LUIZ GABRIEL SOUZA MENCUCINI

Ferramentas em Python para pré-processamento e análise de dados de metabolômica baseada em espectrometria de massas

Trabalho de Conclusão de Curso apresentado à Faculdade de Ciências Farmacêuticas de Ribeirão Preto

Nome do orientador: Ricardo Roberto da Silva

Ribeirão Preto
2025

Mencucini, Luiz

Ferramentas em Python para pré-processamento e análise de dados de metabolômica baseada em espectrometria de massas, Ribeirão Preto, 2025

Número de páginas: 36 p.

Trabalho de Conclusão de Curso apresentado à Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo.

Nome do orientador: Ricardo Roberto da Silva

1-Bioinformática; 2-Metabolômica; 3-Python

Trabalho autorizado pela comissão de graduação.

Nome do autor: Mencucini, Luiz

Título do trabalho: Ferramentas em Python para pré-processamento e análise de dados de metabolômica baseada em espectrometria de massas.

Trabalho de conclusão de curso apresentado à Faculdade de Ciências Farmacêuticas de Ribeirão Preto.

Este trabalho foi apresentado e aprovado pela Comissão de Graduação (Coordenadora do Curso) de Farmácia em 17/07/2025.

Banca examinadora

Título e Nome: _____
Instituição: _____
Julgamento: _____

Título e Nome: _____
Instituição: _____
Julgamento: _____

Título e Nome: _____
Instituição: _____
Julgamento: _____

RESUMO

MENCUCINI, Luiz. Ferramentas em Python para pré-processamento e análise de dados de metabolômica baseada em espectrometria de massas.

As ciências ômicas emergiram com o avanço das tecnologias para detecção e caracterização de moléculas que compõem as células, permitindo investigações de sistemas biológicos por meio do inventário abrangente dos conjuntos de genomas, proteomas e metabolomas. A metabolômica tem como objetivo estudar as pequenas moléculas orgânicas, sendo essencial para aplicações em medicina personalizada, desenvolvimento de fármacos e agricultura. Para acessar e analisar essas moléculas, técnicas como a espectrometria de massas (MS, do inglês *Mass Spectrometry*) têm se destacado pela capacidade de detectar uma grande variedade de compostos presentes em baixas concentrações em amostras biológicas complexas. A aplicação de MS, frequentemente combinada com métodos de separação, resulta em grandes volumes de dados, demandando ferramentas computacionais robustas. Nesse cenário, a linguagem de programação Python vem se consolidando como uma escolha proeminente por sua simplicidade, comunidade ativa e ampla gama de bibliotecas, sendo especialmente útil para resolver os desafios computacionais impostos pela complexidade dos dados metabolômicos. Neste contexto, o presente trabalho tem por objetivo revisar e contextualizar as principais ferramentas disponíveis em Python para realização do pré-processamento de análise de dados metabolômicos oriundos da MS.

Palavras-chave: Bioinformática; Metabolômica; Python

SUMÁRIO

1. INTRODUÇÃO..... 6
1.1. A METABOLÔMICA NO CONTEXTO DAS CIÊNCIAS ÔMICAS.....6
1.2. ESPECTROMETRIA DE MASSAS E METABOLÔMICA..... 8
1.3. LINGUAGEM PYTHON E A BIOINFORMÁTICA..... 10

2. OBJETIVOS..... 13

3. LIDANDO COM DADOS METABOLÔMICOS: DO PRÉ-PROCESSAMENTO À ANOTAÇÃO DE COMPOSTOS.....14
3.1. ESTRATÉGIA PARA REVISÃO DE LITERATURA.....14
3.2. CONVERSÃO DE FORMATO..... 14
3.3. PRÉ-PROCESSAMENTO DE DADOS..... 15
3.4. PROCESSAMENTO DE DADOS E ANÁLISE ESTATÍSTICA.....18
3.5. ANOTAÇÃO DE METABÓLITOS.....19
3.6. GERAÇÃO, VISUALIZAÇÃO E MANIPULAÇÃO DE REDES MOLECULARES..... 22
3.7. FLUXOS DE TRABALHO MULTIFUNCIONAIS..... 24
3.7.1. MASSCUBE.....24
3.7.2. MASS-SUITE.....25
3.7.3. RODIN..... 26
3.7.4. UMETAFLOW.....28
3.8. CONSTRUÇÃO DE FLUXOS DE TRABALHO AUTOMATIZADO..... 28
3.9. OUTROS UTILITÁRIOS PYTHON PARA METABOLÔMICA.....29
3.9.1 JUPYTER NOTEBOOK.....29
3.9.2 SMITER..... 30

4. CONCLUSÃO..... 31

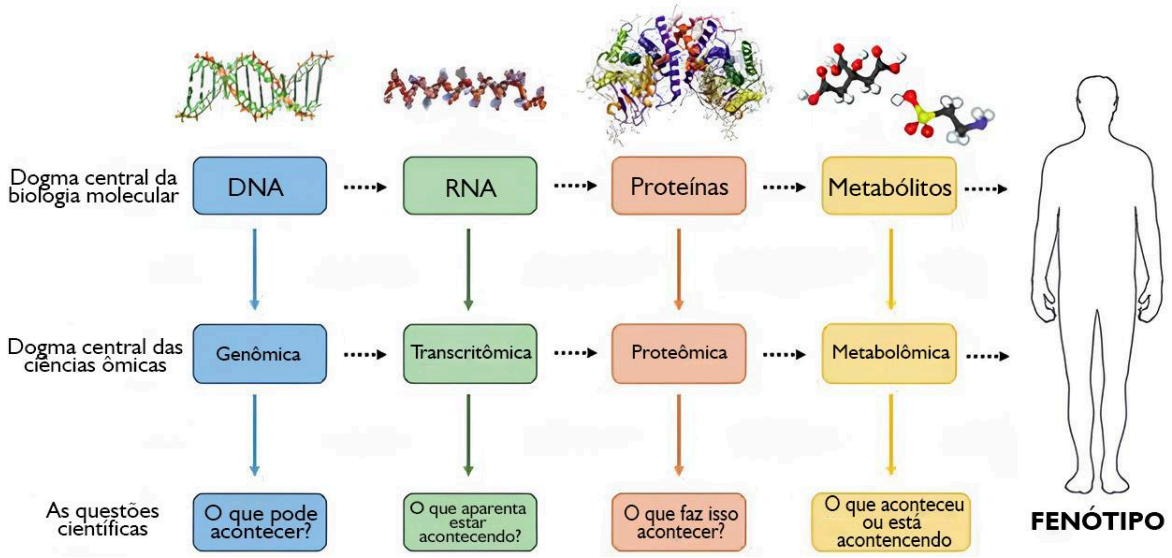
BIBLIOGRAFIA.....32

1. INTRODUÇÃO

1.1. A METABOLÔMICA NO CONTEXTO DAS CIÊNCIAS ÔMICAS

As ciências ômicas constituem um conjunto de áreas de estudos que buscam, em conjunto, explicar e entender o funcionamento de sistemas biológicos a nível molecular (Canuto et al., 2018). Essa compreensão é atingida por meio do uso de tecnologias analíticas de alto desempenho que geram uma grande quantidade de dados sobre um determinado conjunto de moléculas de um organismo (Dai; Shen, 2022), desde ácidos nucleicos, passando por proteínas, até metabólitos e lipídeos. As ômicas surgiram primariamente com o desenvolvimento de tecnologias que conduziram ao sequenciamento genômico, o qual permitiu uma ampla exploração de várias informações codificadas nos genomas, através da leitura da sequência de bases nitrogenadas que compõem o DNA (Ácido Desoxirribonucleico, do inglês *Deoxyribonucleic acid*), levando a grandes avanços na compreensão do papel dos genes em microrganismos, doenças genéticas, patógenos, plantas e outros organismos (Figura 1) (Hasin; Seldin; Lusis, 2017).

Figura 1 - Cascata Ômica: Principais moléculas investigadas por cada ciência ômica.



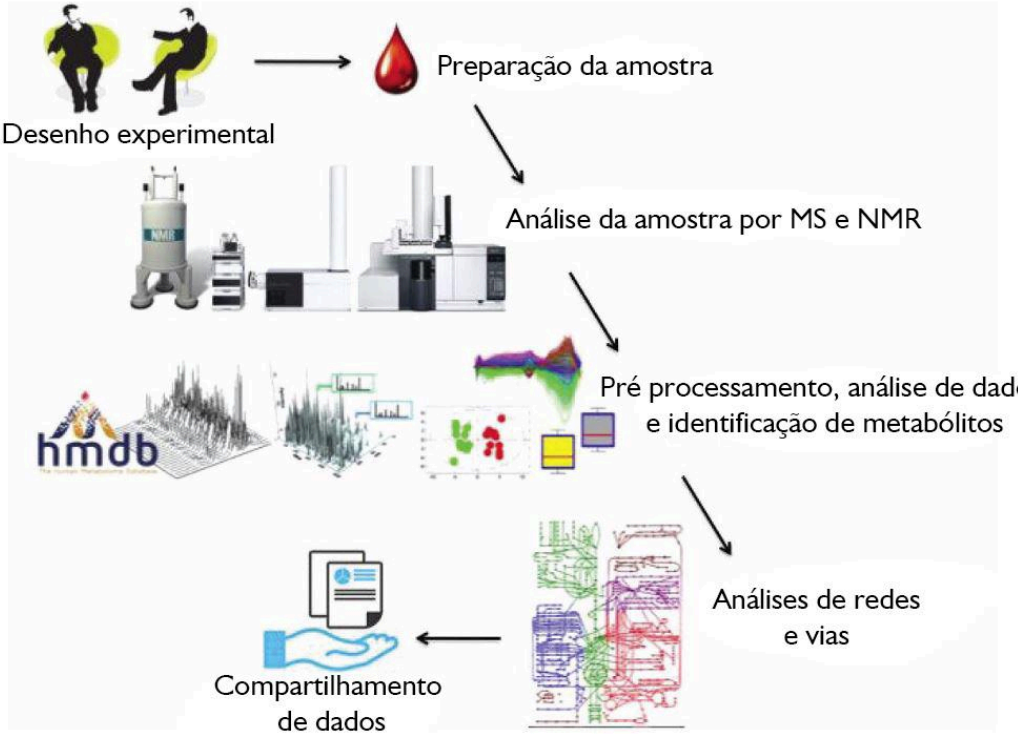
Fonte: Adaptado de (Araújo et al., 2017)

Dentre esse universo de possibilidades fornecido pelas ciências ômicas, a metabolômica se restringe ao estudo das pequenas moléculas produzidas pelo

metabolismo dos organismos (≤ 2000 Da) (Rakusanova; Fiehn; Cajka, 2023), seus processos químicos, seus derivados ambientais, além do estudo de marcadores químicos que caracterizam processos celulares, identificando, caracterizando e quantificando esses compostos (Vailati-Riboni; Palombo; Loor, 2017). Esta ciência ômica presta importante contribuição nas mais diversas áreas do conhecimento, promovendo avanços, por exemplo, para a medicina personalizada (Jacob et al., 2019), descoberta de novos fármacos (Newman; Cragg, 2020; Wishart, 2016), melhoramento vegetal (Fernie; Schauer, 2009), dentre as mais variadas áreas do conhecimento que envolvem seres vivos e suas mais diversas relações.

