Universidade Federal do Rio de Janeiro

Eduardo De Souza Matos

METODOLOGIA DE PROCESSAMENTO DE DADOS E ANÁLISE ESTATÍSTICA PARA ESTUDO METABOLÔMICO GLOBAL POR CROMATOGRAFIA GASOSA ACOPLADA A ESPECTROMETRIA DE MASSAS

RIO DE JANEIRO

2024

Eduardo de Souza Matos

METODOLOGIA DE PROCESSAMENTO DE DADOS E ANÁLISE ESTATÍSTICA PARA ESTUDO METABOLÔMICO GLOBAL POR CROMATOGRAFIA GASOSA ACOPLADA A ESPECTROMETRIA DE MASSAS

Volume único.

Dissertação de Mestrado apresentada ao Programa de
Mestrado Profissional de Formação Para a Pesquisa Biomédica,
Instituto de Biofísica Carlos Chagas Filho,
Universidade Federal do Rio de Janeiro,
como requisito parcial à obtenção do título de
Mestre em Ciências Biológicas.

Orientadora: Profa. Dra. Graciela Maria Dias

Coorientadora: Profa. Dra. Virgínia Martins Carvalho

RIO DE JANEIRO

2024

CIP - Catalogação na Publicação

d433m

de Souza Matos, Eduardo
METODOLOGIA DE PROCESSAMENTO DE DADOS E ANÁLISE ESTATÍSTICA PARA ESTUDO METABOLÔMICO GLOBAL POR CROMATOGRAFIA GASOSA ACOPLADA A ESPECTROMETRIA DE MASSAS / Eduardo de Souza Matos. -- Rio de Janeiro, 2024.
70 f.

Orientadora: Graciela Maria Dias.
Coorientadora: Virgínia Martins Carvalho.
Dissertação (mestrado) - Universidade Federal do Rio de Janeiro, Instituto de Biofísica Carlos Chagas Filho, Programa de Mestrado Profissional em Formação para a Pesquisa Biomédica, 2024.

1. metabolômica global. 2. espectrometria de massas. 3. cromatografia gasosa. 4. cannabis. 5. processamento de dados. I. Dias, Graciela Maria, orient. II. Martins Carvalho, Virgínia, coorient. III. Título.

Elaborado pelo Sistema de Geração Automática da UFRJ com os dados fornecidos pelo(a) autor(a), sob a responsabilidade de Miguel Romeu Amorim Neto - CRB-7/6283.

"METODOLOGIA DE PROCESSAMENTO DE DADOS E ANÁLISE ESTATÍSTICA PARA ESTUDO METABOLÔMICO GLOBAL POR CROMATOGRAFIA GASOSA ACOPLADA A ESPECTROMETRIA DE MASSAS"

EDUARDO DE SOUZA MATOS

DISSERTAÇÃO DE MESTRADO PROFISSIONAL DE FORMAÇÃO PARA A PESQUISA BIOMÉDICA SUBMETIDA À UNIVERSIDADE FEDERAL DO RIO DE JANEIRO VISANDO A OBTENÇÃO DO GRAU DE MESTRE EM FORMAÇÃO PARA A PESQUISA BIOMÉDICA.

APROVADA POR:

Rio de Janeiro, 26 de agosto de 2024.

DRA. FLAVIA FONSECA BLOISE (DOUTORA – UFRJ)
(COORDENADORA DO CURSO DE MESTRADO PROFISSIONAL DE FORMAÇÃO PARA PESQUISA BIOMÉDICA)

VÍDEOCONFERÊNCIA

DRA. GRACIELA MARIA DIAS (DOUTORA – UFRJ) – ORIENTADORA

DRA. VIRGÍNIA MARTINS CARVALHO (DOUTORA – UFRJ) – COORIENTADORA

DRA. ANA BEATRIZ FURLANETTO PACHECO (DOUTORA - UFRJ)

VÍDEOCONFERÊNCIA

DR. ANTONIO JORGE RIBEIRO DA SILVA - (DOUTOR – UFRJ)

DRA. PATRICIA SOSA ACOSTA (DOUTOR – UFRJ)

DRA. NAOMI KATO SIMAS (DOUTORA – UFRJ) - REVISORA

RESUMO

A análise do perfil metabólico por cromatografia acoplada a espectrometria de massas requer posterior processamento dos dados adquiridos na análise instrumental. Tal processamento envolve um fluxo de trabalho composto pela aplicação de diversos algoritmos de forma sequencial. A complexidade envolvida no racional desse processo é uma barreira de entrada de pesquisadores iniciantes na análise metabolômica, bem como a restrição do uso de dados em softwares *fechados* pelo fabricante, com exigência de uma licença comercial. Além disso, a configuração de parâmetros envolvidos no processamento de dados analíticos e do tratamento pré-estatístico causam grande impacto na qualidade do resultado final, o que pode levar a conclusões errôneas acerca do estudo realizado. Nesse contexto, o estudo em questão teve como objetivo principal o desenvolvimento de um fluxo de trabalho para análise de dados de Cromatografia Gasosa acoplada a Espectrometria de Massas (CG-EM) em um estudo metabolômico global de dois grupos de *Cannabis sp.* de diferentes procedências, utilizando o *software* de código livre Mzmine em conjunto com a plataforma online e gratuita de análise estatística MetaboAnalyst. Com a metodologia desenvolvida, foi possível detectar 22 diferentes compostos, sendo 20 desses anotados com provável identidade química. Avaliando-se diferentes abordagens de normalização e se aplicando ferramentas estatísticas univariadas e multivariadas, foi possível distinguir os dois grupos estudados com base em seu perfil químico de fitocanabinóides e terpenos, além de distinguir dois subgrupos de cannabis provenientes de contexto medicinal. A metodologia descrita de forma minuciosa tende a auxiliar pesquisadores na área de metabolômica na análise de dados de CG-EM de diferentes estudos a partir de dados abertos adquiridos em equipamento próprio ou em repositórios, permitindo a avaliação e comparação dos dados existentes.

Palavras-chave: metabolômica global, espectrometria de massas, cromatografia gasosa, cannabis, processamento de dados.

ABSTRACT

The analysis of the metabolic profile by mass spectrometry coupled with chromatography requires subsequent processing of data acquired from instrumental analysis. This processing involves a workflow composed of the sequential application of various algorithms. The complexity involved in the rationale of this workflow is a barrier to entry for new researchers in metabolomic analysis, as well as the restriction of data use in proprietary software requiring a commercial license. Moreover, the configuration of parameters involved in the processing of analytical data and pre-statistical treatment significantly impacts the quality of the results, potentially leading to erroneous conclusions about the conducted study. In this context, the main objective of the work is to develop a workflow for analyzing Gas Chromatography-Mass Spectrometry (GC-MS) data in a global metabolomic study of two groups of *Cannabis sp.* from different origins, using the open-source software Mzmine in conjunction with the free online statistical analysis platform MetaboAnalyst. With the developed methodology, it was possible to detect 22 different compounds, 20 of which were annotated with probable chemical identities. By evaluating different normalization approaches and applying univariate and multivariate statistical tools, it was possible to distinguish the two studied groups based on their chemical profiles of phytocannabinoids and terpenes, as well as to distinguish two subgroups of cannabis from a medicinal context. The meticulously described methodology aims to assist researchers in the metabolomics field in analyzing GC-MS data from different studies using open access data acquired from their own equipment or data repositories, allowing for the evaluation and comparison of existing data.

Keywords: untargeted metabolomics, mass spectrometry, gas chromatography, cannabis, data processing.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

CBCA	ácido canabicromênico
CBD	canabidiol
CBDA	ácido canabidiólico
CBG	canabigerol
CBGA	ácido canabigerólico
CBN	canabinol
CG-EM	cromatografia gasosa acoplada à espectrometria de massas
CLAE-DAD	cromatografia líquida de alta eficiência acoplada a detector de arranjo de diodos
m/z	relação massa/carga
NIST	*National Institute of Standards and Technology*
ONG	organização não-governamental
PCA	análise de componentes principais
PLS-DA	analise discriminante por mínimos quadrados parciais
THC	delta-9 tetraidrocanabinol
THCA	ácido delta-9 tetraidrocanabinólico
UFRJ	Universidade Federal do Rio de Janeiro
v/v	proporção volume por volume

SUMÁRIO

1) INTRODUÇÃO

A metabolômica é uma ciência interdisciplinar que se propõe a estudar o perfil metabólico de amostras de origem biológica em níveis qualitativo e/ou quantitativo. Frente à complexidade inerente à composição metabólica, são empregadas técnicas analíticas de alto rendimento em conjunto com ferramentas bioinformáticas e estatísticas a fim de se obter informações que caracterizem e/ou expliquem o objeto de estudo (WISHART, 2016).

Um dos pontos cruciais do estudo metabolômico é quanto à confiabilidade dos dados obtidos após análise, tendo em vista a grande variabilidade das metodologias analíticas executadas entre diferentes laboratórios (NISHIUMI et al., 2022). Aspectos que impactam na qualidade da interpretação dos resultados, após a análise instrumental, são o processamento dos dados brutos, a anotação criteriosa de identidade dos compostos detectados, o tratamento e formatação dos dados e a execução de ferramentas estatísticas adequadas junto com sua correta interpretação.

Enquanto tais práticas são bem definidas em estudos de metabolômica alvo, cujos compostos de interesse são definidos previamente, o estabelecimento daquelas em estudos de metabolômica global é desafiante. Devido à sua natureza exploratória que suporta a geração de hipóteses sobre o assunto estudado, é incompatível o emprego de práticas que exijam a caracterização prévia de todos os compostos de uma amostra ainda desconhecida. Assim, há um grande esforço da comunidade científica para o estabelecimento de práticas que abranjam as especificidades de tal abordagem (EVANS et al., 2020). Esse esforço se justifica pelo potencial que essa possui em estudos de caracterização de compostos desconhecidos em amostras de origens diversas, além de suportar estudos de larga escala para elucidação mecanística de fenômenos biológicos e na identificação de biomarcadores de doenças.

Nesse contexto, a determinação de um fluxo de trabalho racional básico de processamento de dados utilizando *softwares* de código aberto e plataformas gratuitas de análise permite ao pesquisador compreender o processo como um todo e o impacto de cada etapa no resultado final. Desse modo, o pesquisador pode optar por diferentes abordagens e ponderar os prós e contras de cada decisão ao longo do processo. Além disso, o pesquisador que domina os princípios do processamento de dados de cromatografia e espectrometria de massas é capaz de realizar o escrutínio de informações depositadas em reservatórios de dados e realizar estudos de meta-análise.

2) OBJETIVOS

2.1) Objetivo Geral:

- Estabelecer uma metodologia de tratamento de dados e análise estatística em um estudo de metabolômica global baseada em Cromatografia Gasosa acoplada à Espectrometria de Massas (CG-EM).

2.2) Objetivos Específicos:

- Definir procedimento geral da análise a partir dos dados brutos obtidos da análise instrumental de amostras de cannabis cultivadas para fins medicinais e apreendidas pela Polícia Civil do Rio de Janeiro;
- Avaliar a qualidade e a quantidade das anotações dos *features* detectados utilizando dois bancos de dados de espectros distintos;
- Realizar o tratamento de dados pós-processamento em *software* para análise estatística na plataforma *online* MetaboAnalyst;
- Avaliar dados do perfil químico obtido dos dois tipos de amostras de cannabis de duas procedências.

3) REVISÃO BIBLIOGRÁFICA

3.1) Metabolômica

A Metabolômica é uma área das ciências da natureza que visa a análise para compreensão do perfil metabólico de sistemas biológicos, de formas qualitativa e quantitativa. O interesse em estudar metabólitos nas ciências biológicas é visto desde a década de 40 (WILLIAMS; KIRBY, 1948). Entretanto, o termo "Metaboloma" veio a ser utilizado pela primeira vez por Stephen Oliver em 1998. Sob influência das ciências Ômicas já estabelecidas, como a genômica e a proteômica, o termo foi aplicado a estudos iniciais do metabolismo de organismos unicelulares como *Escherichia coli* e *Saccharomyces cerevisiae* (FERNIE; SCHAUER, 2009).

Metaboloma é o conteúdo de pequenas moléculas, de massa molecular até cerca de 1500 *Da*, presente em materiais biológicos como plasma, urina, saliva, fezes, suor, cultivos celulares e tecidos, entre outros (HOLMES; WILSON; NICHOLSON, 2008). Classes de metabólitos, como sacarídeos, ácidos orgânicos, nucleotídeos, lipídios e aminoácidos, estão incluídas nesse conjunto. Porém, o perfil metabólico não se limita a elas, vista a riqueza de metabólitos secundários produzidos por organismos como plantas e fungos e dos produtos de biotransformação de compostos exógenos, como fármacos (WISHART, 2019). As classes apresentadas diferem substancialmente em suas características físico-químicas, constituindo um desafio analítico a detecção e quantificação das mesmas em uma única abordagem. A tamanha complexidade envolvida é exemplificada pelo Banco de Dados do Metaboloma Humano (sigla em inglês HMDB), que possui mais de 200 mil entradas de metabólitos anotados (WISHART *et al*., 2022).

Desde então, interesse da comunidade científica no perfil metabolômico se fundamenta na premissa de que o metabolismo de um organismo sofre influência de aspectos genéticos, epigenéticos e ambientais, sendo o metaboloma o reflexo do fluxo de vias bioquímicas e de possíveis perturbações em elementos participantes desses sistemas (CAMPBELL; XIA; NIELSEN, 2017; JOHNSON; IVANISEVIC; SIUZDAK, 2016). Assim, o perfil metabólico e sua dinâmica são objetos de estudo nas ciências da saúde na busca pela elucidação dos mecanismos de processos fisiológicos e fisiopatológicos e por biomarcadores para diagnóstico ou prognóstico na clínica (BECKER et al., 2012). Com tais aplicações, a metabolômica é utilizada proeminentemente em áreas como oncologia (KOWALCZYK *et al*., 2020), cardiologia (XIE *et al*., 2023), endocrinologia (ABU-FARHA *et al*., 2024) e toxicologia (LOCCI et al., 2020), entre outras.

