]2[c:9][c:10][c:11]1)-[c:12]3[c:13][c:14][c:15][c:16][c:17]3>>[#6:1]1-[#6:2](=[#8:3])-[#6:8](=[#6:6](=[O:7]))-[#6:9]=[#6:10]-[#6:11]=1.[#6:5]#[#6:4]-[c:12]1[c:13][c:14][c:15][c:16][c:17]1”

导入两个表格后，用户点击“Train Model”即可训练模型，STIWS 将自动完成数据集的划分、谱图特征及分子特征的提取、模型的训练与测试等。所有步骤执行完毕后模型的测试结果将展示在“Training Results”框中，而已训练完毕的模型将自动保存在 STIWS 根目录下的“models”文件夹中，用户可自行修改模型的名称。在已有模型的情况下，用户可直接通过“Choose Model”导入已训练的模型。需要注意的是，由于在保存模型时会分别将模型的各个部分独立保存，因此在导入模型时用户需要选择的是包含所有模型的文件夹而非某一个文件。训练完毕或导入已有模型后，用户可重新切换到母离子匹配界面进行母离子匹配，STIWS 将在母离子匹配的基础上进一步基于模型计算候选分子和二级谱图的匹配得分。最后模型会在母离子匹配输出的结果中格外添加两列，即匹配得分和基于裂解规则解释的碎片离子的 *m/z*。此外，与靶向识别的结果类似，在用户额外导入一级质谱的情况下，STIWS 同样会输出一个精简结果，该结果仅包含各 EIC 峰中碎片离子最多一张二级谱图。

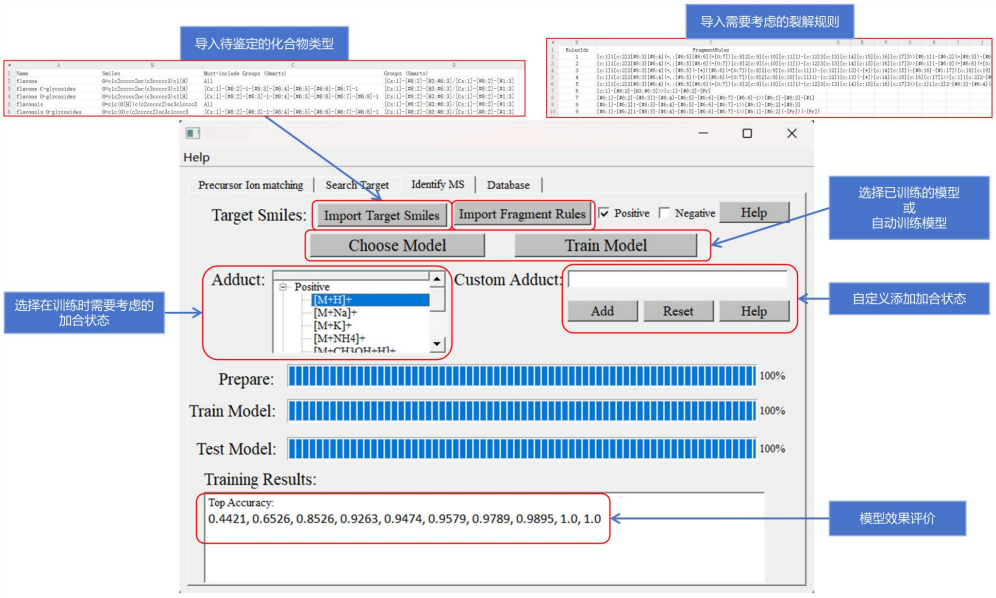

图 5 STIWS 的可视化界面