Dentre as possibilidades da utilização da metabolômica, uma das maiores divisões se dá entre as abordagens de metabolômica direcionada e não direcionada. Na primeira, o foco se dá no acompanhamento de um metabólito específico ou um pequeno grupo destes, e é utilizada principalmente em estudos que são guiados por hipóteses relacionadas por esse pequeno grupo de metabólitos e envolve aperfeiçoar as técnicas de aquisição dos dados para esse grupo, ou ainda, o perfil de apresentação dos mesmos em um determinado contexto de análise, como na relação com uma dada patologia, por exemplo (Chouchani et al., 2014). Já a metabolômica não direcionada consiste em adquirir dados de uma grande quantidade de metabólitos presentes nas amostras, geralmente focando em caracterizar este ambiente químico (Wehrens; Salek, 2019), sem, todavia, ter um alvo específico em vista. Via de regra, a metabolômica não direcionada é amplamente utilizada dada sua possibilidade de descoberta, não apenas de novas estruturas, mas de novas vias metabólicas, e como consequência, desvelar novos processos biológicos (Arini et al., 2025).

Figura 2 - Fluxo de trabalho básico em metabolômica não direcionada.



Fonte: Adaptado de (Wehrens; Salek, 2019)

O fluxograma básico de um experimento de metabolômica não direcionada está representado na Figura 2, e consiste basicamente em um desenho experimental que busca acessar os compostos de forma abrangente, preparar a amostra de maneira que metabólitos de interesse não sejam afetados, adquirir os dados pela técnica escolhida, pré-processar e processar os dados adquiridos de maneira a conseguir anotar os metabólitos contidos na amostra, e, por fim, usar técnicas estatísticas para explorar esses dados e gerar hipóteses, que podem ser testadas e possibilitar a geração de novo conhecimento (Wehrens; Salek, 2019).

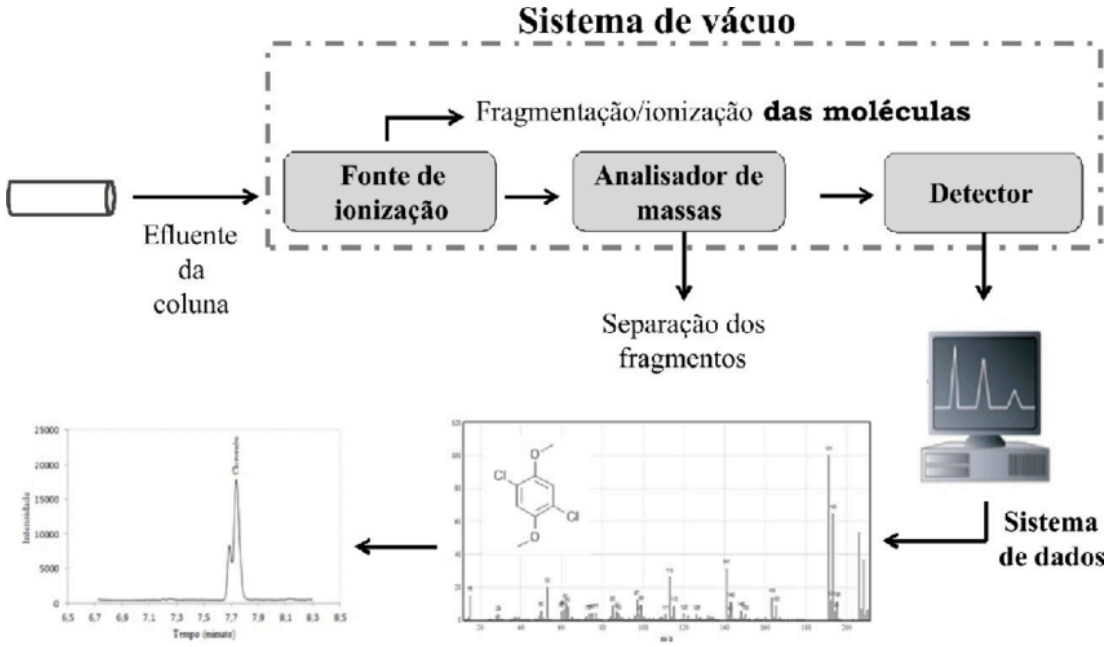
1.2. ESPECTROMETRIA DE MASSAS E METABOLÔMICA

Para acessar os dados dos ambientes químicos discutidos anteriormente, é possível lançar mão de algumas técnicas analíticas, dentre as quais a mais utilizada

é a Espectrometria de Massas (MS, do inglês *Mass Spectrometry*), que será o foco principal deste trabalho, por sua maior adoção pela comunidade científica no contexto da metabolômica e a disponibilidade de *software* aberto para processar os dados resultantes dos experimentos disponíveis (Baidoo; Teixeira Benites, 2019). Todavia, cabe o destaque de que a Ressonância Magnética Nuclear (NMR, do inglês *Nuclear Magnetic Resonance*) também é amplamente utilizada e, hoje, essa técnica é considerada o padrão ouro na identificação de compostos, por seu conteúdo informativa para elucidação estrutural, sua maior reprodutibilidade e potencial para experimentos quantitativos, apesar de ser menos sensível que a MS (Markley et al., 2017). Além de MS e NMR, outras técnicas analíticas também podem ser utilizadas, muito embora representam uma pequena fração dos estudos em metabolômica, como as espectroscopias no infravermelho, UV-VIS e RAMAN, cada uma destas com suas vantagens e desvantagens (Kruk et al., 2017). Dentre as vantagens, no tocante à espectrometria de massas, destacam-se sua capacidade de análises em larga escala (González-Domínguez et al., 2019), bem como a alta sensibilidade, possibilitando análises de amostras com pequena massa do material biológico de interesse (Anh et al., 2024; Dettmer; Aronov; Hammock, 2007).

Em síntese, o princípio da espectrometria de massas consiste na conversão do analito em íon no estado gasoso ao passar pela fonte de ionização, adquirindo carga, negativa ou positiva (Figura 3). Tais íons, então, adentram o espectrômetro de massas, passam pelo analisador de massas, onde estes são separados de acordo com sua razão massa/carga (m/z), chegando ao detector de massas, onde têm sua razão m/z registrada. O detector produz sinais que são interpretados computacionalmente, gerando dados, que fornecem tanto informações qualitativas, acerca da estrutura das moléculas presentes na amostra, como quantitativas, acerca de sua abundância relativa (Ho et al., 2003).

Figura 3 - Esquema simplificado de um espectrômetro de massas.



Fonte: (Nascimento et al., 2018)

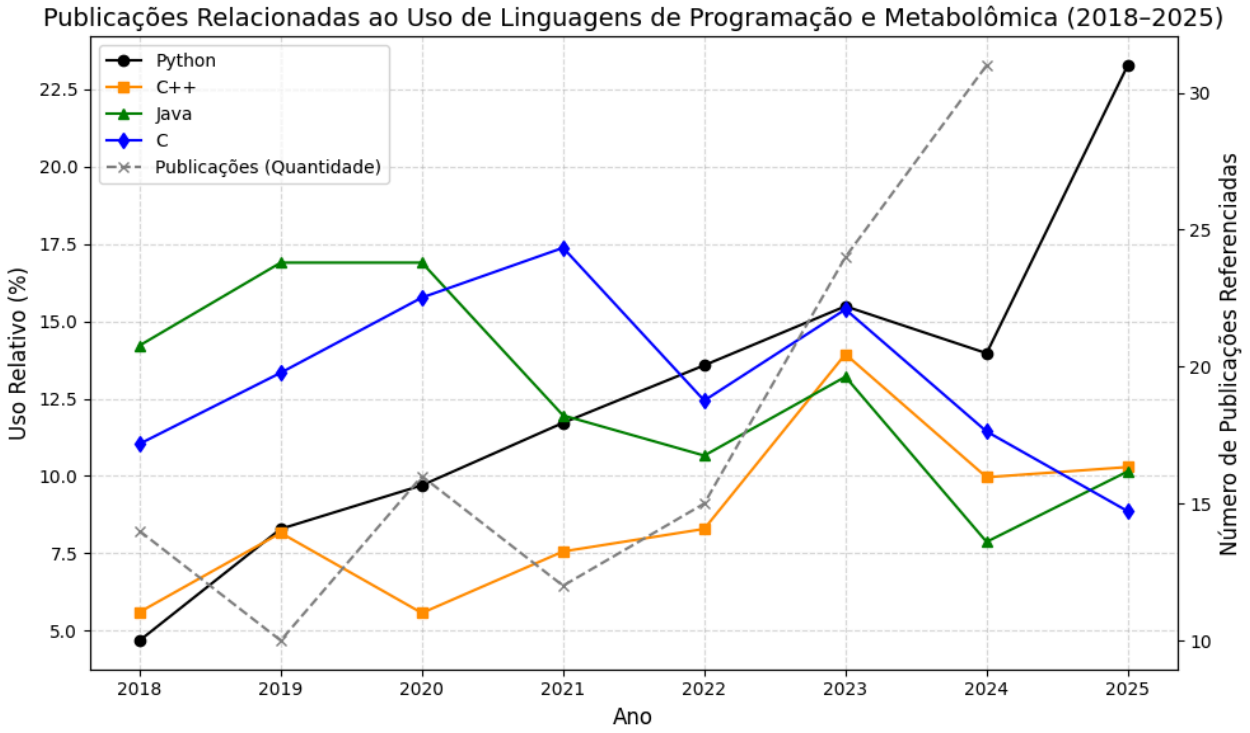
Além disso, geralmente são acopladas técnicas cromatográficas de separação de misturas em conjunto com as técnicas de detecção, permitindo que amostras complexas sejam analisadas com esta técnica. A cromatografia líquida (LC, do inglês *Liquid Chromatography*) ou cromatografia gasosa (GC, do inglês *Gas Chromatography*), a depender do desenho experimental e das características individuais da amostra, são amplamente utilizadas e perfazem parte significativa do escopo de análise em experimentos metabolômicos baseados em espectrometria de massas.