Atualmente, o poder analítico de equipamentos utilizados para análise metabolômica vem sendo adaptado para uso *in loco* em pacientes para procedimentos de diagnóstico durante biópsias e cirurgias. Essa abordagem permite obter informações clínicas de grande importância, por exemplo, a distinção entre tecidos sadios e doentes em tempo real com especificidade e praticidade, aumentando o grau de sucesso frente a análises convencionais. Tais técnicas permitem esse feito a partir de análises espectrométricas realizadas *ex vivo* e *in vivo* associadas a ferramentas para a classificação baseada em algoritmos de aprendizado de máquina (m*achine learning*) (PROCOPIO *et al*., 2023).

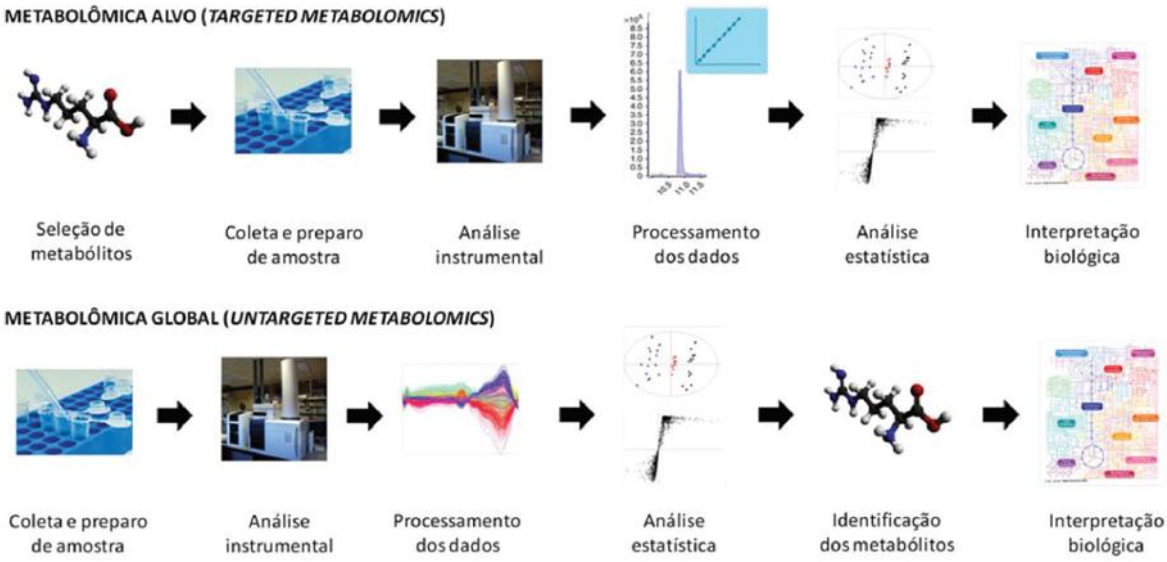
O estudo metabolômico também é realizado integrado a outras ciências ômicas, como a genômica, a transcriptômica e a proteômica, compondo o estudo da biologia de sistemas. Tal abordagem surge do conceito de que processos biológicos não são advindos de uma alteração pontual de um desses níveis de funcionamento, numa compreensão reducionista do fenômeno, mas sim da interação dinâmica entre os mesmos, trazendo uma visão holística do sistema biológico de estudo (MEDIANI; BAHARUM, 2024; PROCOPIO *et al*., 2023). A abordagem integrada é um importante fator para a obtenção de assinaturas biológicas para diagnóstico e prognóstico personalizados para uso clínico em hospitais em medicina de precisão (MUSSAP *et al*., 2021).

Devido à sua complexidade química, o estudo do metabolismo de uma forma ampla só foi possível com o desenvolvimento e aplicação de técnicas analíticas instrumentais de alto rendimento. Uma dessas é a cromatografia acoplada a espectrometria de massas, que permite a identificação e a quantificação da ordem de dezenas até milhares de compostos específicos de uma amostra complexa em uma única análise (DUNN; ELLIS, 2005).

Existem duas abordagens principais em estudos metabolômicos (Figura 1): a metabolômica alvo (no inglês, *targeted*) e a metabolômica global (no inglês, *untargeted*) (CANUTO *et al*., 2018). A primeira visa a quantificação absoluta de metabólitos específicos, geralmente utilizando padrões de pureza analítica como referência das características dos compostos de interesse. Tais características são o tempo de retenção, relação massa/carga, perfil de fragmentação e a relação entre a resposta medida do detector e a concentração dos analitos dentro de uma faixa de concentração determinada. Tal abordagem conta com diversos guias de boas práticas e de validação metodológica, sendo aplicada como ferramenta para validação de hipóteses em pesquisas científicas com

base na quantificação absoluta de metabólitos estritos. (EUROPEAN MEDICINES AGENCY ICH, 2005; INMETRO, 2020).

Figura 1:Etapas de análise metabolômica dos tipos alvo e global. Retirado de CANUTO *et al*., 2018.



A segunda abordagem tem como objetivo as análises qualitativa e quantitativa de metabólitos detectados sem conhecimento prévio de suas características analíticas a partir de padrões de referência. Essa abordagem tem como finalidade exploratória e aplicada para a geração de hipóteses em estudos científicos. Sob o aspecto qualitativo, a anotação de possíveis identidades é feita a partir das características intrínsecas dos compostos frente a metodologia usada e conhecimento presumível do tipo de amostra analisada (KIND; FIEHN, 2007). Em termos quantitativos, valores medidos de área ou altura do pico cromatográfico de todos os analitos detectáveis são usados em análises estatísticas a fim de encontrar um ou mais metabólitos correlacionados com o fenômeno estudado (VIANT et al., 2019).

Uma abordagem mista pode ser aplicada em casos onde a metodologia analítica permite o uso de padrões de referência dos compostos conhecidos junto com a detecção de metabólitos desconhecidos na mesma análise, sendo cunhada como abordagem semi-alvo (do inglês *semitargeted).* Entretanto, tal metodologia depende de uma capacidade maior de distinção entre os sinais de compostos similares, como a obtida por uso de cromatografia líquida de ultra alta eficiência e espectrômetros de massas de alta resolução (ZHOU; YIN, 2016).

A abordagem global, devido ao fato de se desconhecer quais compostos estão presentes na amostra *a priori*, não se beneficia da aplicação dos guias existentes para a abordagem alvo. Assim, após esforço no levantamento das formas de controle e garantia de qualidade por diferentes grupos de pesquisa que utilizam análises metabolômicas, foram definidas algumas práticas no que concerne o planejamento experimental, a fim de se obter maior confiança nos resultados qualitativo e quantitativo (MOSLEY *et al*., 2024; VIANT *et al*., 2019).

A análise metabolômica pode ser dividida nas seguintes etapas: preparação de amostra, análise instrumental, processamento dos dados brutos, análise estatística, anotação de identidade dos metabólitos e interpretação biológica. No caso da análise metabolômica alvo, a identidade dos metabólitos de interesse já é conhecida e caracterizada pelo uso de padrões analíticos referentes aos mesmos (CANUTO *et al*., 2018).

Como primeira etapa, a preparação da amostra visa a extração dos metabólitos de interesse da amostra biológica. Ao se escolher o método de extração, é importante levar em consideração a natureza físico-química dos compostos de interesse, visto que características como hidrofobicidade, carga elétrica e volatilidade variam entre as distintas classes químicas presentes. Dessa forma, não existe um método de extração universal, o que limita o pesquisador a escolher aquele onde se obtém o melhor rendimento de extração de uma parte do metaboloma da amostra. Vale destacar que o método de extração deve cessar, idealmente, toda a atividade enzimática na amostra, mantendo estável o perfil dos metabólitos ao longo do processo. Esse processo é denominado em inglês de *quenching* e pode ser realizado com a adição de solventes orgânicos à amostra ou seu congelamento rápido com uso de nitrogênio líquido. Dessa forma, as proteínas presentes são desnaturadas, reduzindo a atividade enzimática (ALSEEKH *et al*., 2021).

3.2) Espectrometria de massas

A segunda etapa da análise metabolômica é relacionada à instrumentação analítica usada na aquisição dos dados. As técnicas mais utilizadas na análise metabolômica são a espectrometria de massas acoplada à cromatografia e a ressonância magnética nuclear. Ambas são utilizadas por serem técnicas de alto rendimento, ou seja, são capazes de detectar, identificar e quantificar de dezenas a milhares de compostos em uma única análise. A ressonância magnética nuclear, embora muito utilizada nessa área, carece da

mesma sensibilidade oferecida pela espectrometria de massas, sendo a última aplicada na maior parte dos estudos de moléculas de diferentes naturezas químicas (DUNN; ELLIS, 2005).

A espectrometria de massas é uma técnica analítica que mede a relação massa/carga dos analitos ionizados presentes na amostra. Para isso, a análise espectrométrica ocorre em três etapas distintas: a ionização/dessorção dos analitos em íons gasosos; a separação dos íons de acordo com a razão entre sua massa molecular e a carga elétrica obtida; e a detecção dos íons em sinais proporcionais à quantidade de íons detectadas. As duas primeiras etapas possuem maior impacto no desempenho do espectrômetro na análise de diferentes moléculas ou amostras (GROSS, 2004).

A primeira etapa é realizada por uma parte do espectrômetro de massas chamada de fonte de ionização. De acordo com o seu mecanismo de funcionamento, ela apresenta maior ou menor eficiência nesse processo. Para analitos com maior volatilidade, a fonte de ionização por elétrons é utilizada e tem como mecanismo a ionização enérgica de analitos gasosos, com a maior formação de íons radicalares (com a remoção de um elétron de pares eletrônicos não-ligantes ou de ligações químicas das moléculas). Esse processo provoca a dissociação parcial dos íons em fragmentos relacionados à estrutura química da molécula analisada. Para analitos de menor volatilidade e de natureza polar a moderadamente apolar, a fonte de ionização por *eletrospray* é usada, com formação de íons protonados ou desprotonados (positivamente ou negativamente carregados, respectivamente) ou íons formados por adutos. Tal ionização ocorre de forma energeticamente branda e é menos suscetível a gerar dissociação (GROSS, 2004).

A segunda etapa é realizada pelo analisador de massas, que separa os íons com uso de campos elétricos e/ou magnéticos e, por fim, determina a relação massa/carga (m/z) dos mesmos. A separação e medida de m/z podem ser feitas com menores resolução de massas e acurácia, de forma a não poder distinguir compostos de massas moleculares próximas, obtendo medidas de m/z de baixa acurácia (geralmente com erro unitário), sendo o quadrupolo um exemplo desse tipo de analisador de massas. Alternativamente, a separação de íons e a medida do valor de m/z podem ser feitas com altas resolução e acurácia, sendo possível separar íons com massas muito próximas (com menos de 5 partes por milhão ou ppm de diferença de massa), tendo como exemplos de analisadores o tempo-de-vôo e o orbitrap (GROSS, 2004).

Embora a espectrometria de massas tenha desempenho satisfatório se usada individualmente para a análise de amostras complexas, ela possui duas desvantagens que

reduzem sua capacidade analítica: uma é a impossibilidade de resolução de analitos isoméricos constitucionais; a segunda é a perda de sensibilidade na detecção relativa ao efeito de supressão da ionização por *eletrospray* causada por sais e alguns compostos (KIRWAN *et al*., 2014). Assim, seu acoplamento com as cromatografias líquida ou gasosas e a possibilidade de uma análise ortogonal à espectrométrica constituem o estado da arte na análise metabolômica (CHACKO; HASEEB; HASEEB, 2021).

3.3) Processamento de dados

O processamento dos dados adquiridos pela análise da cromatografia acoplada à espectrometria de massas é um ponto crítico para a interpretação biológica de uma análise metabolômica. As empresas fabricantes dos equipamentos geralmente dispõem de *softwares* e fluxos de trabalho que utilizam como arquivos de entrada aqueles com formato e extensão de arquivo próprios. Considerando que tal situação implica na aquisição comercial do *software* próprio do fabricante, isso é um fator limitante para a análise de dados por outros pesquisadores que não possuem tais *softwares* ou suas licenças comerciais de uso. Como consequência, tal situação prejudica iniciativas relacionadas à comparação de dados interlaboratoriais ou estudos de meta-análise (LIOTTA *et al*., 2005).

Uma das iniciativas para o uso aberto e gratuito de dados analíticos é a conversão dos dados para um formato aberto. A geração de arquivos de dado aberto, como os de extensão *mzxml* (PEDRIOLI *et al*., 2004) e *mzml* (MARTENS *et al*., 2011), foi adotada por algumas empresas fabricantes de espectrômetros de massas a fim de universalizar o acesso e análise dos dados pela comunidade científica. Somado a isso, ferramentas de conversão de dados específicos da fabricante para formatos de dado aberto são utilizadas em casos onde só há disponíveis os arquivos brutos contendo os dados fechados (CHAMBERS *et al*., 2012).

Em conjunto com essa iniciativa, o desenvolvimento e disponibilidade de *softwares* gratuitos e de códigos fonte abertos permitem realizar o processamento dos dados analíticos de formato genérico e aberto. Dessa forma, é possível obter características como o valor quantitativo de compostos detectados no método (como a intensidade máxima do pico cromatográfico ou sua área) e que contribuem na elucidação da identidade dos compostos *a posteriori* (como tempo de retenção, valor da relação massa/carga e perfil de fragmentação) (CHEN; LI; XU, 2022).

Atualmente, existe uma gama de *softwares* livres para análise de dados espectrométricos disponíveis. Há grande interesse da comunidade científica em apoiar o uso de tais ferramentas, com a oferta de treinamentos, desenvolvimento de fluxos de trabalho e suporte para os desenvolvedores (CHANG *et al*., 2021). Dentre os *softwares* mais utilizados, considerando o número de citações dos artigos de apresentação, estão o Mzmine (PLUSKAL *et al*., 2010), o XCMS (XU *et al*., 2019) e o MSDIAL (TSUGAWA *et al*., 2015).

3.4) Cannabis

O gênero *Cannabis sp.* compreende um grupo de plantas herbáceas da família das *Cannabaceae*, da qual também pertence o gênero *Humulus*, representado pelo lúpulo. A espécie *Cannabis sativa* L. é considerada atualmente a única representante de seu gênero, originária dos planaltos da Ásia Central, havendo evidências de domesticação da planta há mais de 12 mil anos na atual China para usos têxteis e medicinais (REN *et al*., 2021). O termo "cannabis" pode se referir às flores (ou inflorescências) da espécie *C. sativa* ou à planta como um todo (UNODC, 2022).

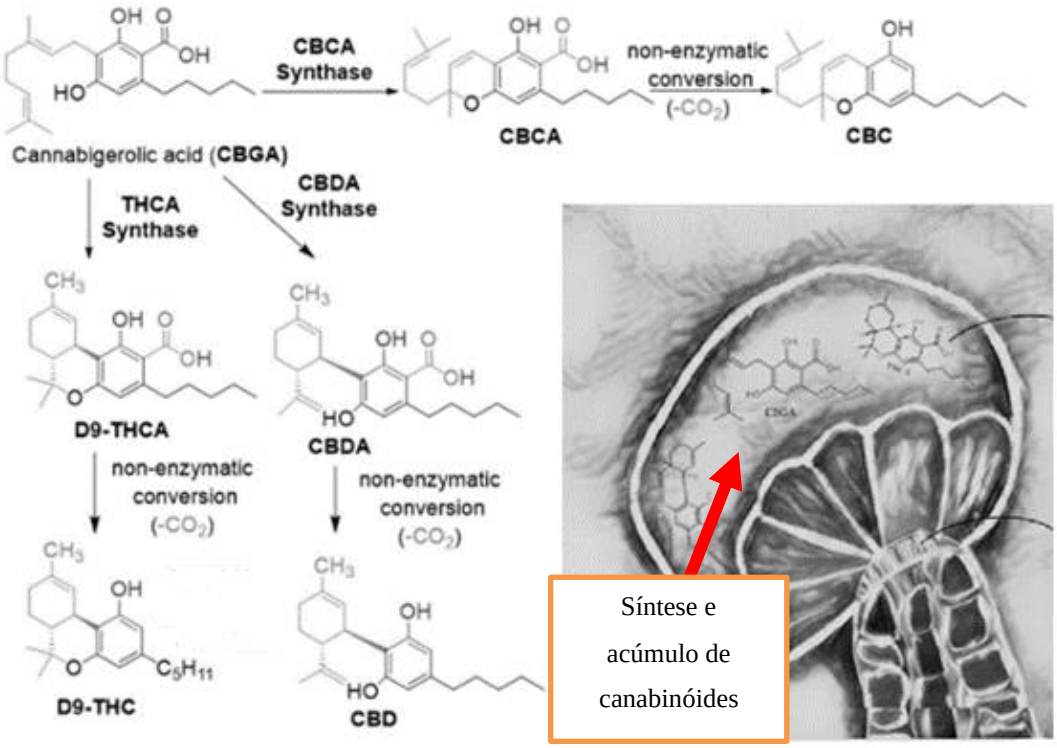
A cannabis, desde sua domesticação na China, foi submetida à seleção artificial e hibridização entre diferentes cultivares ao longo dos anos. Embora parte da sua história de domesticação tenha sido ocultada devido ao crescimento clandestino, hoje é possível fazer sua classificação taxonômica a partir de genes específicos que cada variedade apresenta em quatro grandes grupos: tipo basal, tipo droga, tipo droga-selvagem e tipo cânhamo (REN *et al*., 2021).

A cannabis possui importância econômica na produção de fibras de cânhamo, usadas pela indústria de tecidos, e principalmente como fonte de substâncias psicoativas classificadas como fitocanabinóides, sendo os mais importantes o delta-9 tetraidrocanabinol (THC) e o canabidiol (CBD). Além dessa classe, a cannabis também produz uma série de compostos bioativos, como terpenos, compostos fenólicos e alcalóides (ALIFERIS; BERNARD-PERRON, 2020; RADWAN *et al*., 2021).

Os fitocanabinóides são produtos do metabolismo secundário da planta, sendo produzidos e acumulados nos tricomas das inflorescências. Os precursores iniciais de sua biossíntese são o ácido olivetólico, derivado da condensação e ciclização do malonil-CoA, e o geranil-pirofosfato, um derivado isoprênico ativado. A transferência do grupamento isoprênico para o ácido olivetólico produz o ácido canabigerólico (CBGA), do qual se originam o ácido delta-9 tetraidrocanabinólico (THCA), o ácido canabidiólico

(CBDA) e o ácido canabicromênico (CBCA) (Figura 2). Todos os fitocanabinóides ácidos podem sofrer descarboxilação não-enzimática catalisada por aquecimento ou ao longo do tempo, dando origem às suas espécies neutras, CBD e THC (TAHIR *et al*., 2021).

Figura 2: Mecanismo de biossíntese dos principais canabinoides. Estrutura do tricoma glandular da cannabis, local de biossíntese de fitocanabinóides (quadro interior à direita). Adaptado de TAHIR *et al*., 2021.



Devido a diversidade de compostos de interesse econômico produzida por essa planta, o perfil químico metabólico da cannabis tem sido estudado em pesquisas metabolômicas por ressonância magnética nuclear (MARKLEY et al., 2017) ou por cromatografia associada à espectrometria de massas (FUHRER; ZAMBONI, 2015). Alguns tópicos estudados nessa linha de pesquisa são o perfil terpênico e canabinoide da inflorescência durante os estágios de desenvolvimento, os efeitos da temperatura e da polaridade de solventes de extração no conteúdo canabinoide de extratos (POLITI et al., 2008), a discriminação de variedades de cannabis de acordo com o perfil terpênico de óleos essenciais (HILLIG, 2004) e a quimiotaxonomia de linhagens de cannabis quanto a seu conteúdo de THC e CBD (ALIFERIS; BERNARD-PERRON, 2020; HILLIG; MAHLBERG, 2004).

Diferentes classificações da cannabis quanto a composição de seus canabinóides principais, ou seja, seus quimiotipos, são propostas na literatura. A classificação mais

antiga é proposta por Small e Beckstead (1973), de acordo com as proporções das concentrações de THC e CBD (SMALL; BECKSTEAD, 1973). Mandolino e Carboni (2004) propuseram uma classificação mais extensa, dividindo as variedades de cannabis em 5 quimiotipos, de acordo com as concentrações de três canabinóides principais: THC, CBD e canabigerol (CBG) (MANDOLINO; CARBONI, 2004). Outra classificação, usada para fins forenses e econômicos, é adotada pelo Escritório das Nações Unidas sobre Drogas e Crime (*United Nations Office on Drugs and Crimes*, UNODC) de acordo com a razão entre a soma das áreas dos picos cromatográficos do THC e de seu produto de oxidação, o canabinol (CBN), e a área do pico cromatográfico do CBD. No caso de a razão ser maior que 1, a cannabis é considerada do tipo droga, e, no caso de ser menor que 1, é considerada para fins industriais (UNODC, 2022).

3.5) Atividade biológica

O THC e o CBD são os fitocanabinóides da cannabis com efeitos bioativos mais estudados. No organismo humano, essas substâncias interagem com o sistema endocanabinóide, um conjunto de receptores que são ativados por substâncias endógenas derivadas de ácidos graxos, tendo a anandamida e o 2-araquidonoilglicerol como seus representantes principais. Os receptores principais são o CB1 e o CB2, interagindo com diversos processos fisiológicos do sistema nervoso central e do sistema imunológico. Além desses receptores, foram descritos recentemente outros receptores que endocanabinóides interagem: os receptores vanilóides e um grupo de receptores acoplados à proteína G (REZENDE *et al*., 2023).

Os receptores canabinóides participam de um tipo específico de regulação pós-sináptica entre neurônios: o mecanismo de regulação retrógrada, sendo altamente expressos nos neurônios pré-sinápticos. Esses receptores metabotrópicos são ativados pelos endocanabinóides liberados pelo neurônio pós-sináptico, desencadeando a ativação da proteína G inibitória na superfície interna da membrana celular e inibição da síntese de adenosina 3’,5’-monofosfato cíclico pela enzima adenilato ciclase (KANO, 2014).

Estudos apontam que o sistema endocanabinóide está envolvido em diversos processos fisiológicos: homeostase, ansiedade, regulação do apetite, comportamento emocional, depressão, funções nervosas e neurogênese, cognição, aprendizagem, memória, nocicepção e fertilidade. Devido a isso, esse sistema se caracteriza como um conjunto de alvos terapêuticos potenciais para desordens associadas a esses processos (LOWE *et al*., 2021).

3.6) Regulamentação do uso de cannabis

Embora recentemente o uso medicinal da cannabis e os estudos científicos movimentem vultuosos recursos econômicos (NEW FRONTIER DATA, 2023), seu uso como droga recreativa ao longo da história é cercado de estigmas sociais e de discriminação de seus usuários. Existem, atualmente, diferentes posicionamentos relacionados às legislações que tratam da produção e do porte da cannabis em vários países. O impacto do uso recreativo de cannabis na saúde pública vem sendo estudado nos locais onde o seu uso é permitido (SCHEIER; GRIFFIN, 2021; WATSON *et al*., 2023) e, embora seu uso crônico não apresente riscos graves associados à letalidade (CONNOR *et al*., 2021), a conscientização tem sido adotada como uma forma de redução de danos (FISCHER *et al*., 2022).

A cannabis utilizada como droga de abuso é denominada maconha, havendo uma extensa lista de sinônimos e gírias relacionadas ao seu nome e uso. A maconha é constituída da flor da cannabis seca ou prensada, com ou sem a presença de outras partes da planta como galhos, caules, folhas e sementes. Outra apresentação da droga é a resina obtida dos tricomas glandulares, conhecida como haxixe (UNODC, 2022).

O uso dessas apresentações da maconha ocorre principalmente por inalação da fumaça do cigarro, embora haja uso por via oral de alimentos contendo a resina da planta. Os efeitos do uso são euforia, alteração da percepção da realidade e do tempo, letargia, boca seca, aumento do apetite, alterações cognitivas e de memória (CONNOR *et al*., 2021).

A exploração econômica da cannabis é feita em países da Europa e da América do Norte, sendo o plantio de variedades contendo menos que 0,3% de THC, caracterizadas como cânhamo, legalizado em âmbito federal no Canadá e nos Estados Unidos da América, para uso pelas indústrias têxtil, cosmética e de suplemento alimentar (UNODC, 2022). Nos Estados Unidos, até 2020, 31 dos 56 estados e territórios americanos descriminalizaram o uso de cannabis, sendo o uso do canabidiol legalizado em todos, exceto 2, deles (ALHARBI, 2020).

Enquanto se observa uma tendência mundial de legalização do uso da cannabis e de produtos associados, no Brasil, são proibidos pela lei federal nº 11343 de 2006 - a chamada Lei Antidrogas - a importação, a exportação, o comércio, a manipulação e uso de substâncias consideradas narcóticas, sem especificar quais substâncias. A aplicação da Lei, considerada uma norma penal em branco, se faz através da resolução de diretoria

colegiada nº 344, de 1998, da Agência Nacional de Vigilância Sanitária, que define a planta do gênero *Cannabis* e o fármaco THC como narcóticos e, assim, de uso proibido. A resolução permite como exceções a prescrição e importação de produtos derivados de cannabis contendo CBD e/ou THC para fins de tratamento médico (ANVISA, 2024). Entretanto, a importação das plantas ou de partes delas *in natura* foi proibida (ANVISA, 2023). A maioria dos produtos atualmente registrados na ANVISA é à base de CBD, embora o Mevatyl, conhecido como Sativex internacionalmente, indicado para tratamento de sintomas da esclerose múltipla, contenha em sua formulação teores aproximadamente equivalentes de CBD e THC (25 e 27 mg/mL, respectivamente) (ANVISA, 2018).

Embora apresente flexibilização do uso medicinal da cannabis, a regulamentação ainda restringe o acesso a terapias à base da planta ou de THC e/ou CBD e o cultivo. O custo de aquisição do Mevatyl é considerado elevado, bem como a importação de produtos à base de cannabis (O GLOBO, 2023). Enquanto isso, pacientes que não têm condições para compra do produto buscam permissão por via judicial para plantio e produção de extratos medicinais (com proteção jurídica por meio do mecanismo de *Habeas corpus*), ainda assim possuindo como entrave o acesso à assistência técnico-científica para a produção dos medicamentos, bem como à caracterização química e ao controle de qualidade dos mesmos (CARVALHO *et al*., 2022a).

Outro problema enfrentado pelos pacientes consiste na diferenciação dos produtos à base de cannabis para fins medicinais daqueles para fim recreativo, incluindo a maconha apreendida do tráfico de drogas. Nesse ponto, há esforços para a caracterização de ambos os produtos por técnicas de cromatografia líquida ou cromatografia gasosa com base nos canabinóides principais e terpenos (CARVALHO *et al*., 2022b, 2022a; ROCHA *et al*., 2020). Entretanto, não há estudos mais abrangentes sobre o perfil químico de ambas as apresentações da cannabis.

4) METODOLOGIA

4.1) Obtenção de amostras

Foram utilizadas amostras de cannabis provieram de duas fontes distintas (ilustradas na figura 3): cannabis cultivadas em contexto medicinal; e apreendidas pela Policia Civil do Estado do Rio de Janeiro. As cultivadas foram provenientes advinda de pacientes**,** ou seus familiares com autorização judicial para seu plantio e uso para tratamento de doenças como epilepsia refratária aos tratamentos tradicionais, câncer, dor crônica, entre outros (ROCHA *et al*., 2020). Tais amostras foram adquiridas a partir do projeto EtnoS-Farmacannabis, da Faculdade de Farmácia da UFRJ, o qual é responsável pelo suporte farmacêutico aos pacientes, sendo uma das suas atividades a caracterização química das plantas de cannabis cultivadas. As amostras foram caracterizadas morfologicamente e secas conforme trabalho anterior (CARVALHO *et al*., 2022b), sendo denominadas neste trabalho como "cannabis medicinal" e individualmente referidas pelo código SP.

As amostras de apreensão, sendo nesse contexto ilícito chamadas de maconha, foram inicialmente submetidas à análise forense pelo Instituto de Criminalística Carlos Éboli para produção da prova material e posteriormente cedidas para pesquisa de doutorado do pesquisador Fernando Gomes de Almeida do Programa de Pós-Graduação em Ciências Farmacêuticas da UFRJ. Essas amostras foram caracterizadas como prensado de partes da planta como folhas, flores, galhos e sementes, sendo denominadas neste trabalho como "cannabis de apreensão" e individualmente referidas pelo código AP.