1.3. A LINGUAGEM PYTHON E A BIOINFORMÁTICA

A quantidade de dados ômicos gerados cresce ano após ano, o que demanda cada vez mais a criação de ferramentas computacionais que permitam lidar com o volume e a complexidade dos dados ômicos (Perez-Riverol et al., 2019). A linguagem Python se tornou popular pela sua sintaxe de simples compreensão por conta de sua semelhança com a língua inglesa no ano de 2024, alcançando o posto de linguagem de programação mais utilizada globalmente (Staff, 2024). Em se tratando de ciências ômicas, esta linguagem de programação acompanhou o

crescimento da popularidade, entre outros fatores, pela sua natureza de código aberto, ampla e ativa comunidade, ecossistema poderoso de pacotes úteis para processar uma grande quantidade de dados, além de tirar vantagem do crescimento exponencial das técnicas de inteligência artificial, avançando em um contexto que antes era dominado pela linguagem de programação R, conhecida por seus poderosos pacotes estatísticos. Em comparação com outras linguagens de programação, como C++, Java e C, o Python teve um maior crescimento na sua adoção na área da metabolômica (Figura 4), tendo saído do posto de menos utilizada para mais utilizada nos últimos 7 anos, se comparado a essas 4 linguagens de programação.

Figura 4 - Gráfico representando o crescimento das publicações em metabolômica que se utilizam de linguagens de programação (eixo das ordenadas à direita) bem como o crescimento do uso da linguagem Python para esta ciência ômica (eixo das ordenadas à esquerda).



Fonte: Elaborado pelo Autor

Tais fluxos de trabalho automatizados geralmente podem ser definidos como sequências de análise automatizadas, nas quais os dados de saída de uma etapa são ou um dos resultados de análise ou a entrada para a próxima etapa de análise. Este nível de automatização e personalização permite ao mesmo tempo adequar os

parâmetros de análise ao necessário para a resolução do problema de pesquisa específico e reduzir drasticamente o tempo de processamento dos dados (Leipzig, 2017).

2. OBJETIVOS

A presente monografia tem por objetivo, por meio de revisão da literatura, identificar, catalogar e apresentar as ferramentas desenvolvidas em Python mais utilizadas, atuais e otimizadas para a utilização em fluxos de trabalho de metabolômica baseada em espectrometria de massas, explorando o funcionamento e os principais diferenciais de cada uma delas ao discorrer acerca das principais etapas de um fluxo de trabalho em metabolômica.

3. LIDANDO COM DADOS METABOLÔMICOS: DO PRÉ-PROCESSAMENTO À ANOTAÇÃO DE COMPOSTOS

3.1. ESTRATÉGIA PARA REVISÃO DE LITERATURA

O presente trabalho optou por uma “revisão de literatura narrativa”, tentando resumir ou sintetizar o que foi escrito sobre o tópico para contextualizar a aplicação das ferramentas em um fluxo de trabalho padrão, sem a pretensão de generalizar o conhecimento acumulado no tópico.

3.2. CONVERSÃO DE FORMATO

Os dados de LC-MS ou GC-MS gerados em experimentos de metabolômica são geralmente compostos de três dimensões: tempo de retenção, razão massa/carga (m/z) e intensidade. Todavia, recentemente os novos aparelhos têm apresentado a capacidade de agregar ainda uma quarta dimensão, com a possibilidade de medida da seção de colisão cruzada (CCS, do inglês *Cross Collision Section*), a qual permite a obtenção de dados referentes aos isômeros de uma dada molécula. Os dados gerados pelo espectrômetro de massas geralmente são armazenados no formato de dados da fabricante do equipamento. Via de regra, cada fabricante desenvolve seu próprio formato e conjunto de ferramentas computacionais para processar (ler, analisar e escrever) os dados gerados por seus equipamentos, com o uso de *software* proprietário (Adusumilli; Mallick, 2017).

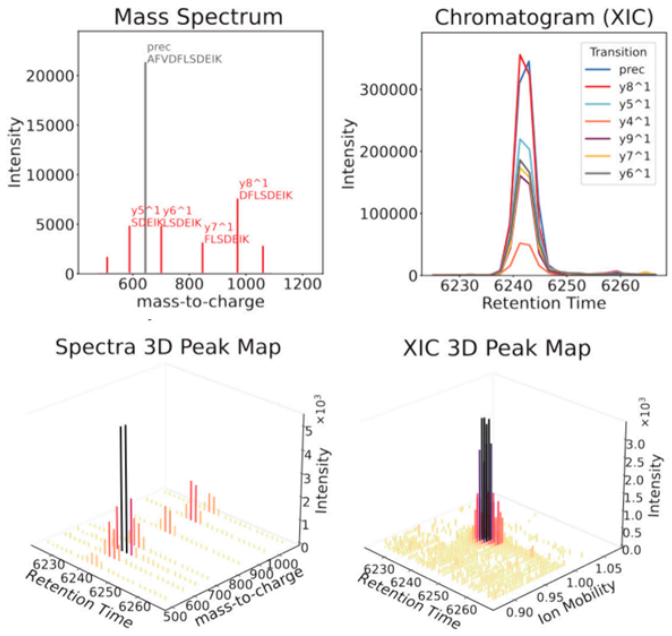
Por isso, as ferramentas *open-source* desenvolvidas pela comunidade científica adotaram um padrão de organização dos dados para que os arquivos proprietários de todos os fabricantes possam ser utilizados, o que é um fator determinante para o desenvolvimento das ômicas (Quackenbush, 2004).

Portanto, para que possa ser iniciado o pré-processamento dos dados em metabolômica, faz-se necessária a conversão de dados do formato proprietário para um formato público, geralmente .mzML ou .mzXML. Para esta conversão, a principal ferramenta utilizada atualmente é o *MSConvert*, contido no pacote de ferramentas *ProteoWizard* (Chambers et al., 2012). O *ProteoWizard MSConvert*, apesar de não ser escrito em Python, é muitas vezes utilizado em fluxos de trabalho

computacionais em metabolômica, podendo ser facilmente empregado nesse contexto em seu formato de interface de linha de comando.

Alternativamente a esta ferramenta, o *software OpenChrom* possibilita importar dados de espectrometria de massas de diversos fabricantes e exportar em formatos públicos (Wenig; Odermatt, 2010), sendo uma ferramenta de interface gráfica que ainda dispõe de ferramentas de visualização e análise de dados provenientes desta técnica analítica. Como forma de analisar os dados brutos de espectrometria de massas convertidos, pode se utilizar o *pyOpenMS-viz*, uma plataforma para visualização de dados espectrométricos, através das funções de gráficos de espectros, mapa de picos, cromatogramas ou mobilogramas (Figura 5) (Sing et al., 2025).

Figura 5 - Exemplos de visualizações disponíveis ao usuário no pacote *pyOpenMS-viz*.



Fonte: Adaptado de (Sing et al., 2025)

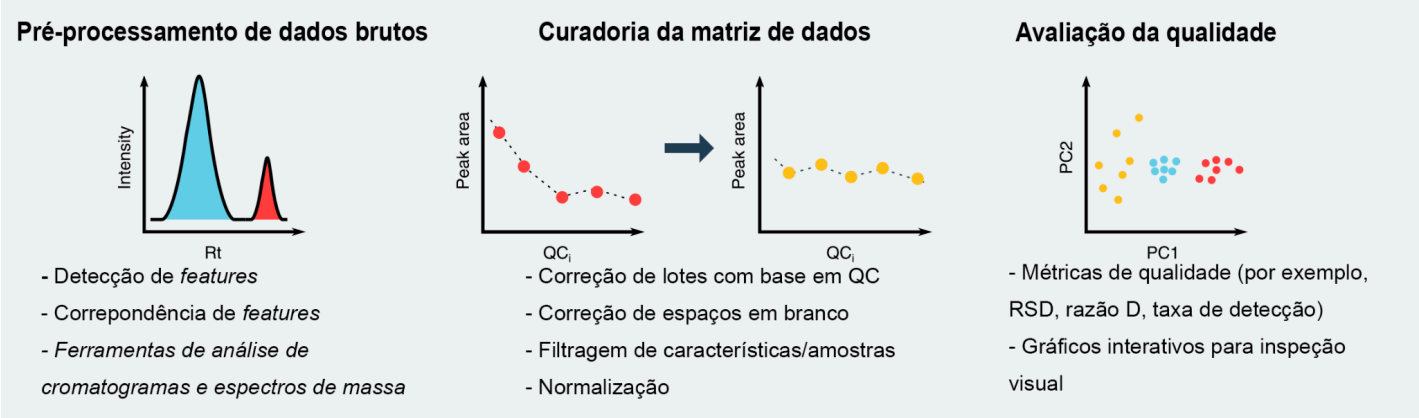
3.3. PRÉ-PROCESSAMENTO DE DADOS

Com a conversão realizada, é possível acessar os dados brutos do experimento através das ferramentas *open-source*. Então é iniciado o pré-processamento propriamente dito, que é composto, via de regra, pelas fases de (i) filtragem de ruído e correção do tempo de retenção, que objetiva principalmente

separar um composto químico presente na amostra de ruídos obtidos por diversas interferências possíveis em um ambiente de altíssima sensibilidade, (ii) detecção dos picos, (iii) alinhamento e (iv) normalização (Katajamaa; Orešič, 2007).