No total, dados de análise de 23 amostras de cannabis medicinal e 44 amostras de cannabis de apreensão foram gentilmente cedidos pelo pesquisador Fernando Almeida para uso neste trabalho de mestrado. Os métodos de extração e análise instrumental encontram-se descritos a seguir.

Figura 3: Apresentação de duas amostras representativas dos grupos de cannabis medicinal acima) e de apreensão (abaixo).

4.2) Extração de metabólitos

A amostra triturada (100 mg) foi transferida para tubos de polipropileno de 15 mL (tipo *Falcon*) e foram adicionados 10 mL de solvente para extração (metanol:n-hexano, 9:1, v/v). A mistura foi submetida à homogeneização em vórtex (1 min), ultrassonicação (10 min) e centrifugação (2007 g/5 min). O sobrenadante foi transferido para balão volumétrico de 25 mL. O processo, a partir da adição do solvente, foi repetido duas vezes, com os volumes de 10 mL e 5 mL de solvente. Por fim, avolumou-se o sobrenadante recolhido para 25 mL com o mesmo solvente.

O extrato obtido (8 mL) foi concentrado até secagem do solvente em evaporador automático a base de fluxo de gás nitrogênio modelo Auto EVA-20 Plus (marca RayKol), nas seguintes condições: temperatura do banho-maria de 40ºC, pressão de 3 psi, velocidade de descida da agulha de 0,7 mm/min, tempo total de concentração de 45 minutos. Os extratos foram ressuspensos em 4 mL de n-hexano, sendo 1 mL transferido para frascos *vials* para análise por CG-EM.

4.3) Análise por CG-EM

Os extratos foram analisados em uma abordagem metabolômica do tipo global. O equipamento utilizado foi o cromatógrafo a gás acoplado ao espectrômetro de massas do tipo quadrupolo modelo GC-MS QP 2010 Ultra, marca Shimadzu. A coluna capilar usada na análise foi a DB-5MS, 0,25 µm de espessura do ligante, 5% de fenilmetilpolisiloxano, 30 m x 0,25 mm de dimensão. O método instrumental foi constituído das seguintes condições: temperatura da fonte de íons a 220 ºC, temperatura da interface de 280 ºC, gradiente de temperatura iniciando em 100 ºC, com rampa de 10 ºC/min até 300 ºC, com manutenção da mesma por 4 min, constituindo assim 24 minutos de análise. O volume de injeção foi de 1 µL, com *split ratio* de 1/20. O filamento de ionização foi ligado somente a partir de 4 minutos, a fim de esperar a saída do sinal do solvente. O filamento foi desligado novamente em dois períodos de tempo, correspondentes aos tempos de retenção dos canabinoides majoritários, THC ou CBD, de acordo com o perfil da amostra que foi previamente determinado por CLAE-DAD, a fim de não saturar o detector com a concentração elevada do canabinoide majoritário. Para as amostras de cannabis que apresentaram perfil majoritário de CBD, maioria das cultivadas em contexto medicinal, o período de supressão de sinal foi ajustado entre 16,60 e 17,30 minutos (tempo de retenção correspondente ao canabidiol), enquanto as amostras de perfil majoritário THC, quase todas as amostras de cannabis de apreensão, o período de supressão de sinal foi entre 17,60 e 18,10 minutos (tempo de retenção correspondente ao tetrahidrocanabinol).

Cada lote de amostras analisado foi constituído por uma análise de amostra controle contendo somente o solvente (n-hexano, grau HPLC) empregado na reconstituição dos extratos, sendo considerada um controle de branco do sistema. Tais amostras foram denominadas ao longo do trabalho como “amostras branco” e foram 16 no total. Após tal análise, no mesmo lote, um conjunto de amostras de um dos grupos foi analisada três vezes de modo seriado, sendo assim consideradas replicatas técnicas.

4.4) Processamento dos dados brutos

Os sinais gerados de cada amostra por GC-MS foram submetidos a conversão de formato, utilizando o *software* GCMS ChemStation para o formato genérico mzXML a fim de serem importados para o software Mzmine versão 2.53. Em seguida, foi realizado o processamento dos dados até a obtenção de *features* relativos a cada composto detectado analiticamente, estando o caminho para a função e os parâmetros usados descritos no

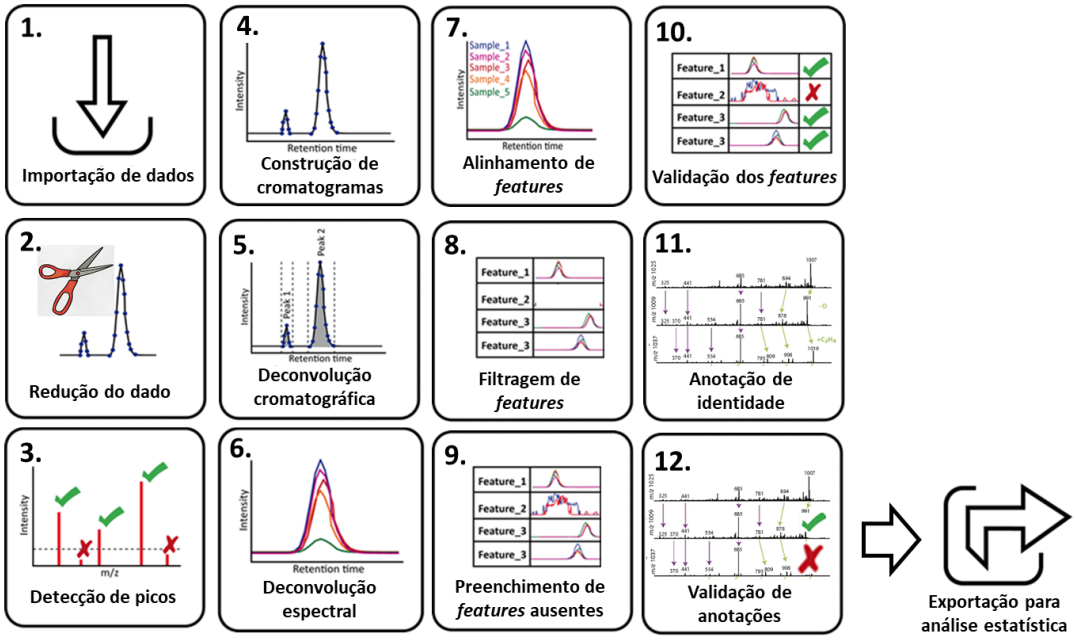
Anexo I. Os dados brutos foram processados e analisados como um todo, conforme o ilustrado na Figura 4.

Após importação dos dados, foi realizada a redução dos dados cromatográficos em uma faixa de tempo de retenção (ferramenta *Crop Filter*), para remoção de artefatos dos dados convertidos. Posteriormente, foi realizada a detecção dos picos espectrométricos acima de uma intensidade mínima (conforme Anexo I), recomendando ser acima do que é visualmente considerado ruído (*Peak Detection*).

Após isso, foi realizada a construção de cromatogramas de valores de m/z que apresentassem comportamento cromatográfico (ex: cromatogramas de determinado m/z com picos aparentes acima de uma intensidade mínima (conforme Anexo I), considerando o perfil de intensidades do pico cromatográfico e do ruído. Ferramenta *ADAP Chromatogram Builder*). Nos cromatogramas gerados de cada análise, foi executado um algoritmo de identificação de picos cromatográficos com configuração otimizada (Anexo I) para selecionar picos com formato e intensidade (*Chromatogram Deconvolution*).

Com o conjunto de picos identificados foi executado, para cada análise individualmente, o algoritmo de deconvolução espectral *Multivariate Curve Resolution* para o agrupamento de picos de m/z diferentes, mas com perfil cromatográfico similar. Assim, foram agrupados os íons precursores e seus respectivos fragmentos oriundos da ionização por elétrons de cada composto num tempo de retenção específico. Em seguida, foi realizado o alinhamento dos picos entre amostras e duas filtragens de *features:* a primeira para remoção de *features* presentes no tempo de retenção onde o filamento da fonte de ionização era desligado nos dois métodos instrumentais e a segunda para remoção de *features* que ocorriam em menos de 4 amostras. Em seguida, utilizou-se um algoritmo de preenchimento de picos faltantes em cada amostra para cada *feature (Same RT and m/z range gap filler),* a partir da integração dos sinais da mesma faixa de tempo de retenção. Obteve-se como resultado a tabela de todos os *features* detectados em cada amostra, com seu perfil espectrométrico e tempo de retenção definidos. Por fim, realizou-se a inspeção manual de todos os resultados de forma a ajustar possíveis processos de integração errônea de picos e remover *features* duplicados ou de baixa qualidade.

Figura 4: Fluxo de trabalho de processamento dos dados brutos de CG-EM no programa Mzmine



4.5) Anotação de identidade dos *features*

Para anotação dos *features* presentes na tabela, foi utilizada a opção *Local spectra database search*, utilizando o algoritmo de similaridade espectral *Weighted dot-product Cosine*, com o valor mínimo de cosseno de similaridade de 0,7. Como bancos de dados, foram utilizados dois conjuntos de bancos de dados de espectros de massas obtidos por ionização por elétrons.

O primeiro conjunto foi o do *Mass Bank of North America* (MoNA), de acesso aberto e que está disponível no endereço da web https://mona.fiehnlab.ucdavis.edu/downloads. O segundo conjunto foi o do banco de espectros do NIST versão 2020, acesso via aquisição de licença, o qual foi obtido por conversão do banco de dados em seu formato bruto para o formato *MSP* pelo *software* LIB2NIST, obtido no endereço web https://chemdata.nist.gov/mass-spc/ms-search/Library_conversion_tool.html. O arquivo convertido foi posteriormente processado por uso de script na linguagem R (Anexo II), para correção da formatação e reconhecimento do banco pelo Mzmine. As anotações obtidas foram analisadas manualmente a fim de verificar possíveis correlações de baixa similaridade entre os espectros experimentais e dos bancos de dados.

4.6) Tratamento de dados pré-análise estatística

A tabela contendo os *features* do experimento foi exportada para o formato .csv para análise na plataforma MetaboAnalyst (PANG et al., 2021), sendo importada a partir do modo *Statistical Analysis (One Factor)*. A importação da tabela foi feita configurando o tipo de dados para *Peak Intensities* e formato de amostras em colunas não-pareadas. Na etapa de *Data Integrity Check,* selecionou-se o botão *Missing Values* e foi escolhida especificamente a substituição de valores ausentes por 1/5 do mínimo valor de detecção do *feature* dentre as amostras de seu grupo, sem realizar a remoção de *features.*

Após isso, foi escolhida a opção *Download* para os dados normalizados, sendo baixada a tabela com todas as intensidades preenchidas. Esta tabela foi submetida a um processamento via *script* de programação na linguagem R (anexo III), a fim de gerar uma tabela contendo a média das intensidades de cada *feature* entre as replicatas técnicas de cada unidade experimental, que é a amostra extraída de origem, junto com todas as amostras branco.

4.7) Análise estatística

A tabela processada foi importada para o MetaboAnalyst novamente, sendo feita como processamento básico a centralização pela média dos valores e escalonamento pelo desvio padrão. Foram utilizados para comparação os dados sem normalização, com normalização pela soma e com normalização pela mediana. Também, foram comparados os dados não-transformados e transformados por logaritmo na base 10. Para comparação, foi utilizada a análise por componentes principais (PCA).

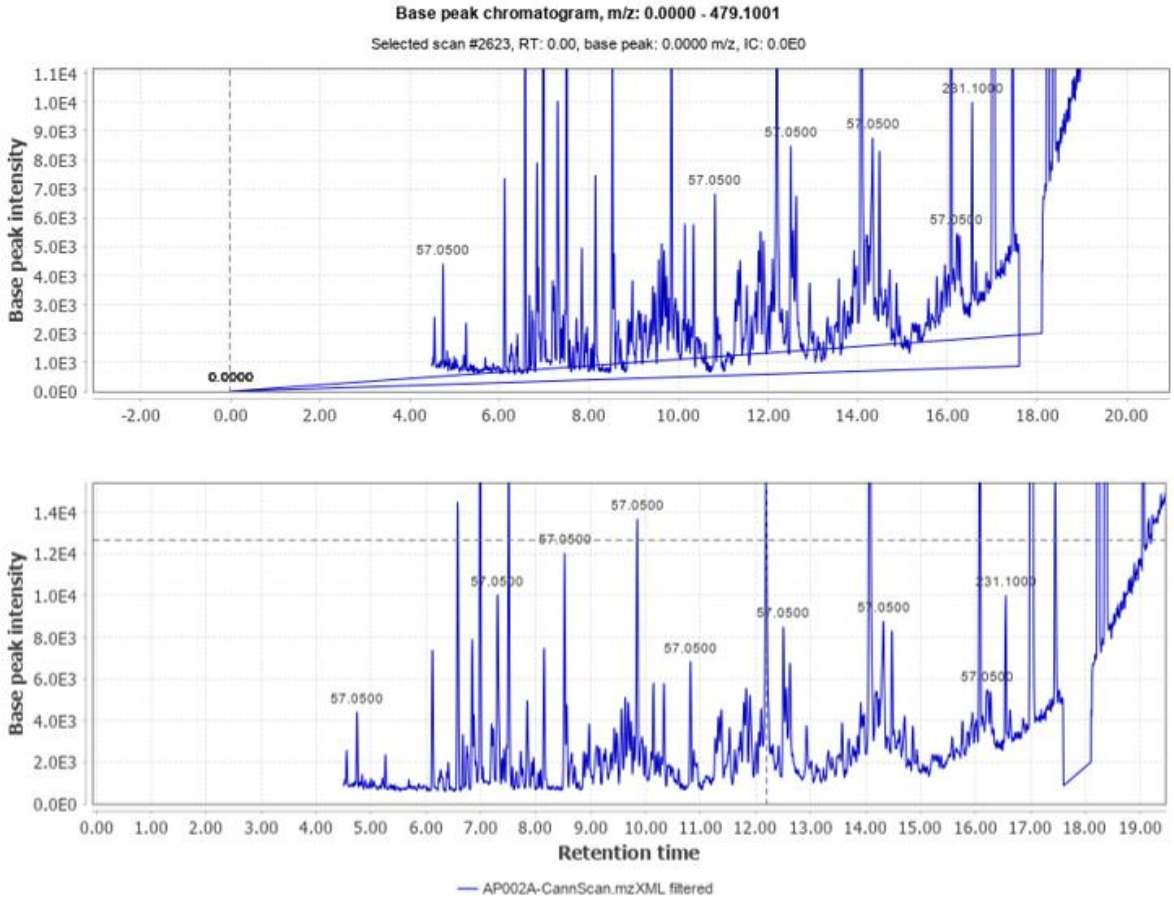
Os dados contendo somente os grupos de cannabis de apreensão e medicinal foram analisados centrados pela média e escalonados pelo desvio padrão, utilizando-se a análise de componente principal. Posteriormente foram usadas as áreas de confiança calculadas pelo teste T^2 Hotelling como critério de exclusão de amostras aberrantes e os grupos foram submetidos à análise de componentes principais novamente, além de serem analisados por métodos de aprendizado de máquina não supervisionados, como clusterização e *K-Means,* e métodos supervisionados PLS-DA (Análise discriminante por mínimos quadrados parciais). Os resultados foram ilustrados em gráficos do tipo *heatmap* e *volcano plot*. pela análise de componentes principais novamente, além de serem analisados por outras ferramentas como clusterização por *Heat Map , K-Means*, PLS-DA e *volcano plot*.