Entre as ferramentas Python para o pré-processamento de dados metabolômicos está o *TidyMS* (Riquelme et al., 2020), voltado ao pré-processamento de dados de LC-MS em metabolômica não direcionada, com ênfase na curadoria baseada em controle de qualidade. Suas principais funcionalidades são descritas na Figura 6. Esta biblioteca utiliza o algoritmo baseado em CWT (do inglês *Continuous Wavelet Transform*) para realizar as funções de detecção de picos (*peak picking)* e o algoritmo centWave (Tautenhahn; Böttcher; Neumann, 2008) para detecção de *features* (agrupamentos de íons, ou picos, derivados do mesmo composto) em leituras do espectrômetro consecutivas (*scans)*. O pacote permite a aplicação de filtros estatísticos, correção de deriva instrumental e remoção de contaminantes, promovendo matrizes de dados mais robustas e confiáveis. Embora possua funcionalidades para análise de dados brutos, sua utilização é especialmente recomendada na etapa de curadoria após a extração de *features*.

Figura 6 - Funcionalidades básicas da biblioteca TidyMS.

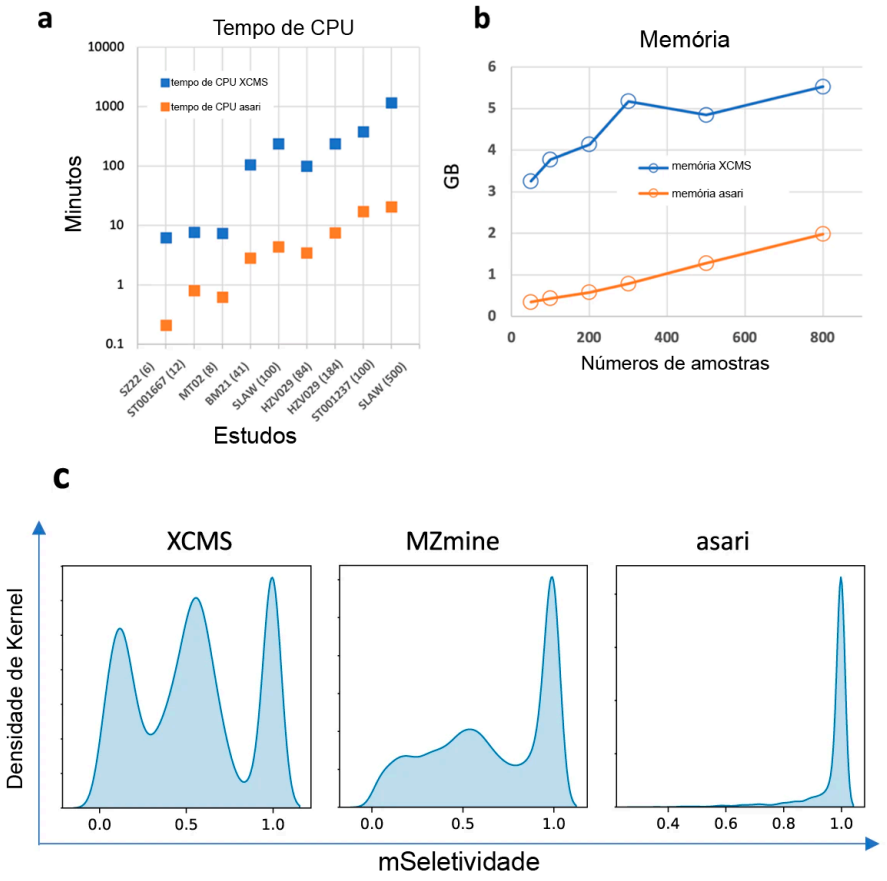


Fonte: Adaptado de (Riquelme et al., 2020)

O *asari* (Li et al., 2023) é outro dos *softwares* utilizados nessa etapa de metabolômica por LC-MS. Os autores da ferramenta demonstraram uma precisão e desempenho maior do que as ferramentas mais utilizadas atualmente para a etapa de pré-processamento dos dados de LC-MS, *MZmine* (Schmid et al., 2023), *XCMS*

(Smith et al., 2006) e *MS-DIAL* (Tsugawa et al., 2015), no que diz respeito à qualidade dos picos identificados, utilizando o parâmetro de mSeletividade, que demonstra o quão únicas são as *features* detectadas, o que leva a menos falsos positivos nesse quesito (Figura 7c), qualidade da quantificação dos picos cromatográficos e eficiência computacional, utilizando menos memória e levando menos tempo de análise que as outras ferramentas (Figuras 7a e 7b).

Figura 7 - Comparação do desempenho computacional do asari com as principais ferramentas utilizadas em pré-processamento de dados metabolômicos.



Fonte: Adaptado de (Li et al., 2023)

O *PyMS* (O'Callaghan et al., 2012) é um conjunto de ferramentas disponibilizado com um pacote Python que objetiva pré-processar dados oriundos de GC-MS, contendo métodos de leitura de dados brutos, correção da linha de base, filtragem de ruído, detecção de picos, alinhamento de tempo de retenção e normalização dos dados.

Ainda no tocante aos pacotes Python de código aberto para pré-processamento de dados de MS, o *pyOpenMS* implementa uma série de rotinas disponibilizadas originalmente na linguagem C++. Com esse pacote, é possível, dentre uma série de outras funções, importar arquivos de MS em formato público e submeter ao fluxograma básico de pré-processamento, possuindo funções para auxiliar tanto com experimentos em metabolômica direcionada quanto em metabolômica não direcionada (Röst et al., 2014).

Outro importante pacote Python com funcionalidades para o pré-processamento de dados de metabolômica é o *spectrum_utils* (Bittremieux, 2020), que contém principalmente funções para pré-processar dados de espectros de fragmentação (MS/MS) para realização de busca espectral em bibliotecas e também para criar imagens de alta resolução que possam ser utilizadas em exploração de dados, publicações ou visualização em páginas *web*. Recentemente o pacote passou por atualizações que adiciona funcionalidades para integrar os novos padrões de dados da comunidade de espectrometria de massas (Bittremieux et al., 2023).

Outras bibliotecas com múltiplas funcionalidades, além do pré-processamento, incluem a *Mass-Suite* (Hu et al., 2023), *MassCube* (Fiehn et al., 2025), *NP*^{3} (Bazzano et al., 2024) e *PyMS* (O'Callaghan et al., 2012), que serão melhor explicitadas na seção dedicada a ferramentas com múltiplas funções. Além dos citados, outros aplicativos primariamente desenvolvidos para serem utilizados como GUI (interface gráfica do usuário, do inglês *graphical user interface)*, como o *MZmine* (Schmid et al., 2023), podem ser conectados em fluxos de trabalhos escritos em Python.

3.4. PROCESSAMENTO DE DADOS E ANÁLISE ESTATÍSTICA

Após o pré-processamento, os arquivos de dados brutos obtidos por cromatografia acoplada à espectrometria de massas, contendo milhões de registros de m/z, tempo de retenção e intensidade, são resumidos em um formato tabular consolidando o resultado na matriz de *features.* Esta contém, fundamentalmente, a área sob a curva cromatográfica em cada amostra para cada *feature* detectada, juntamente com informações cruciais para a anotação desses íons, como o tempo

de retenção (TR) e a razão massa/carga (m/z). De posse dessa matriz de dados, emerge a necessidade de um *software* de código aberto que seja não apenas capaz de ler e editar tal matriz, mas que também possibilite a realização de múltiplas análises subsequentes. Essas análises podem abranger desde a normalização e filtragem avançada dos dados, passando por análises estatísticas univariadas e multivariadas para a descoberta de biomarcadores, até a integração com bancos de dados para anotação de metabólitos e interpretação biológica dos resultados. No ecossistema Python, o pacote mais proeminente e amplamente adotado para essa finalidade é o *Pandas* (McKinney, 2010).

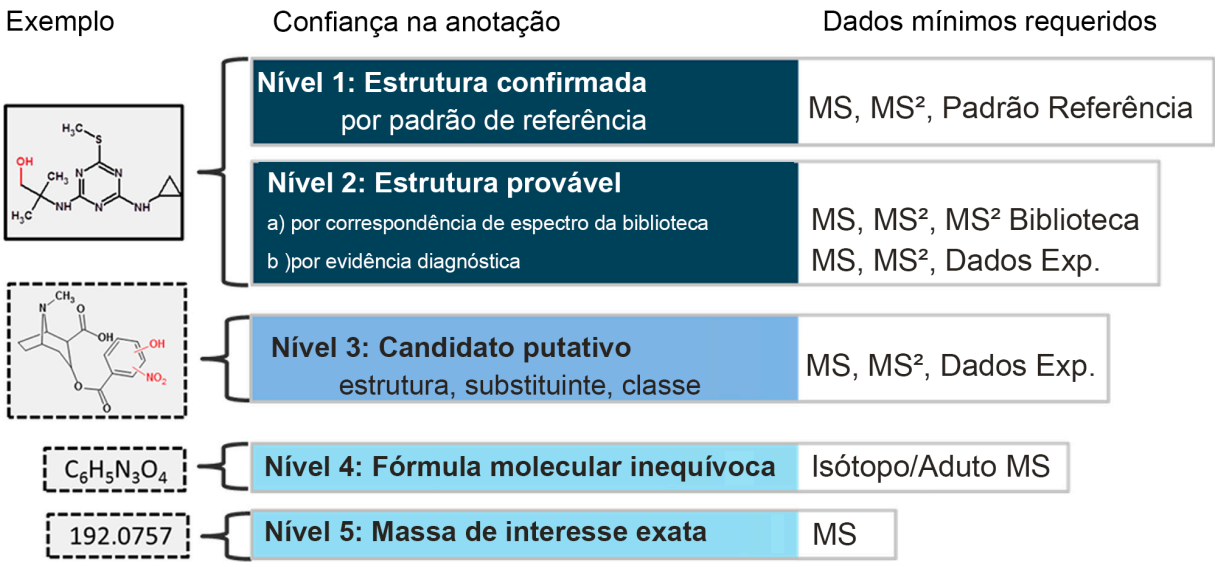
Para complementar as ferramentas de análise estatística, o pacote *metbit* (Aeiwz, 2025) surge como uma solução em Python dedicada à exploração de dados metabolômicos por meio de técnicas multivariadas. Seu foco principal reside na aplicação e visualização de métodos como a Análise de Componentes Principais (PCA, do inglês *Principal Component Analysis*) e a Análise Discriminante por Mínimos Quadrados Parciais Ortogonais (PLS-DA, do inglês *Partial Least-Squares Discriminant Analysis*), que são análises de fácil interpretação e possibilitam identificar padrões e gerar hipóteses para caracterização de biomarcadores potenciais em conjuntos de dados complexos. Ao processar a matriz de *features* gerada nas etapas anteriores, o *metbit* auxilia na redução da dimensionalidade e na interpretação das variações entre diferentes grupos experimentais, fornecendo gráficos informativos que facilitam a compreensão dos resultados e a seleção de *features* de interesse para investigações subsequentes.