5) RESULTADOS E DISCUSSÃO

5.1) Processamento dos dados brutos

Os dados de GC-EM brutos provindos de conversão para o formato *mzxml* foram importados para a execução do fluxo de trabalho de processamento para análise metabolômica global. Inicialmente, apresentaram um erro na sua leitura, contendo um espectro no tempo zero que é artefato do modo de aquisição utilizado (Figura 5, cromatograma superior) uma vez que houve a supressão do sinal no tempo de retenção de THC ou CBD, a depender do perfil da amostra analisada, rica em THC ou CBD, para não haver saturação do detector . Dessa forma, a ferramenta *crop filter* foi utilizada com essa finalidade alternativa de remover tal artefato e manter o cromatograma contínuo (Figura 5, cromatograma inferior). À exceção desse fato, o processamento ocorreu sem apresentar erros em sua execução.

Figura 5: Perfis cromatográficos representativos do dado de uma amostra analisada apresentando o erro (cromatograma superior) e com o mesmo corrigido depois do uso da ferramenta *crop filter* (cromatograma inferior).



O processamento dos dados brutos gerados nas análises por espectrometria de massas e cromatografia líquida é uma etapa chave para a obtenção de informações válidas a respeito do perfil químico de amostras de análise metabolômica. Assim, a configuração correta dos parâmetros de cada algoritmo utilizado em todo o processamento é fundamental. O comportamento geral observado é que os parâmetros podem ser configurados de forma mais flexível ou mais restringente, sendo o equilíbrio entre esses dois extremos necessário para o assinalamento correto da maior parte dos *features* presentes no dado, mas evitando a adição de *features* errôneos que contribuem para o aumento de variáveis sem relevância ou significado analítico e que, consequentemente, aumentam o ruído computacional e prejudicam o poder estatístico de análises posteriores (BORGES et al., 2022; YU; CHEN; HUAN, 2021).

O ajuste dos diversos parâmetros é dependente das características da metodologia analítica aplicada em sua pesquisa. Tais características são a largura do pico cromatográfico, o número de pontos que compõem o pico, a intensidade absoluta mínima dos sinais espectrométricos e cromatográficos, o nível de sinal/ruído mínimo dos mesmos, a tolerância das medidas de tempo de retenção ou da relação massa/carga de acordo com o desvio esperados desses, entre outros (BORGES *et al*., 2022).

Sendo assim, é necessário que o pesquisador tenha conhecimento dessas características, que podem ser avaliadas a partir do uso da metodologia empregada para a análise de uma amostra conhecida, como uma mistura de padrões analíticos. Na ausência dessa, pode-se utilizar os dados de uma amostra representativa do conjunto de amostras analisadas, em comparação com um controle negativo do sistema (por exemplo, injeção do solvente da amostra) (BORGES *et al*., 2022).

Neste trabalho, foi utilizado o *software* Mzmine, versão 2.53, como plataforma para processamento dos dados analíticos. O Mzmine é um dos *softwares* mais utilizados para análise metabolômica global e o mais citado em sua área, tendo como grande vantagem a inspeção em tempo real do resultado de cada processamento em seu fluxo de trabalho. Isso permite a otimização dos parâmetros de cada etapa de forma melhor direcionada, aumentando a cobertura dos *features* presentes em todo o dado. Isso também reduz possíveis assinalamentos espúrios que contribuem para o aumento de ruído computacional no dado que será submetido a análise estatística (YU; CHEN; HUAN, 2021).

Embora existisse a versão 3.0 do software no momento da execução desse trabalho, optou-se pelo uso da versão anterior devido a menor exigência de recursos

gráficos e de processamento computacional, o que permite o uso do software em computadores de menor desempenho.

Observaram-se alguns problemas na aplicação dessa metodologia. Um deles foi a conversão errada do banco de dados do formato padrão do NIST para o formato *MSP*, de forma a impedir o uso do banco de dados na etapa de anotação de identidade. Outro é o erro existente nos arquivos *mzxml* provindos especificamente de *softwares* da empresa que produz e comercializa as plataformas analíticas, nesse caso, a Shimazdu. Outro ponto foi a necessidade de se formatar a tabela de *features* obtidos do Mzmine e calcular a média dentre as replicatas técnicas das abundâncias dos *features* em cada amostra. Em todos os exemplos, foram propostas resoluções utilizando ferramentas inerentes do Mzmine e utilizando *scripts* de programação na linguagem R, sendo esse último uma ferramenta valiosa na formatação e automação de procedimentos da análise de dados.

5.2) Anotação dos *features*

Após aplicação do fluxo de trabalho básico nos dados utilizando o Mzmine, obteve-se uma lista de 22 *features* no total, identificados no conjunto de amostras (Tabela 1), com área sob a curva, tempo de retenção e espectro de fragmentação definidos. Utilizando-se o algoritmo de identificação por similaridade espectral, dos 22 *features* totais, 16 anotações foram obtidas a partir do banco MoNA e 20 foram obtidas pelo banco de dados NIST (Tabela 1), sendo que em ambos houve casos com uma ou mais identificações. Para as análises posteriores, foi considerada a anotação com maior pontuação de similaridade por *feature* em conjunto com avaliação manual dos espectros experimentais e do banco de dados.

A qualidade da anotação de identidade dos *features* advindos de análise metabolômica por cromatografia e espectrometria de massas pode ser classificada em níveis (DUNN *et al*., 2017; SCHYMANSKI *et al*., 2014). Tais níveis são estabelecidos, conforme descrito na Figura 6, indicando ordens crescentes de confiança e especificidade do nível 5 ao nível 0. No caso deste trabalho, por conta do uso de uma biblioteca espectral contendo dados sobre as substâncias esperadas, todas as identificações são consideradas como de nível 2, descritas na Tabela 2.

Tabela 1:*Features* detectados após processamento no Mzmine, com suas anotações provindas dos bancos de dados MoNA e NIST e seus respectivos valores de similaridade.

ID	*m/z*[1]	TR[2]	Anotação MoNA	Cos[3] MoNA	Anotação NIST	Cos[3] NIST
1	275.162	18.32	1,5-diaminoantraquinona	0.849	canabinol	0.890
2	149.121	10.98	1β-hidroxi-1α,(4a)β-dimetill-7beta-(1-metil -1-hidroxietil)-(8a)β-decaidronaftaleno	0.815	beta-eudesmol	0.761
3	63.585	8.90	1,3,5-triisopropilbenzeno	0.808	10-epi-γ-eudesmol	0.854
4	59.018	16.34	oleamida	0.701	oleamida	0.861
5	68.050	11.24				
6	116.452	8.58	elemol	0.804	guaiol	0.901
7	93.082	7.47	cariofileno	0.736	germacreno B	0.722
8	82.074	11.69	11-eicosenol	0.736	acetato de fitila	0.752
9	93.084	6.56	cariofileno	0.893	β-cariofileno	0.907
10	93.085	6.96	Formiato de linalila	0.748	α-humuleno	0.881
11	84.488	18.95	docosano	0.761	Δ9-tetrahidrocanabinol	0.734
12	69.053	9.60	α-cedreno	0.756	α-bisabolol	0.897
13	241.154	16.06	4-aminoantipirina	0.717	tetrahidrocanabivarina	0.869
14	59.029	9.32	elemol	0.808	α-eudesmol	0.907
15	150.100	5.56				
16	175.069	14.46	2,3:5,6-bis(trimetileno)piran-4-ona	0.711	canabicromeorcina	0.775
17	83.055	10.12			2-dodecoxyethanol	0.774
18	203.065	15.17	4-aminoantipirina	0.830	canabidivarol	0.888
19	71.055	14.05			fitol	0.852
20	73.035	12.49	ácido palmítico	0.805	ácido palmítico	0.821
21	231.108	15.76			canabicitrano	0.798
22	57.045	13.12			nitrato de potássio	0.731

Nota: [1]Relação *m/z* do pico base. [2]Tempo de retenção, em minutos. [3]Cosseno.

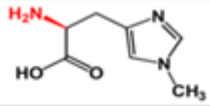
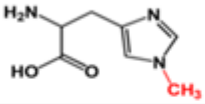
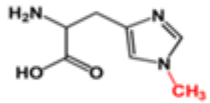
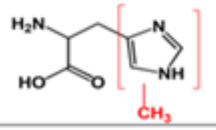
O nível 0 de qualidade de anotação, por conta do uso de um padrão de referência da substância de interesse, é a condição de maior confiança em análises metabolômicas. Tal qualidade é atribuída pelo fato de se obter as informações relativas à caracterização da substância de interesse (como tempo de retenção, relação massa/carga e perfil de fragmentação) nas exatas condições da metodologia analítica ao se analisar o padrão de referência e sem a existência de isômeros constitucionais indistinguíveis pela técnica (por exemplo, estereoisômeros). Assim, há de se considerar a existência de substâncias com estruturas similares aos analitos mesmo com uso de padrões analíticos na etapa de anotação de identidade (DUNN *et al*., 2017).

No caso da análise global, a anotação é mais desafiante, pois se utiliza de informações (por exemplo, tempo de retenção, medida de massa exata ou perfil de fragmentação de uma substância) obtidas de outros equipamentos. Isso resulta em menor confiança na anotação por conta dos pontos supracitados e adicionado da variação de tais características intrínsecas da molécula analisada.

Especificamente, o perfil de fragmentação obtido pelo processo de fragmentação induzida por ionização por elétrons a 70 eV é bem reprodutível e utilizado como referência na anotação de análises de cromatografia gasosa acoplada a espectrometria de

massas. Entretanto, nas análises realizadas nesse trabalho, observou-se diferença na anotação de identidade dos *features* ao se usar dois bancos de dados diferentes (MoNA e NIST). Possíveis causas são a ausência de espectros das substâncias de interesse, a qualidade dos espectros presentes no banco de dados e associação errônea entre espectros experimental e teórico devido a forma de funcionamento do algoritmo de similaridade na análise de espectros ambíguos (GROSS, 2004).

Figura 6: Níveis de qualidade de anotação de identidade de compostos por análise metabolômica. Adaptado de CUYKX *et al*., 2018.

	Nível de confiança	Requerimentos
	**Nível 0:** Estrutura sem ambiguidade	Estrutura completa e estereoquímica do composto isolado
	**Nível 1:** Estrutura confiável	Padrão de referência anotado usando pelo menos duas técnicas ortogonais Análise de EM,EM/EM e tempo de retenção correspondentes ao de padrões de referencia
	**Nível 2:** Estrutura provável	Correspondência com literatura ou banco de dados com pelo menos duas técnicas Ex: Análise de EM,EM/EM e tempo de retenção correspondentes a bancos de dados
	**Nível 3:** Estrutura ou classe provavel, com mais de um possível candidato	Correspondência com pelo menos duas técnicas suportando estrutura ou classe Ex: Análise de EM,EM/EM e tempo de retenção (indicativo)
C7H11N3O2	**Nível 4:** Fórmula química inequívoca, mas sem identificação de classe	Correspondência do espectro com a fórmula proposta (com picos do envelope isotópico)
169.0851	**Nível 5:** Desconhecido	Sem identificação, somente medida de m/z

Nota: EM: Medida de relação *m/z* exata. EM/EM: perfil de fragmentação em análises sequenciais, com sua dissociação por ionização eletrônica ou induzida por colisão. Tempo de retenção se refere ao tempo de retenção característico do composto de interesse em uma análise cromatográfica quando comparado a um padrão da substância usando o mesmo método.

Com relação ao primeiro caso, considerada a espécie analisada neste trabalho, é presumível a presença de compostos da classe de fitocanabinóides no extrato analisado. Entretanto, o banco de dados MoNA não possui entradas de substâncias dessa classe, assim, o algoritmo retornou anotações consideradas errôneas para maior parte dos *features* (exemplo na Figura 7). Em comparação, o banco de dados NIST apresentou anotações de identidades melhor relacionadas ao perfil químico esperado do extrato analisado (Tabela 2).

Figura 7: Comparação de espectros de fragmentação experimental e teórico presente no banco de dados (acima e abaixo, respectivamente) presentes no MoNA e NIST. Os picos azuis indicam estar em comum entre os dois espectros e os picos laranjas só estão presentes em um dos espectros. (A): resultado obtido com o banco de dados MoNA, com identificação do composto como heneicosano. (B): resultado obtido com o banco de dados NIST, com identificação do composto como canabidiol. Cabe ressaltar que, devido ao método de análise estar programado para desligar o filamento de detecção nos períodos que os compostos TCH e CBD eluem, tal identificação de CB seria possível pela detecção de algum composto isomérico com perfil de fragmentação similar.

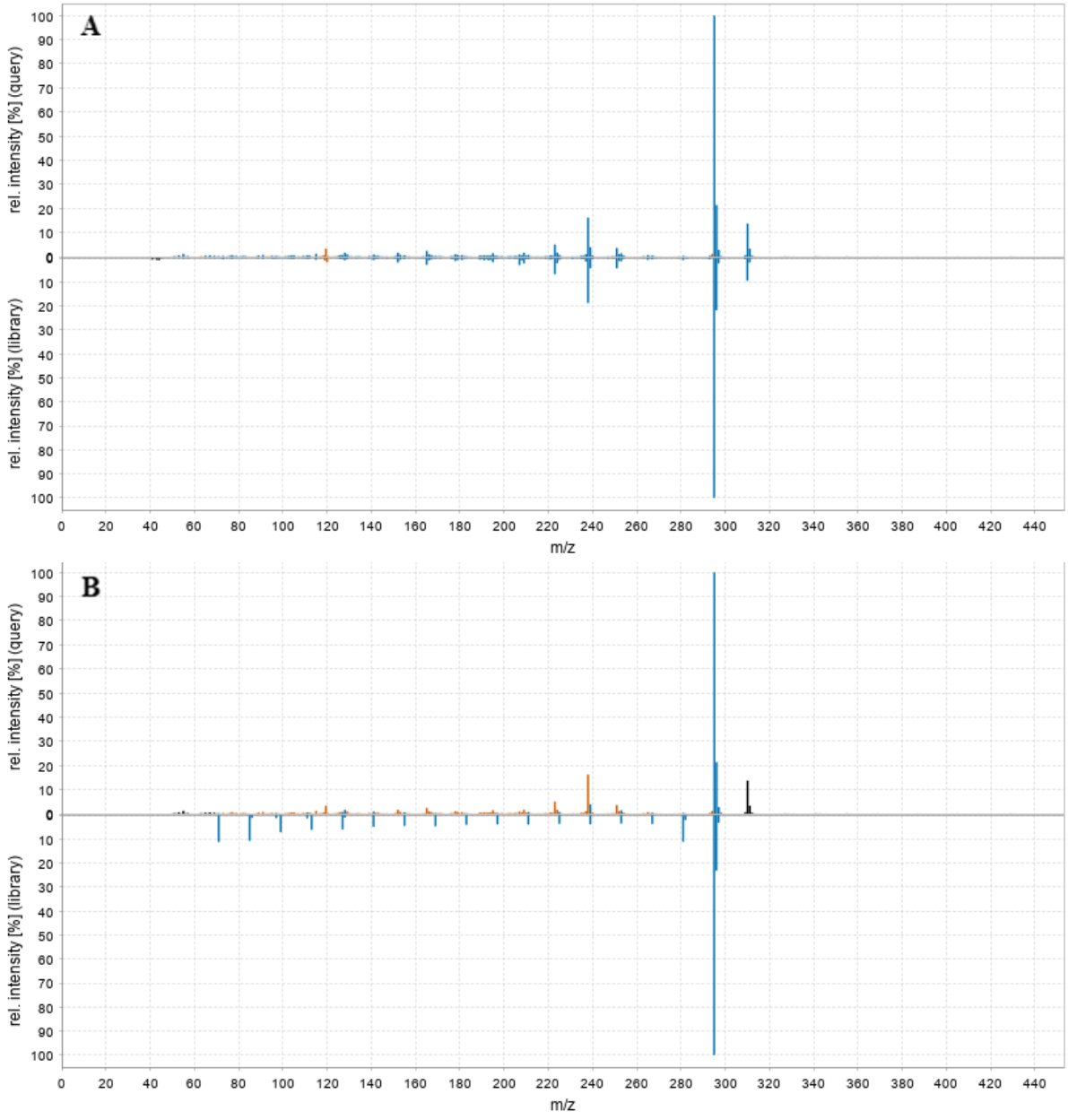


Tabela 2: Anotações de identidade dos *features* detectados na análise de amostras de cannabis, sendo os picos azuis coincidentes entre os espectros e os picos laranjas os que são exclusivos e um espectro.

ID	m/z[1]	TR[2]	Anotação Final	cos[3]	Número CAS	Fórmula Química
1	275.162	18.32	canabinol	0.890	521-35-7	C21H26O2
2	149.121	10.98	β-eudesmol	0.761	473-15-4	C15H26O
3	63.585	8.90	α-eudesmol	0.820	473-16-5	C15H26O
4	59.018	16.34	oleamida	0.861	301-02-0	C18H35NO
5	68.05	11.24	sem notação	-	-	-
6	116.452	8.58	guaiol	0.901	489-86-1	C15H26O
7	93.082	7.47	β-selineno	0.721	17066-67-0	C15H24
8	82.074	11.69	acetato de fitila	0.752	10236-16-5	C22H42O2
9	93.084	6.56	β-cariofileno	0.907	87-44-5	C15H24
10	93.085	6.96	α-humuleno	0.881	6753-98-6	C15H24
11	84.488	18.95	docosano	0.761	629-97-0	C22H46
12	69.053	9.60	α-bisabolol	0.897	23178-88-3	C15H26O
13	241.154	16.06	tetraidrocanabivarina	0.869	31262-37-0	C19H26O2
14	59.029	9.32	α-eudesmol	0.907	473-16-5	C15H26O
15	150.1	5.56	verbenona	0.747	80-57-9	C10H14O
16	175.069	14.46	canabicromeorcina	0.775	55824-09-4	C17H22O2
17	83.055	10.12	2-dodeciloxietanol	0.774	4536-30-5	C14H30O2
18	203.065	15.17	canabidivarol	0.888	24274-48-4	C19H26O2
19	71.055	14.05	fitol	0.852	150-86-7	C20H40O
20	73.035	12.49	ácido palmítico	0.821	57-10-3	C16H32O2
21	231.108	15.76	canabicitrano	0.798	31508-71-1	C21H30O2
22	57.045	13.12	sem notação	-	-	-

Nota: [1]m/z de referência. [2]Tempo de retenção em minutos. [3]Valor de cosseno da busca de similaridade espectral.

O algoritmo de busca de similaridade espectral baseia-se no cálculo do produto escalar entre os vetores multidimensionais referentes aos espectros experimental e teórico. Esses vetores são intercambiavelmente representados por matrizes, onde seus valores estão associados à presença de picos de m/z específicos e suas intensidades. Como resultado, encontra-se o valor de similaridade de cosseno de 0 até 1, sendo mais próximo de 1 o cálculo entre dois vetores mais próximos no espaço multidimensional, o que significa maior similaridade entre os espectros relacionados (ARON *et al.*, 2020).

Embora seja um processamento que automatiza a busca de similaridade de uma grande quantidade de espectros experimentais frente a um banco de dados, por conta dos motivos explicitados no parágrafo anterior, mostra-se necessária uma inspeção manual dos espectros associados. Observou-se neste trabalho algumas associações presumidas errôneas, considerando o conhecimento associado à natureza química da amostra e do

comportamento esperado da metodologia analítica usada. Um exemplo é a anotação de dois *features* como a substância alfa-eudesmol, indiferenciáveis pelo espectro de fragmentação e com tempos de retenção diferentes (8,90 e 9,32 min). Uma abordagem confirmatória para esse caso seria a análise de um padrão da substância em questão com a mesma metodologia analítica.

Conclui-se que o procedimento de anotação de identidade de *features* numa análise metabolômica global deve ser feita criteriosamente, sendo direcionada de forma racional considerando os conhecimentos prévios da amostra analisada, da qualidade do banco de dados usados e do mecanismo de funcionamento do algoritmo de similaridade espectral (SCHYMANSKI et al., 2014). Recomenda-se, ainda, a exploração de dados analíticos em repositórios de dados de metabolômica, de forma a avaliar previamente o comportamento das substâncias de interesse na pesquisa ou o desempenho da metodologia usada como informações adicionais a suportar as anotações dos analitos.

5.3) Normalização e transformação dos dados

Avaliaram-se os gráficos das análises de componentes principais com a comparação das 3 classes de amostras: cannabis de apreensão, cannabis medicinal e amostras branco (Figura 8). Em todos os gráficos foi possível observar os agrupamentos com tendência de separação dos 3 grupos. Também foi possível verificar a tendência do grupo "cannabis medicinal" apresentar maior variância dentre os três pela observação do maior espalhamento das amostras ao longo dos eixos dos dois primeiros componentes. Outra característica dos dados como um todo é a presença de amostras fora das áreas de confiança calculadas pelo teste T^2 Hotelling, consideradas desse modo discrepantes do comportamento geral de seu grupo.

Por outro lado, observaram-se alterações na característica dos dados ao se realizar a normalização em comparação com o dado não-normalizado (Figura 8A). Uma das alterações foi o aumento da variância do grupo branco, caracterizada pelo maior espalhamento das amostras desse grupo no gráfico do PCA. Enquanto no dado não-normalizado essas amostras encontram-se basicamente sobrepostas, ou seja, com grande similaridade, ao se aplicar a normalização tanto por soma quanto por mediana (Figuras 8B e 8C), observou-se o aumento da variância nesse grupo, comparável ao grupo de cannabis de apreensão. Já a transformação dos dados por logaritmo em base 10 (Figura 8D), observou maior separação entre os três grupos, com o mesmo aumento de variância no grupo branco em comparação ao grupo de cannabis de apreensão.

O grupo branco analisado se constitui como uma forma de análise de qualidade dos dados frente ao experimento como um todo, porém, esse foi retirado para a comparação direta dos grupos de estudo AP e SP. Ainda na PCA, observaram-se amostras fora da área de confiança de seus grupos, sendo tais amostras consideradas aberrantes (Figura 9). A fim de analisar a diferença dos dois grupos com maior precisão nas análises seguintes, foi feita a remoção de amostras aberrantes de ambos os grupos (SP004, SP017, SP035 e SP049 no grupo de cannabis medicinal e AP006 e AP007 no grupo de cannabis de apreensão) e se observou o perfil anterior sendo mantido (Figura 10).

Figura 8: Análises de componentes principais (PCA) dos dados de metabolômica da cannabis, centrados pela média e escalonados pelo desvio padrão, comparando o dado bruto não-normalizado (A) com diferentes formas de normalização (soma e mediana, B e C, respectivamente) e de transformação em logaritmo de base 10 (D).

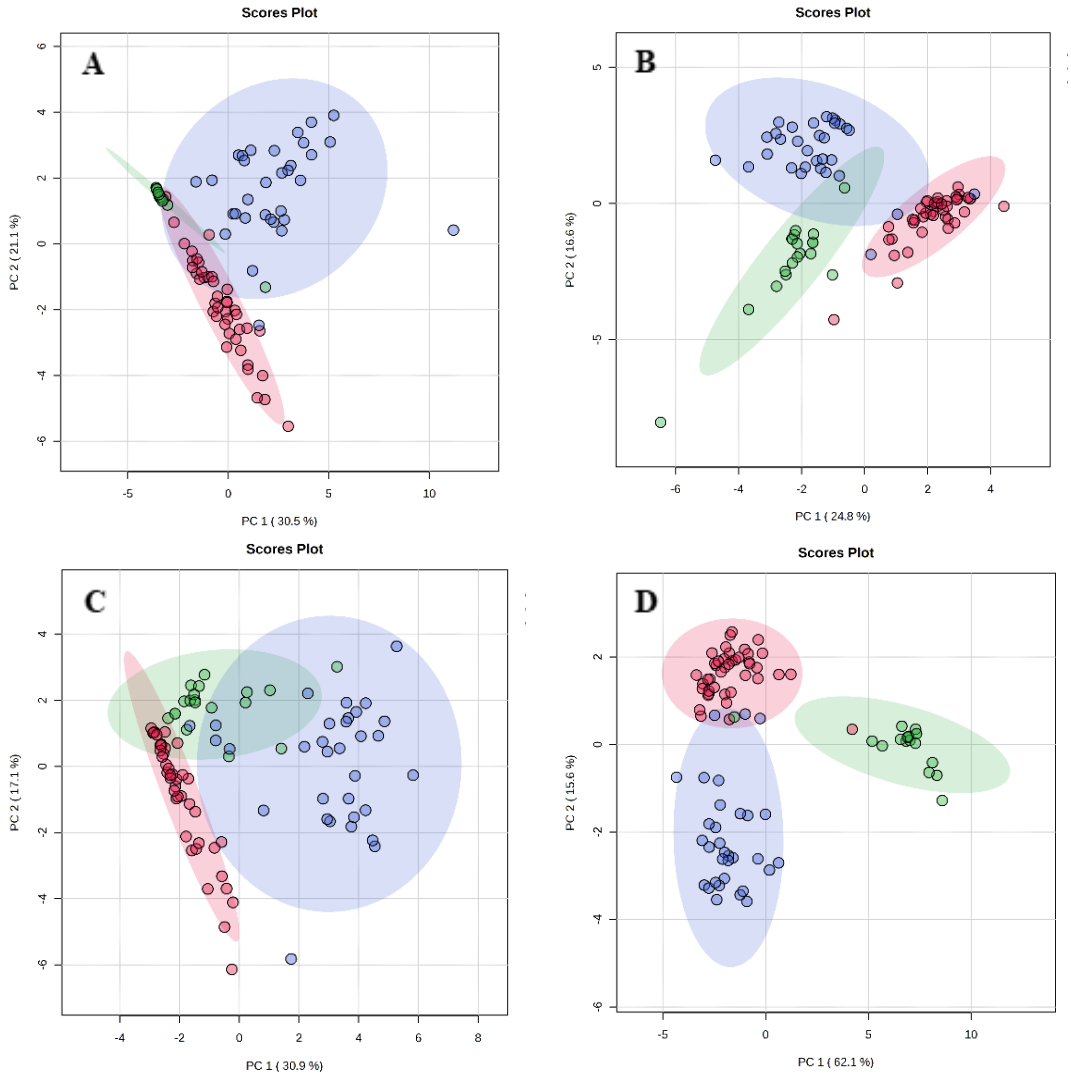


Figura 9: Análises de componentes principais dos dados somente dos grupos de cannabis medicinal e de apreensão, centrados pela média e escalonados pelo desvio padrão, sem normalização (Figura 9A) e com normalização baseada pela mediana (Figura 9B).