3.5. ANOTAÇÃO DE METABÓLITOS

A anotação de metabólitos em metabolômica é sistematizada de forma hierárquica nos níveis: nível 1, estrutura confirmada, no qual a estrutura proposta foi confirmada por meio de uma apropriada avaliação de um padrão referência por correspondência em MS, MS/MS e tempo de retenção, além de algum outro método ortogonal, se possível, como a NMR; nível 2, estrutura provável, a qual indica que é possível propor uma estrutura exata usando diferentes evidências, auxiliado principalmente pelo pareamento espectral; nível 3, candidato(s) provisório(s), uma área em que há evidências para possíveis estruturas, mas informação insuficiente

para uma exata; nível 4, fórmula molecular inequívoca, quando há dados para se chegar a uma única fórmula molecular, porém sem informações para possíveis estruturas; nível 5, massa exata, quando é possível determinar a massa molecular do composto, porém não há informações suficientes para ser atribuída uma fórmula molecular (Figura 8) (Schymanski et al., 2014).

Figura 8 - Sistematização proposta para a confiança da anotação em metabolômica de acordo com os tipos de dados disponíveis.



Fonte: Adaptado de (Schymanski et al., 2014)

A principal forma de se anotar os compostos se dá pela comparação dos espectros de massa das *features* contidos na amostra com os de bases de dados desses espectros. Há bases que armazenam apenas dados provenientes de apenas uma configuração de equipamento, outras possuem mais de uma destas (GC-MS e LC-MS, por exemplo). Essas bases de dados também se diferenciam pelo nível de curagem, bem como da matriz de origem destes espectros (Bittremieux; Wang; Dorrestein, 2022; Vinaixa et al., 2016).

Essa comparação entre os espectros experimentais e de referência pode ser efetuada através de diferentes maneiras, porém, a mais utilizada é através do valor de similaridade de cosseno e suas variações com base nos espectros de fragmentação dos compostos. O valor de similaridade de cosseno básico é calculado através da representação dos espectros como vetores no espaço n-dimensional, onde cada dimensão corresponde a um valor de razão massa-carga

(m/z) e o valor associado representa a intensidade do pico. A similaridade de cosseno é então determinada pelo ângulo entre esses vetores, sendo calculada pela razão entre o produto escalar das intensidades dos picos pareados e o produto das normas (norma euclidiana) desses vetores. O valor resultante varia entre 0 e 1, sendo 1 indicativo de máxima similaridade (espectros idênticos em termos de distribuição relativa de intensidades) e 0 quando não há qualquer sobreposição ou alinhamento vetorial. Outras métricas utilizadas para o cálculo da similaridade são *the fit*, *reverse fit* e *purity scores,* por exemplo, utilizados no HMDB (Wishart et al., 2021), além de métodos baseados em probabilidade, como o X-Rank (Mylonas et al., 2009).

Entre os pacotes Python que permitem comparar estruturas podemos citar o *matchms* (Huber et al., 2020), que, apesar de possuir outras métricas, utiliza o cálculo de cosseno e alguns derivados para calcular similaridade entre espectros, possuindo ainda métodos de filtragem dos espectros para remoção de ruído. Recentemente, o *matchms* inseriu funcionalidades para inspecionar e corrigir bibliotecas de espectros de fragmentação (que podem ser construídas por qualquer usuário da ferramenta) através da validação das anotações estruturais, resultando em bibliotecas espectrais mais limpas e precisas, melhorando a qualidade nos resultados de busca nessas bibliotecas (De Jonge et al., 2024).

Novos métodos para calcular a similaridade estrutural em espectros de fragmentação têm emergido, demonstrando performance melhor que os algoritmos baseados no valor de cosseno. Como exemplos podemos citar os métodos *Spec2Vec* (Huber et al., 2021a), inspirado no algoritmo de processamento de linguagem natural *Word2Vec*, e o *MS2DeepScore* (Huber et al., 2021b), que utiliza uma rede neural siamesa para prever a similaridade entre as estruturas dos compostos com base em seus espectro de fragmentação, e não apenas a similaridade entre os espectros, assim como na maioria das abordagens. Ambas foram projetadas para serem utilizadas com dados de fragmentação (MS/MS).

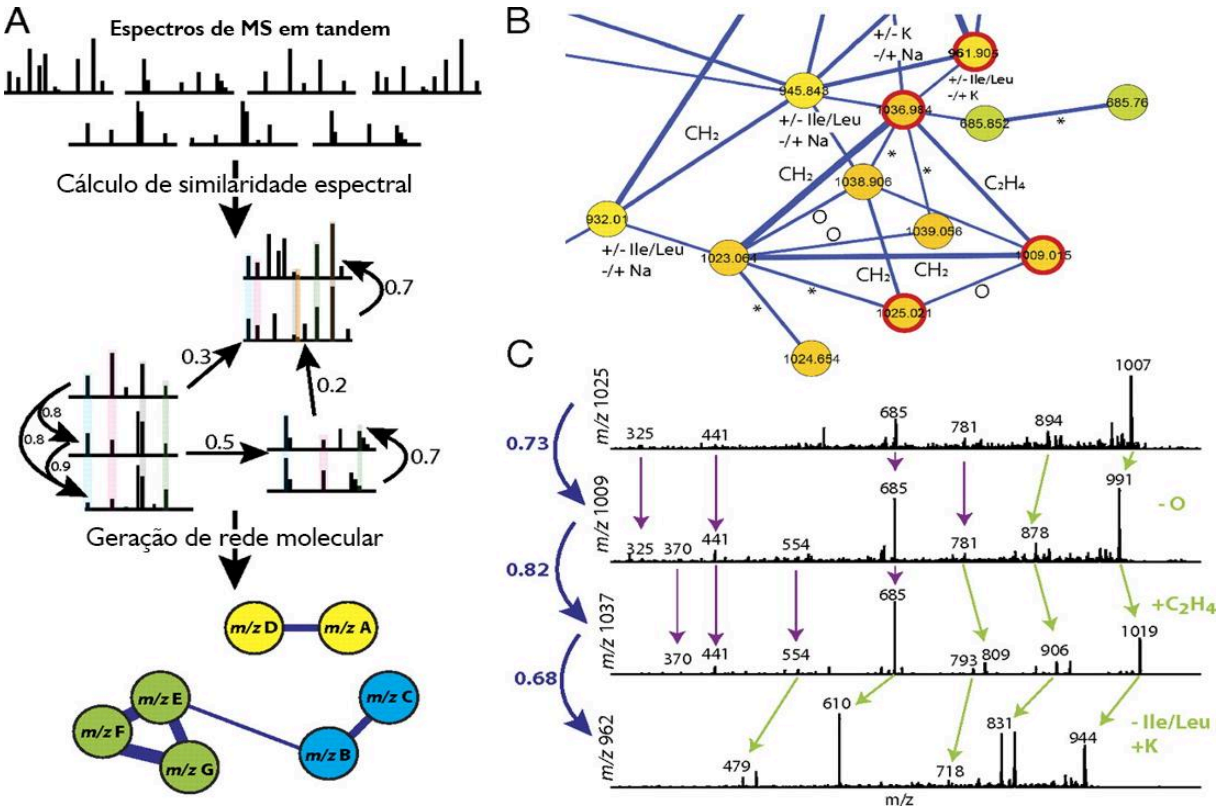
Entretanto, o elevado número de possibilidades estruturais esperados para os metabólitos detectados em amostras biológicas, faz com que apenas bancos de dados manualmente inspecionados e/ou personalizados não sejam suficientes para anotar a maior parte das moléculas detectadas por experimentos de MS. Assim sendo, surgiram outros métodos para se tentar anotar os metabólitos presentes na

amostra sem a utilização de bases de referência, principalmente baseados nos desenvolvimentos recentes da área de aprendizado de máquinas (Ebbels et al., 2023).

3.6. GERAÇÃO, VISUALIZAÇÃO E MANIPULAÇÃO DE REDES MOLECULARES

Mesmo após as etapas de pré-processamento, análise multivariada e anotação, os resultados das análises em metabolômica são complexos, e necessitam de conhecimento avançado para serem interpretados. Para auxiliar a interpretação desses amplos conjuntos de dados, a comunidade científica vem desenvolvendo estratégias analíticas, entre elas, a visualização por meio de redes moleculares (Figura 9). A criação de redes é uma abordagem computacional para organizar e interpretar os dados resultantes dos experimentos em metabolômica, na qual cada nó representa uma *feature*, através do seu espectro de fragmentação, e as conexões entre esses nós indicam a similaridade entre os espectros.

Figura 9 - Construção básica de uma rede molecular.



(A) Esquema de como ocorre a ligação entre os nós de uma rede molecular (B) Representação de uma rede molecular (C) Cálculo do valor de cosseno entre espectros pelos seus picos pareados. Fonte: Adaptado de (Watrous et al., 2012)

Por meio das redes moleculares é possível observar semelhanças e diferenças importantes dentro da amostra biológica analisada, sendo, inclusive, ferramenta importante em várias aplicações, como, por exemplo, na descoberta de novas drogas, entendimento do metabolismo de drogas e medicina personalizada, principalmente pela sua capacidade de relacionar e facilitar a visualização de anotações de moléculas com espectros de fragmentação previamente caracterizados, a moléculas similares, que não tiverem anotação atribuída por busca em biblioteca espectral (Quinn et al., 2017).