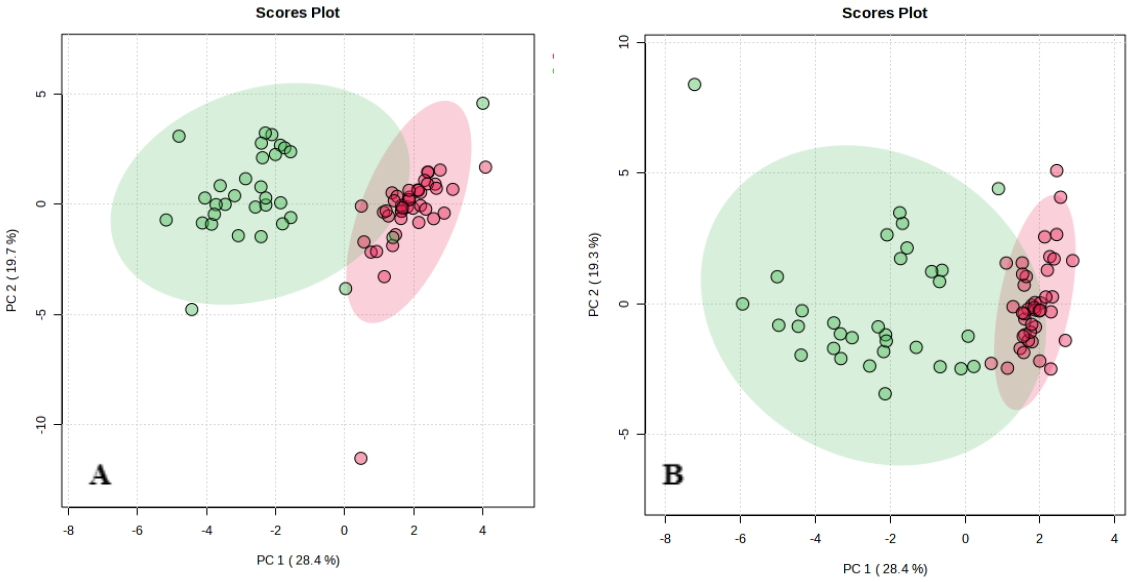
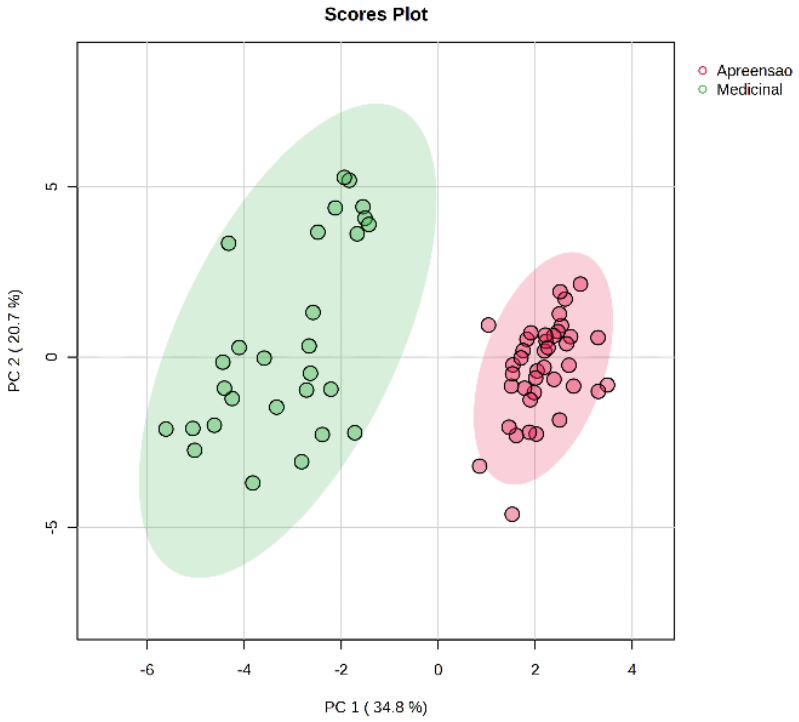


Figura 10: Análises de componentes principais dos dados de amostras aberrantes.



A normalização dos dados analíticos é um processamento que permite corrigir a diferença de sinal dos analitos detectados entre amostras a serem comparadas quantitativamente. Tal diferença tem como fontes possíveis a quantidade inicial de amostra processada, o rendimento da extração dos analitos, o volume de injeção da solução de amostra e da flutuação de resposta analítica do detector. De forma geral, a normalização é feita a partir da multiplicação das abundâncias dos sinais de uma análise por um fator de correção (VAN DEN BERG *et al*., 2006).

Entre as estratégias de normalização usadas, está a normalização baseada na amostra de partida, como a massa ou o volume de amostra processados, ou a concentração de um ou mais compostos de referência previamente conhecidos. A desvantagem dessa estratégia é que ela não permite a correção de variações de sinal oriundas de oscilações de rendimento de extração de cada composto, do efeito da presença de substâncias interferentes em diferentes quantidades e de oscilações da resposta do detector (KATAJAMAA; OREŠIČ, 2007).

Uma estratégia possível para lidar com esses fatores é o uso de padrões internos, substâncias adicionadas no início do processo de extração da amostra e com comportamento químico similar aos analitos, porém estando teoricamente ausentes nas amostras. Assim, espera-se de um bom candidato a padrão interno o rendimento no processo de extração e a resposta do detector próxima à dos analitos (como exemplo, uma substância com tempo de retenção e rendimento de ionização similares aos analitos numa análise de espectrometria de massas acoplada a cromatografia líquida), porém distinguível desses. O uso de padrões em análises sem um analito definido (como as análises de metabolômica global) é possível a partir da presunção da presença e concentração mínima de classes químicas ou de substâncias de referência na amostra de interesse, muito embora seja difícil utilizar padrões que cubram todas as diferentes classes de compostos (BLAISE *et al*., 2021; KATAJAMAA; OREŠIČ, 2007).

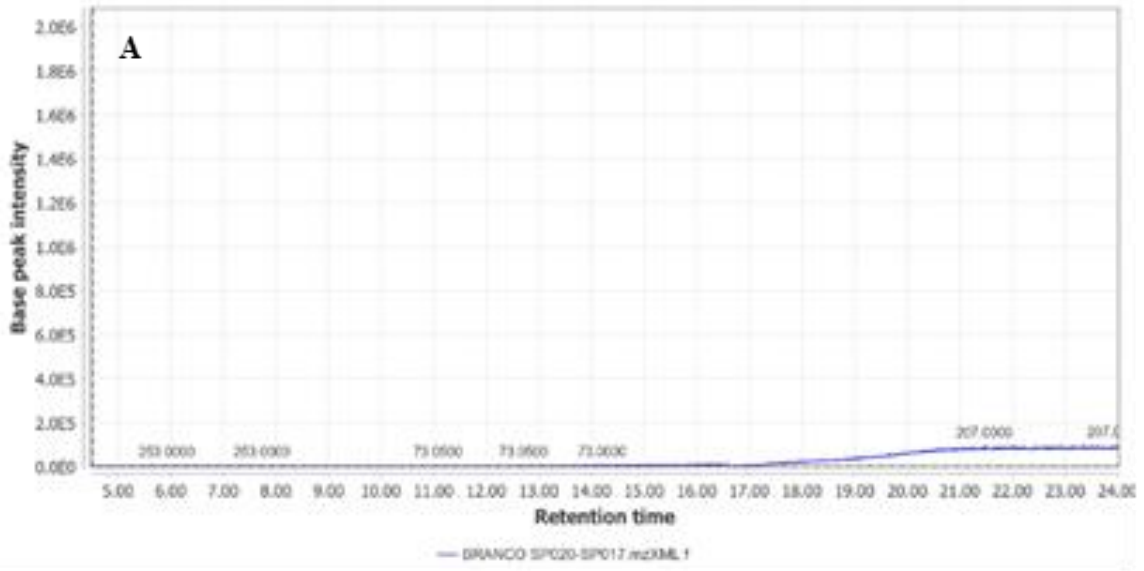
Entretanto, pode não ser possível utilizar essas duas estratégias na análise retrospectiva de dados, devido à falta de informações ou o não-uso de padrões internos na metodologia de preparo de amostra realizada. Assim, para esse tipo de caso, restam somente estratégias de normalização baseadas no dado analítico final. Duas formas de normalização usadas são as baseadas na média (também chamadas de normalização por soma ou por contagem total de íons) e baseadas na mediana. Tais estratégias possuem como premissa o fato que a contagem total das abundâncias dos diferentes analitos

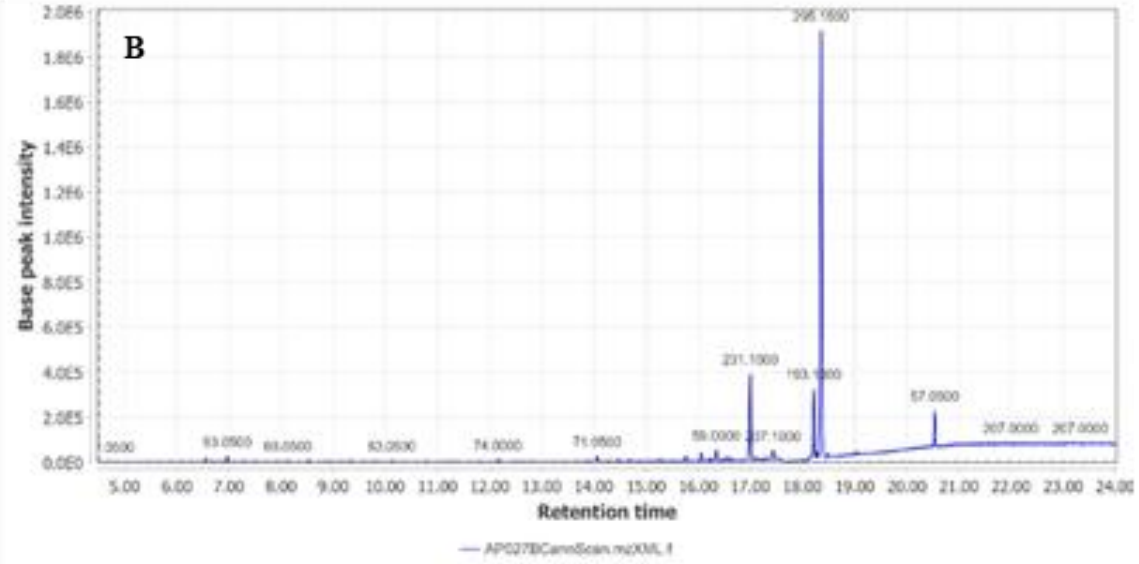
detectados deve-se manter similar entre todas as amostras (BLAISE *et al*., 2021; KATAJAMAA; OREŠIČ, 2007).

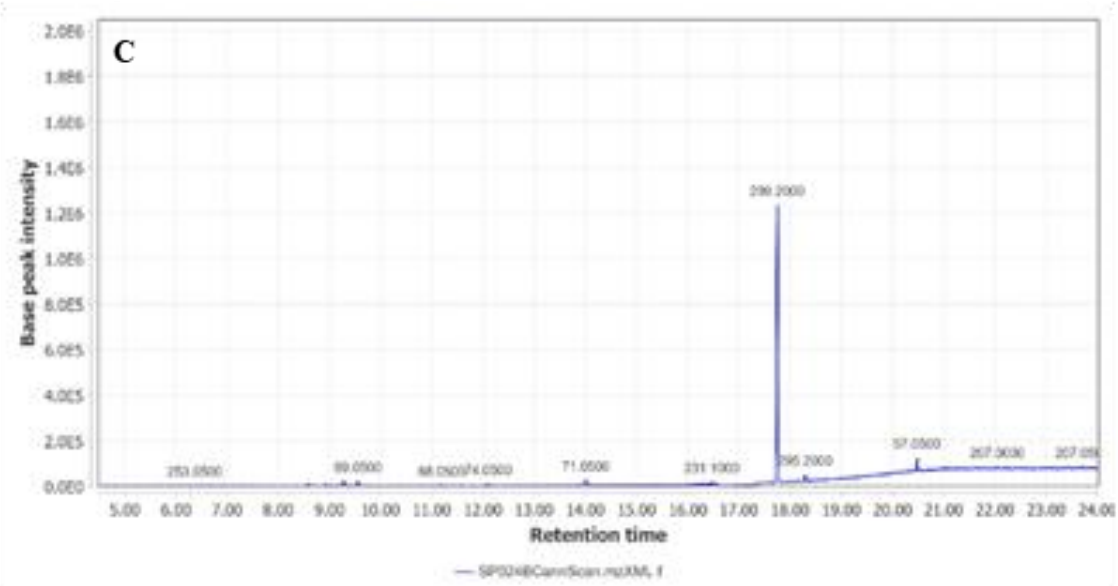
No caso das análises realizadas neste estudo, por se tratar de análises de metabolômica global já realizadas e com ausência de informações utilizáveis para normalização por amostra, os dados obtidos permitiam somente abordagens de normalização baseadas no dado final. Ao se inspecionar os cromatogramas dos três grupos analisados (medicinal, apreensão e branco), verificou-se grande variação da quantidade total de sinais analíticos (nesse caso, o número de picos cromatográficos e sua área, exemplificadas nos cromatogramas da Figura 11). Como consequência, observou-se distorções de características como a variância dos grupos na análise de componentes principal, observada principalmente na comparação entre o grupo branco com os outros dois. Nesse caso, é preferível analisar os três grupos sem o uso da normalização (BLAISE et al., 2021).

Já no caso da comparação dos grupos de apreensão e medicinal, presumido um comportamento similar entre as variáveis das amostras, foi escolhida a normalização pela mediana. Tal normalização, enquanto é aplicável sob o conceito de que a diferença entre grupos se encontra num número pequeno de variáveis, porém com abundância similar à soma de todos os valores das variáveis (nesse caso, a contagem total de íons), a mediana se apresenta menos suscetível à distorção do comportamento das amostras na presença de uma variável que representa maior parte da variação dentro da amostra (BLAISE *et al*., 2021).

Figura 11: Cromatogramas de pico base de amostras representativas dos grupos branco (A), apreensão (B) e medicinal (C).