A ferramenta escrita em Python mais utilizada para a construção de redes é a biblioteca *NetworkX* (Hagberg; Swart; Schult, 2008), que pode utilizar o cálculo de similaridade espectral gerado na etapa de análise para criar a conexão entre os íons detectados. Após a construção das redes moleculares, é possível a utilização da ferramenta *ChemWalker* (Borelli et al., 2023), desenvolvida em Python, para

expandir o escopo de anotação de espectros de massa em experimentos de metabolômica não direcionada a partir da topologia de redes espectrais. Utilizando um algoritmo de caminhada aleatória em redes moleculares, o *ChemWalker* propaga informações de anotações conhecidas para nós não diretamente conectados, superando limitações de ferramentas anteriores, como o *NAP* (Silva et al., 2018). Integrando resultados de fragmentação *in silico* do *MetFrag* (Ruttkies et al., 2016) e métricas de similaridade estrutural, o *ChemWalker* constrói grafos ponderados que permitem uma propagação mais abrangente e eficiente das anotações.

3.7. FLUXOS DE TRABALHO MULTIFUNCIONAIS

Com tantas opções de ferramentas e técnicas de análise para processar dados metabolômicos, surge o desafio de criar soluções que consigam integrar as etapas de análise mencionadas, para garantir um processo mais rápido e eficiente. Nesse sentido, o desenvolvimento de fluxos de trabalho multifuncionais tem se intensificado. Eles se caracterizam pela integração das etapas supradescritas em uma única ferramenta. Destacaremos algumas das ferramentas desenvolvidas recentemente.

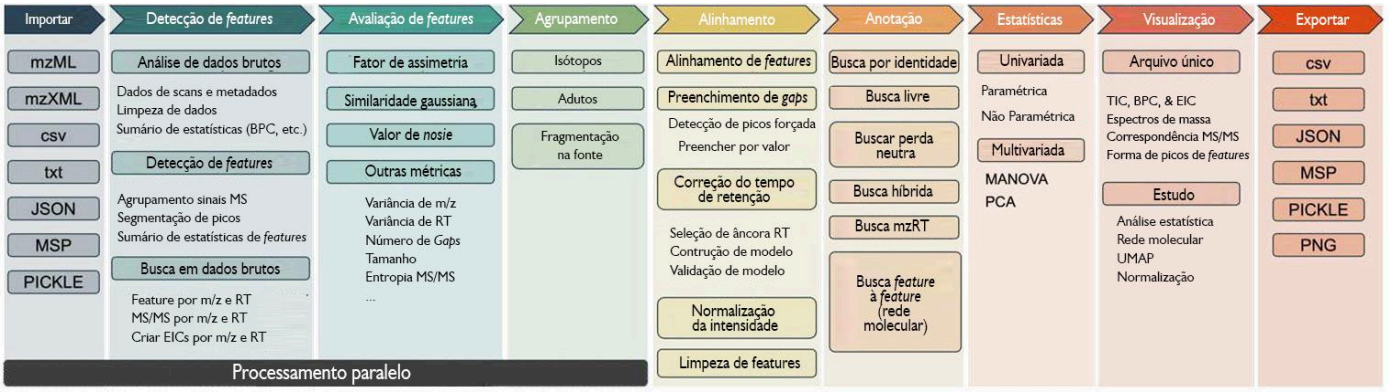
3.7.1 MASSCUBE

O *MassCube* é um *framework* para análise de experimentos de metabolômica não direcionada baseada em MS (Figura 10), que parte dos dados brutos e oferece uma gama de fluxos de trabalho diferentes que podem ser utilizados mesmo por usuários sem experiência em programação, se destacando por sua velocidade de processamento e arquitetura modular, o que permite a implementação de novos algoritmos criados pela comunidade (Fiehn et al., 2025).

Seu algoritmo de detecção de *features* é inspirado na estratégia utilizada pelo *centWave* (Tautenhahn; Böttcher; Neumann, 2008), através do grupamento de sinais juntamente com um algoritmo de detecção de bordas assistido por filtro Gaussiano. Em comparação com outras ferramentas para este fim, como o *MS-DIAL* (Tsugawa et al., 2015), *MZmine3* (Schmid et al., 2023) e *XMCS* (Smith et al., 2006), a

ferramenta mostrou possuir uma maior cobertura de detecção de *features*, precisão e velocidade de processamento.

Figura 10 - Funcionalidades do *MassCube*.



Fonte: Adaptado de (Fiehn et al., 2025)

3.7.2 MASS-SUITE

O *Mass-Suite* é mais um dos recentes lançamentos de pacotes de código aberto em Python que contemplam análises completas em metabolômica baseada em espectrometria de massas. É um fluxo de trabalho desenvolvido prioritariamente para análises de qualidade de água e outras análises ambientais não direcionadas. Foi desenvolvido como uma ferramenta flexível e personalizável, projetada tanto para cumprir as funções básicas de fluxos de trabalho multifuncionais, como extração de *features,* redução dimensional, anotação, visualização de dados e análises estatísticas como para oferecer capacidades avançadas de mineração de dados e modelos preditivos que não são normalmente oferecidos por ferramentas de código aberto, como de análise por agrupamento e algoritmos de modelagem não supervisionada, funções estas indicadas na Figura 11 (Hu et al., 2023).

Figura 11 - Comparação das funcionalidades do Mass-Suite com as de outras ferramentas disponíveis publicamente.

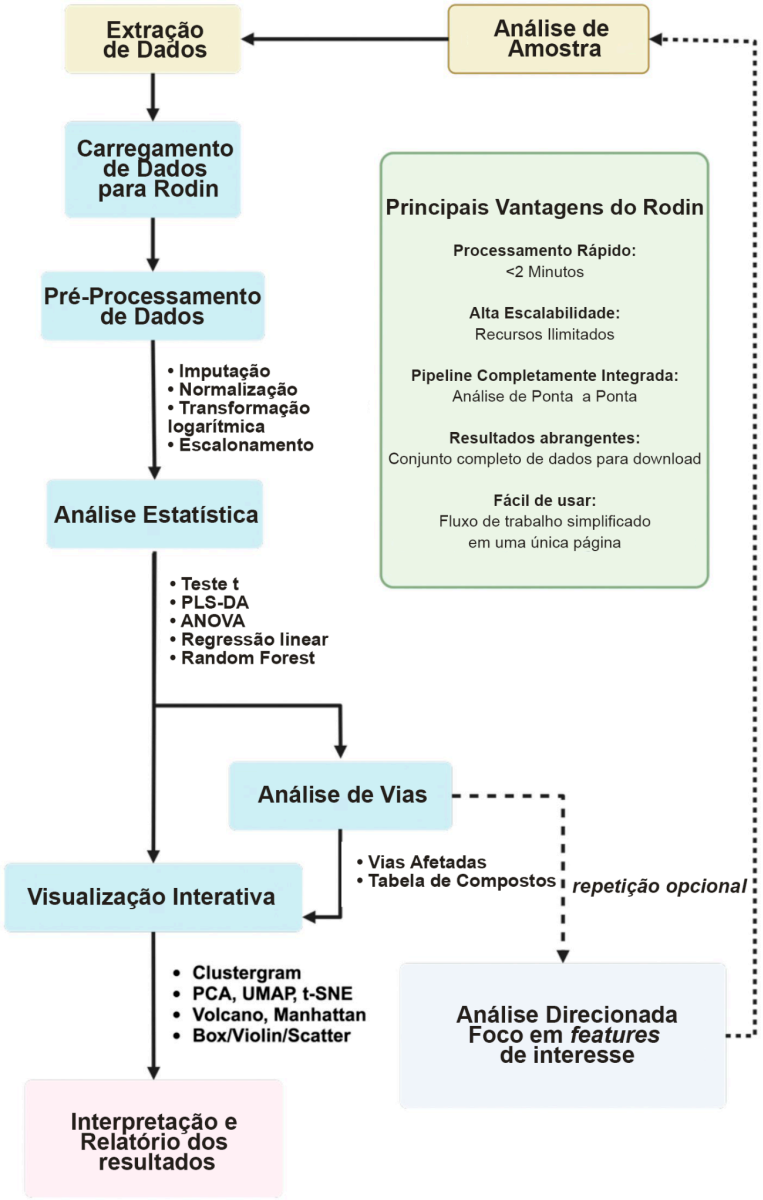
Características	Ferramentas					
	MSS	TidyMS	MZmine2	XCMS*	MSDIAL	PatRoon
Linguagem	Python	Python	Java	R	C#	R
Pré-processamento de dados brutos	√	√	√	√	√	√
Correção de lote baseada em CQ	×	√	×	×	×	×
Relatórios de qualidade	√	√	√	×	√	√
Normalização, imputação, escalonamento	√	√	√	×	√	√
Anotação de features	√	×	√	×	√	√
Agrupamento de isótopos	×	×	√	√	√	√
Gráficos de visualização interativa	√	×	√	×	√	×
Análise estatística de agrupamento	√	×	×	×	×	×
Ferramentas de modelagem	√	×	×	×	×	×

Fonte: Adaptado de (Hu et al., 2023)

3.7.3 RODIN

O *Rodin* possui uma infraestrutura completa para trabalhar com dados de metabolômica não direcionada (Minasenko et al., 2025), desde o pré-processamento usando *softwares* bem estabelecidos como *apLCMS* (T et al., 2009) e *XCMS* (Smith et al., 2006) até análises estatísticas, visualizações diversas e análise por meio de redes, representadas na Figura 12.

Figura 12 - Fluxo de trabalho do pacote Rodin.



Fonte: Adaptado de (Minasenko et al., 2025)

Suas principais características são de ser um fluxo de trabalho rápido e fácil de ser utilizado, além de possuir visualizações interativas e customizáveis em tempo real que mostram todo o conjunto de dados sem ter que efetuar mudanças entre as interfaces, minimizando interrupções no fluxo de trabalho, o que faz com que os usuários foquem na interpretação dos resultados, proporcionando transparência e reprodutibilidade.

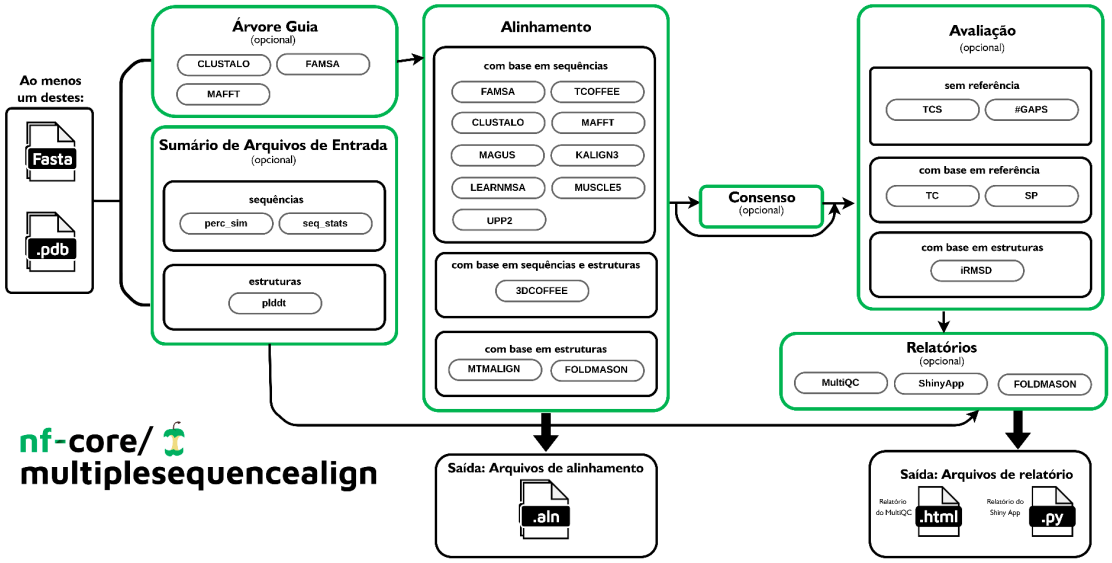
3.7.4 UMETAFLOW

UmetaFlow (Kontou et al., 2023) é um fluxo de trabalho, para o gerenciador *Snakemake* (Köster; Rahmann, 2012), completo para metabolômica não direcionada, unindo vários pacotes em um fluxo contínuo, separado em quatro partes principais: pré-processamento, predições estruturais e de fórmula, exportação para o GNPS (Rede Molecular Social de Produtos Naturais Globais, do inglês *Global Natural Products Social Molecular Networking*) (Wang et al., 2016) e, finalmente, a etapa de pareamento espectral.

3.8. CONSTRUÇÃO DE FLUXOS DE TRABALHO AUTOMATIZADOS

Com o avanço no desenvolvimento de ferramentas para cobrir todas as análises possíveis dentro de experimentos de metabolômica, foi possível pensar na criação de fluxos de trabalho personalizáveis e automatizados, nos quais a saída de uma das ferramentas é automaticamente utilizada como entrada de outra, onde cada etapa pode ser personalizada em vários níveis, inclusive como sendo obrigatórias ou opcionais, como demonstrado na Figura 13, possibilitando o controle de dependências, facilitando o monitoramento de erros e consequente ressubmissão de tarefas e auxiliando a paralelização das análises, entre outros requerimentos frequentemente presentes em análises de bioinformática.

Figura 13 - Exemplo de uma representação visual de fluxo de trabalho automatizado na plataforma *Nextflow*.



Fonte: Adaptado de (Ewels et al., 2020)

Nesse sentido, os orquestradores de fluxo de trabalho, como *Snakemake* (Köster; Rahmann, 2012; Mölder et al., 2021) e *Nextflow* (Di Tommaso et al., 2017), emergem como ferramentas que auxiliam na criação e manutenção desses fluxos. O primeiro foi desenvolvido especificamente para a criação de fluxos de trabalho em bioinformática e possui uma sintaxe muito parecida com o Python. O *Nextflow* é apresentado como uma ferramenta para atender a bioinformatas com experiência em programação e ser uma evolução desse tipo de ferramenta em termos de estabilidade, execução paralela eficiente, tolerância a erros e rastreabilidade.

3.9. OUTROS UTILITÁRIOS PYTHON PARA METABOLÔMICA

3.9.1 JUPYTER NOTEBOOK

O *Jupyter Notebook* é um utilitário que consiste em um formato de documento desenvolvido para a publicação de códigos, resultados e explicações de uma forma que facilite a leitura e a execução. Por meio desses cadernos virtuais é possível armazenar, executar e compartilhar códigos Python com textos explicativos em diferentes formatações (Kluyver et al., 2016). Toda essa facilidade e usabilidade

fizeram com que o *Jupyter Notebook* se tornasse uma ferramenta bastante popular no desenvolvimento, execução e aprendizado para a grande maioria das ciências ômicas. Tal ferramenta, hoje, pode até ser implementada de forma *online* na plataforma Google Colab.

3.9.2 SMITER

O *SMITER* (do inglês *Synthetic mzML writer*) funciona como uma biblioteca Python de linha de comando capaz de simular experimentos de LC-MS/MS e gerar arquivos mzML sintéticos (Kösters; Leufken; Leidel, 2021). Ao operar com base em fórmulas químicas, *SMITER* pode modelar diversas biomoléculas e incorpora modelos customizáveis de ruído e fragmentação, permitindo a criação de conjuntos de dados padrão ouro. Esses dados sintéticos são valiosos para a otimização de parâmetros em algoritmos de processamento e a avaliação de desafios analíticos, como a co-eluição, antes da execução de experimentos laboratoriais complexos e dispendiosos.

4. CONCLUSÃO

Diante do exposto nesta monografia, é possível afirmar que a metabolômica é um campo de estudos em plena expansão, que conta com uma comunidade de desenvolvedores crescente. Em uma área intrinsecamente multidisciplinar, como as demais ciências ômicas, a colaboração de profissionais oriundos das mais diversas áreas se faz muito presente e é de suma importância para o desenvolvimento de ferramentas que além de serem computacionalmente eficientes, e criarem resultados segundo modelos estatísticos adequados, também permitam que os usuários com pouca experiência em computação possam personalizar e aplicar as ferramentas, auxiliando nas potenciais descobertas científicas na área.

BIBLIOGRAFIA

ADUSUMILLI, Ravali; MALLICK, Parag. Data Conversion with ProteoWizard msConvert. *In*: COMAI, Lucio; KATZ, Jonathan E.; MALLICK, Parag (Orgs.). **Proteomics: Methods and Protocols**. New York, NY: Springer, 2017. p. 339–368.

AEIWZ. **aeiwz/metbit**. , 10 jun. 2025. Disponível em: <https://github.com/aeiwz/metbit>. Acesso em: 10 jun. 2025

ANH, Nguyen Ky *et al.* Advancements in Mass Spectrometry-Based Targeted Metabolomics and Lipidomics: Implications for Clinical Research. **Molecules**, v. 29, n. 24, p. 5934, jan. 2024.

ARAÚJO, Ana Margarida *et al.* Metabolomic approaches in the discovery of potential urinary biomarkers of drug-induced liver injury (DILI). **Critical Reviews in Toxicology**, v. 47, n. 8, p. 638–654, 14 set. 2017.

ARINI, Gabriel Santos *et al.* A multi-omics reciprocal analysis for characterization of bacterial metabolism. **Frontiers in Molecular Biosciences**, v. 12, 20 mar. 2025.

BAIDOO, Edward E. K.; TEIXEIRA BENITES, Veronica. Mass Spectrometry-Based Microbial Metabolomics: Techniques, Analysis, and Applications. *In*: BAIDOO, Edward E. K. (Org.). **Microbial Metabolomics: Methods and Protocols**. New York, NY: Springer, 2019. p. 11–69.

BAZZANO, Cristina F. *et al.* NP3 MS Workflow: An Open-Source Software System to Empower Natural Product-Based Drug Discovery Using Untargeted Metabolomics. **Analytical Chemistry**, v. 96, n. 19, p. 7460–7469, 14 maio 2024.

BITTREMIEUX, Wout. spectrum_utils: A Python Package for Mass Spectrometry Data Processing and Visualization. **Analytical Chemistry**, v. 92, n. 1, p. 659–661, 7 jan. 2020.

BITTREMIEUX, Wout *et al.* Unified and Standardized Mass Spectrometry Data Processing in Python Using spectrum_utils. **Journal of Proteome Research**, v. 22, n. 2, p. 625–631, 3 fev. 2023.

BITTREMIEUX, Wout; WANG, Mingxun; DORRESTEIN, Pieter C. The critical role that spectral libraries play in capturing the metabolomics community knowledge. **Metabolomics : Official journal of the Metabolomic Society**, v. 18, n. 12, p. 94, 19 nov. 2022.

CANUTO, Gisele A. B. *et al.* METABOLÔMICA: DEFINIÇÕES, ESTADO-DA-ARTE E APLICAÇÕES REPRESENTATIVAS. **Química Nova**, v. 41, p. 75–91, jan. 2018.

CHAMBERS, Matthew C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. **Nature Biotechnology**, v. 30, n. 10, p. 918–920, out. 2012.

CHOUCHANI, Edward T. *et al.* Ischaemic accumulation of succinate controls reperfusion injury through mitochondrial ROS. **Nature**, v. 515, n. 7527, p. 431–435, nov. 2014.

DAI, Xiaofeng; SHEN, Li. Advances and Trends in Omics Technology Development. **Frontiers in Medicine**, v. 9, p. 911861, 1 jul. 2022.

DE JONGE, Niek F. *et al.* Reproducible MS/MS library cleaning pipeline in matchms. **Journal of Cheminformatics**, v. 16, n. 1, p. 88, 29 jul. 2024.

DETTMER, Katja; ARONOV, Pavel A.; HAMMOCK, Bruce D. Mass spectrometry-based metabolomics. **Mass Spectrometry Reviews**, v. 26, n. 1, p. 51–78, 2007.

DI TOMMASO, Paolo *et al.* Nextflow enables reproducible computational workflows. **Nature Biotechnology**, v. 35, n. 4, p. 316–319, abr. 2017.

EBBELS, Timothy M. D. *et al.* Recent advances in mass spectrometry-based computational metabolomics. **Current Opinion in Chemical Biology**, v. 74, p. 102288, 1 jun. 2023.

EWELS, Philip A. *et al.* The nf-core framework for community-curated bioinformatics pipelines. **Nature Biotechnology**, v. 38, n. 3, p. 276–278, mar. 2020.

FERNIE, Alisdair R.; SCHAUER, Nicolas. Metabolomics-assisted breeding: a viable option for crop improvement? **Trends in Genetics**, v. 25, n. 1, p. 39–48, 1 jan. 2009.

FIEHN, Oliver *et al.* **MassCube: a Python framework for end-to-end metabolomics data processing from raw files to phenotype classifiers**. Research Square, , 7 jan. 2025. Disponível em: <https://www.researchsquare.com/article/rs-5530740/v1>. Acesso em: 3 jun. 2025

GONZÁLEZ-DOMÍNGUEZ, Raúl *et al.* High-Throughput Metabolomics Based on Direct Mass Spectrometry Analysis in Biomedical Research. *In*: D'ALESSANDRO, Angelo (Org.). **High-Throughput Metabolomics: Methods and Protocols**. New York, NY: Springer, 2019. p. 27–38.

HAGBERG, Aric; SWART, Pieter J.; SCHULT, Daniel A. **Exploring network structure, dynamics, and function using NetworkX**. *[S.l.]*: Los Alamos National Laboratory (LANL), Los Alamos, NM (United States), 1 jan. 2008. Disponível em: <https://www.osti.gov/biblio/960616>. Acesso em: 27 maio. 2025.

HASIN, Yehudit; SELDIN, Marcus; LUSIS, Aldons. Multi-omics approaches to disease. **Genome Biology**, v. 18, n. 1, p. 83, 5 maio 2017.

HO, CS *et al.* Electrospray Ionisation Mass Spectrometry: Principles and Clinical Applications. **The Clinical Biochemist Reviews**, v. 24, n. 1, p. 3–12, fev. 2003.

HU, Ximin *et al.* Mass-Suite: a novel open-source python package for high-resolution mass spectrometry data analysis. **Journal of Cheminformatics**, v. 15, n. 1, p. 87, 23 set. 2023.

HUBER, Florian *et al.* matchms - processing and similarity evaluation of mass spectrometry data. **Journal of Open Source Software**, v. 5, n. 52, p. 2411, 31 ago. 2020.

HUBER, Florian *et al.* Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. **PLOS Computational Biology**, v. 17, n. 2, p. e1008724, 16 fev. 2021a.

HUBER, Florian *et al.* MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. **Journal of Cheminformatics**, v. 13, n. 1, p. 84, 29 out. 2021b.

JACOB, Minnie *et al.* Metabolomics toward personalized medicine. **Mass Spectrometry Reviews**, v. 38, n. 3, p. 221–238, 2019.

KATAJAMAA, Mikko; OREŠIČ, Matej. Data processing for mass spectrometry-based metabolomics. **Journal of Chromatography A**, Data Analysis in Chromatography. v. 1158,

n. 1, p. 318–328, 27 jul. 2007.

KLUYVER, Thomas *et al.* Jupyter Notebooks – a publishing format for reproducible computational workflows. *In*: **Positioning and Power in Academic Publishing: Players, Agents and Agendas**. *[S.l.]*: IOS Press, 2016. p. 87–90.

KONTOU, Eftychia E. *et al.* UmetaFlow: an untargeted metabolomics workflow for high-throughput data processing and analysis. **Journal of Cheminformatics**, v. 15, p. 52, 12 maio 2023.

KÖSTER, Johannes; RAHMANN, Sven. Snakemake—a scalable bioinformatics workflow engine. **Bioinformatics**, v. 28, n. 19, p. 2520–2522, 1 out. 2012.

KÖSTERS, Manuel; LEUFKEN, Johannes; LEIDEL, Sebastian A. SMITER—A Python Library for the Simulation of LC-MS/MS Experiments. **Genes**, v. 12, n. 3, p. 396, mar. 2021.

KRUK, Joanna *et al.* NMR Techniques in Metabolomic Studies: A Quick Overview on Examples of Utilization. **Applied Magnetic Resonance**, v. 48, n. 1, p. 1–21, 2017.

LEIPZIG, Jeremy. A review of bioinformatic pipeline frameworks. **Briefings in Bioinformatics**, v. 18, n. 3, p. 530–536, 1 maio 2017.

LI, Shuzhao *et al.* Trackable and scalable LC-MS metabolomics data processing using asari. **Nature Communications**, v. 14, n. 1, p. 4113, 11 jul. 2023.

MARKLEY, John L. *et al.* The future of NMR-based metabolomics. **Current Opinion in Biotechnology**, Analytical biotechnology. v. 43, p. 34–40, 1 fev. 2017.

MCKINNEY, Wes. Data Structures for Statistical Computing in Python. **scipy**, 1 maio 2010.

MINASENKO, Boris *et al.* Rodin: a streamlined metabolomics data analysis and visualization tool. **Bioinformatics Advances**, v. 5, n. 1, p. vbaf088, 1 jan. 2025.

MÖLDER, Felix *et al.* **Sustainable data analysis with Snakemake**. F1000Research, , 18 jan. 2021. Disponível em: <https://f1000research.com/articles/10-33>. Acesso em: 3 jun. 2025

MYLONAS, Roman *et al.* X-Rank: A Robust Algorithm for Small Molecule Identification Using Tandem Mass Spectrometry. **Analytical Chemistry**, v. 81, n. 18, p. 7604–7610, 15 set. 2009.

NASCIMENTO, Ronaldo Ferreira do *et al.* Cromatografia gasosa: aspectos teóricos e práticos. 2018.

NEWMAN, David J.; CRAGG, Gordon M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. **Journal of Natural Products**, v. 83, n. 3, p. 770–803, 27 mar. 2020.

O'CALLAGHAN, Sean *et al.* PyMS: a Python toolkit for processing of gas chromatography-mass spectrometry (GC-MS) data. Application and comparative study of selected tools. **BMC Bioinformatics**, v. 13, n. 1, p. 115, 30 maio 2012.

PEREZ-RIVEROL, Yasset *et al.* Quantifying the impact of public omics data. **Nature Communications**, v. 10, n. 1, p. 3512, 5 ago. 2019.

QUACKENBUSH, John. Data standards for “omic” science. **Nature Biotechnology**, v. 22, n. 5, p. 613–614, maio 2004.

QUINN, Robert A. *et al.* Molecular Networking As a Drug Discovery, Drug Metabolism, and Precision Medicine Strategy. **Trends in Pharmacological Sciences**, v. 38, n. 2, p. 143–154, fev. 2017.

RAKUSANOVA, Stanislava; FIEHN, Oliver; CAJKA, Tomas. Toward building mass spectrometry-based metabolomics and lipidomics atlases for biological and clinical research. **TrAC Trends in Analytical Chemistry**, v. 158, p. 116825, 1 jan. 2023.

RIQUELME, Gabriel *et al.* A Python-Based Pipeline for Preprocessing LC–MS Data for Untargeted Metabolomics Workflows. **Metabolites**, v. 10, n. 10, p. 416, out. 2020.

RÖST, Hannes L. *et al.* pyOpenMS: A Python-based interface to the OpenMS mass-spectrometry algorithm library. **PROTEOMICS**, v. 14, n. 1, p. 74–77, 2014.

RUTTKIES, Christoph *et al.* MetFrag relaunched: incorporating strategies beyond in silico fragmentation. **Journal of Cheminformatics**, v. 8, n. 1, p. 3, 29 jan. 2016.

SCHMID, Robin *et al.* Integrative analysis of multimodal mass spectrometry data in MZmine 3. **Nature Biotechnology**, v. 41, n. 4, p. 447–449, abr. 2023.

SCHYMANSKI, Emma L. *et al.* Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. **Environmental Science & Technology**, v. 48, n. 4, p. 2097–2098, 18 fev. 2014.

SILVA, Ricardo R. da *et al.* Propagating annotations of molecular networks using in silico fragmentation. **PLOS Computational Biology**, v. 14, n. 4, p. e1006089, 18 abr. 2018.

SING, Justin Cyril *et al.* pyOpenMS-viz: Streamlining Mass Spectrometry Data Visualization with pandas. **Journal of Proteome Research**, v. 24, n. 4, p. 2152–2158, 4 abr. 2025.

SMITH, Colin A. *et al.* XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. **Analytical Chemistry**, v. 78, n. 3, p. 779–787, 1 fev. 2006.

STAFF, GitHub. **Octoverse: AI leads Python to top language as the number of global developers surges**. **The GitHub Blog**, 29 out. 2024. Disponível em: <https://github.blog/news-insights/octoverse/octoverse-2024/>. Acesso em: 11 maio. 2025

T, Yu *et al.* apLCMS--adaptive processing of high-resolution LC/MS data. **Bioinformatics (Oxford, England)**, v. 25, n. 15, 8 jan. 2009.

TAUTENHAHN, Ralf; BÖTTCHER, Christoph; NEUMANN, Steffen. Highly sensitive feature detection for high resolution LC/MS. **BMC Bioinformatics**, v. 9, n. 1, p. 504, 28 nov. 2008.

TSUGAWA, Hiroshi *et al.* MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. **Nature Methods**, v. 12, n. 6, p. 523–526, jun. 2015.

VAILATI-RIBONI, Mario; PALOMBO, Valentino; LOOR, Juan J. What Are Omics Sciences? *In*: AMETAJ, Burim N. (Org.). **Periparturient Diseases of Dairy Cows: A Systems Biology Approach**. Cham: Springer International Publishing, 2017. p. 1–7.

VINAIXA, Maria *et al.* Mass spectral databases for LC/MS- and GC/MS-based metabolomics:

State of the field and future prospects. **TrAC Trends in Analytical Chemistry**, v. 78, p. 23–35, 1 abr. 2016.

WANG, Mingxun *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. **Nature Biotechnology**, v. 34, n. 8, p. 828–837, ago. 2016.

WATROUS, Jeramie *et al.* Mass spectral molecular networking of living microbial colonies. **Proceedings of the National Academy of Sciences**, v. 109, n. 26, p. E1743–E1752, 26 jun. 2012.

WEHRENS, Ron; SALEK, Reza. **Metabolomics: Practical Guide to Design and Analysis**. 1. ed. Boca Raton, Florida: Chapman and Hall/CRC, 2019. v. 1

WENIG, Philip; ODERMATT, Juergen. OpenChrom: a cross-platform open source software for the mass spectrometric analysis of chromatographic data. **BMC Bioinformatics**, v. 11, n. 1, p. 405, 30 jul. 2010.

WISHART, David S. Emerging applications of metabolomics in drug discovery and precision medicine. **Nature Reviews Drug Discovery**, v. 15, n. 7, p. 473–484, jul. 2016.

WISHART, David S. *et al.* HMDB 5.0: the Human Metabolome Database for 2022. **Nucleic Acids Research**, v. 50, n. D1, p. D622–D631, 19 nov. 2021.