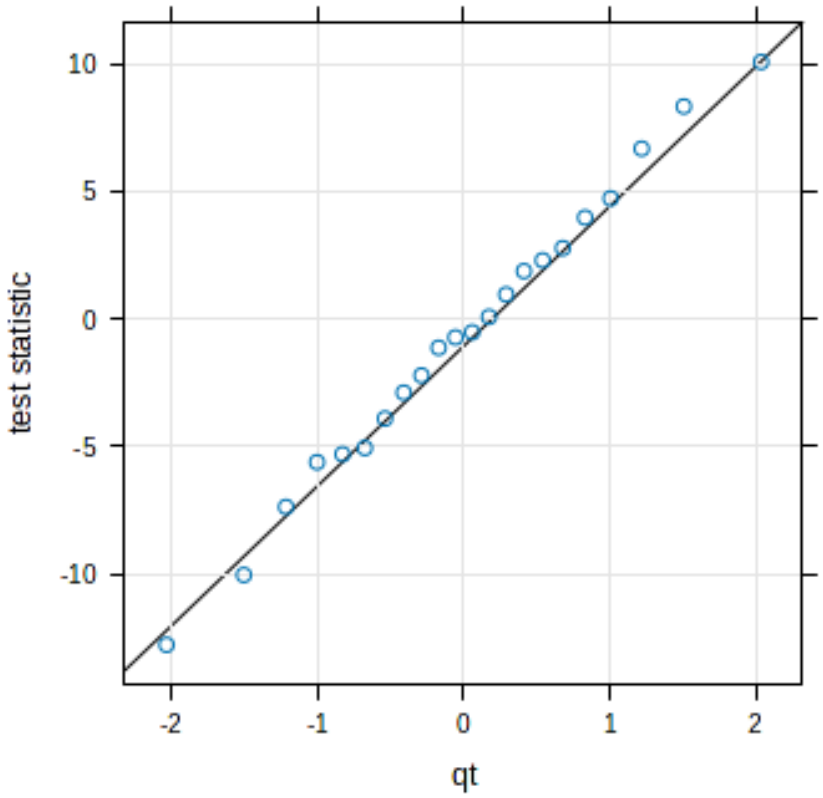


Após a aplicação da normalização, tendo em vista o objetivo de comparar o comportamento dos grupos como um todo, foi possível aplicar um critério de exclusão de amostras aberrantes a partir do teste T^2 Hotelling. Tal teste é utilizado para fazer a comparação estatística entre classes de amostras a partir de dois componentes principais, de forma que é calculada uma área ou região de confiança no gráfico referente a cada classe. O teste indica uma diferenciação estatística entre duas ou mais classes no caso das regiões não se sobreporem. Paralelamente, tal área é usada também para determinar amostras que fogem do comportamento geral da classe à qual pertencem (BLAISE *et al.*, 2021).

Outro aspecto importante da estratégia de normalização é o comportamento gaussiano (ou normal) do somatório de abundâncias de todo o conjunto analítico. Tal requisito é necessário para a aplicação de métodos estatísticos como o teste T de Student, a análise de componentes principal e a análise discriminante por mínimos quadrados parciais. Uma forma utilizada para verificar a presença de normalidade no dado é através do gráfico de quantis-quantis da distribuição normal, onde é esperada uma correlação linear dos quantis do somatório de abundâncias em comparação com os valores de quantis da distribuição normal, graficamente explicitada pela distribuição uniforme dos primeiros quantis ao longo da reta do gráfico (Figura 12) (SHLENS, 2003).

No caso dos dados do trabalho, tal gráfico foi gerado a partir da submissão dos dados na opção de análise de poder estatístico do MetaboAnalyst. Pôde-se observar uma correlação linear das diferentes abordagens de normalização das abundâncias dos grupos de cannabis medicinal e cannabis de apreensão, com maior correlação ocorrendo na normalização por mediana (dados não mostrados).

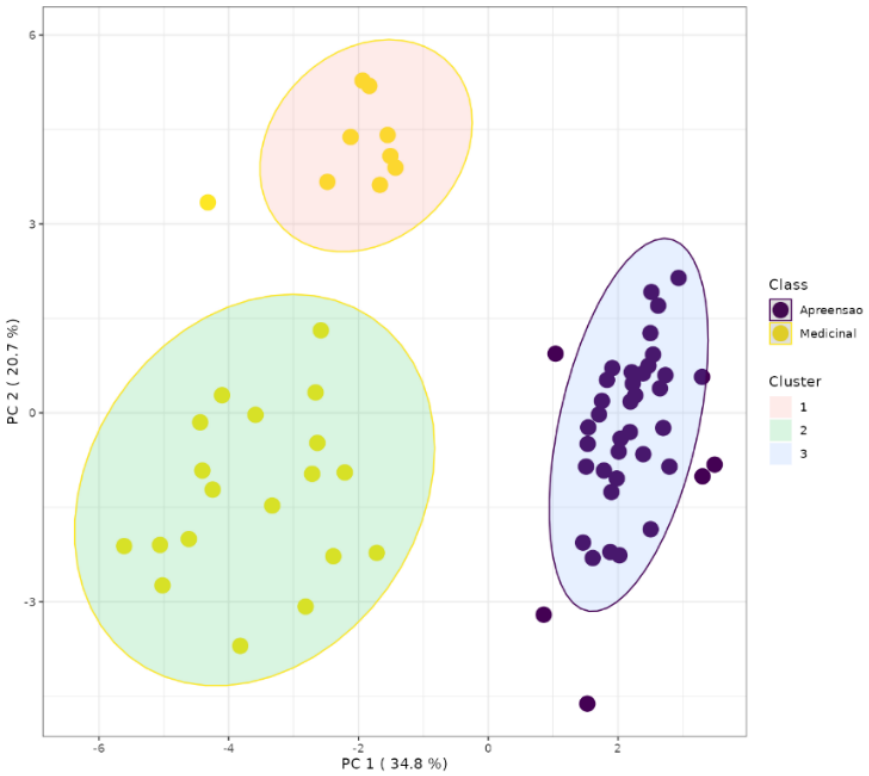
Figura 12: Análise de distribuição do conjunto de amostras de apreensão e medicinal pelo gráfico de quantis-quantis da distribuição normal, após normalização por mediana.



O gráfico obtida da clusterização por *K-means* corroborou também o agrupamento não-supervisionado dos grupos e sugeriu uma heterogeneidade entre as amostras do grupo de cannabis medicinal logo após adicionar uma terceira separação da segmentação realizada, que formaram dois clusters logo após a sua separação total do grupo de cannabis de apreensão (Figura 13), sendo tais clusters descritos no Quadro 1.

O gráfico de *Heat Map* dos grupos (Figura 14) mostrou o agrupamento natural de todas as amostras dentro de seus grupos, e a separação entre os mesmos, resultantes da análise dos 15 *features* de menor p-valor da análise de teste-t.

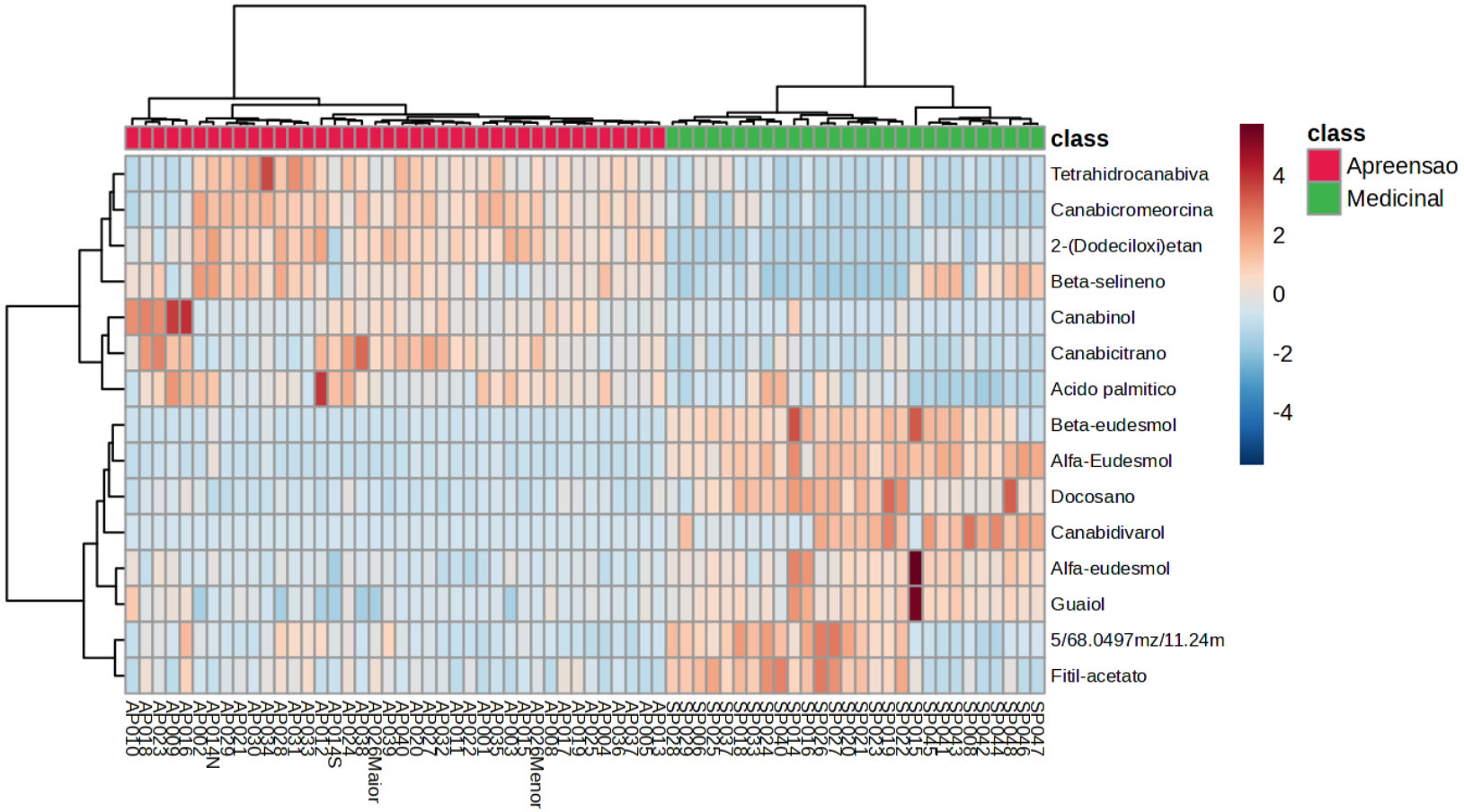
Figura 13: Agrupamento não-supervisionado formado pelo gráfico por *K-means*, gerando três clusters de amostras.



Quadro 1: Amostras de cada grupo formado pela gráfico de *K-means*.

Cluster 1	SP015, SP041, SP042, SP043, SP044, SP045, SP046, SP047, SP048
Cluster 2	SP006, SP008, SP014, SP016, SP018, SP019, SP020, SP021, SP022, SP023, SP024, SP025, SP026, SP027, SP028, SP029, SP033, SP037, SP040
Cluster 3	AP001, AP002, AP003, AP004, AP005, AP008, AP009, AP010, AP011, AP012, AP013, AP014N, AP014S, AP015, AP016, AP017, AP018, AP019, AP020, AP021, AP022, AP023, AP024, AP025, AP026Maior, AP026Menor, AP027, AP028, AP029, AP030, AP031, AP032, AP033, AP034, AP035, AP036, AP037, AP038, AP039, AP040

Figura 14: *Heat Map* formado pelos 15 *features* com maior valor no teste T, mostrando o agrupamento das amostras em seus respectivos grupos AP (apreensão) e SP (medicinal).



5.3) Análises supervisionadas

Foi realizada inicialmente a análise discriminante por mínimos quadrados parciais (*Partial Least Square Discriminant Analysis*, PLS-DA) entre os dois grupos (Figura 15). A análise do primeiro componente, mais relacionada à covariância entre os grupos, apresentou uma variância de 34,6 %, enquanto o segundo componente respondeu por uma variância de 18,6 %. O gráfico demonstra a total separação dos dois grupos.

As cinco variáveis de maior importância do modelo criado dessa análise foram os *features* 15, 18, 2, 17 e 12 (Figura 16). O modelo criado pela análise também foi avaliado quanto a sua performance preditiva, ou seja, o valor de Q2. Esse apresentou valor maior que 0,5, que é o valor desejável para validar o modelo de boa performance (dados não-mostrados).

Figura 15: Análise discriminante por mínimos quadrados parciais (PLS-DA). *Score plot* das amostras observadas nos componentes 1 e 2.

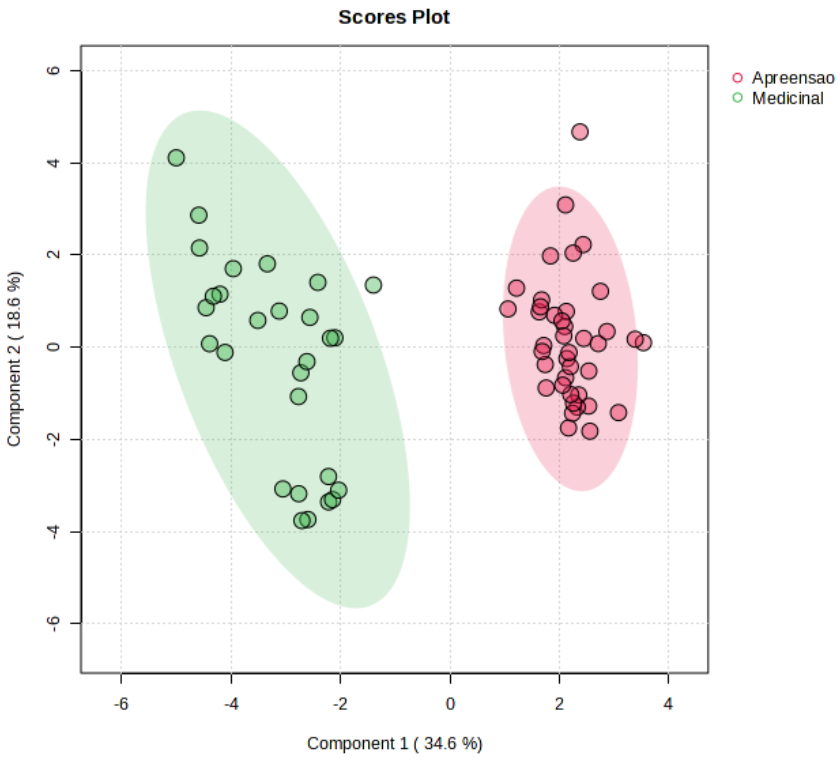
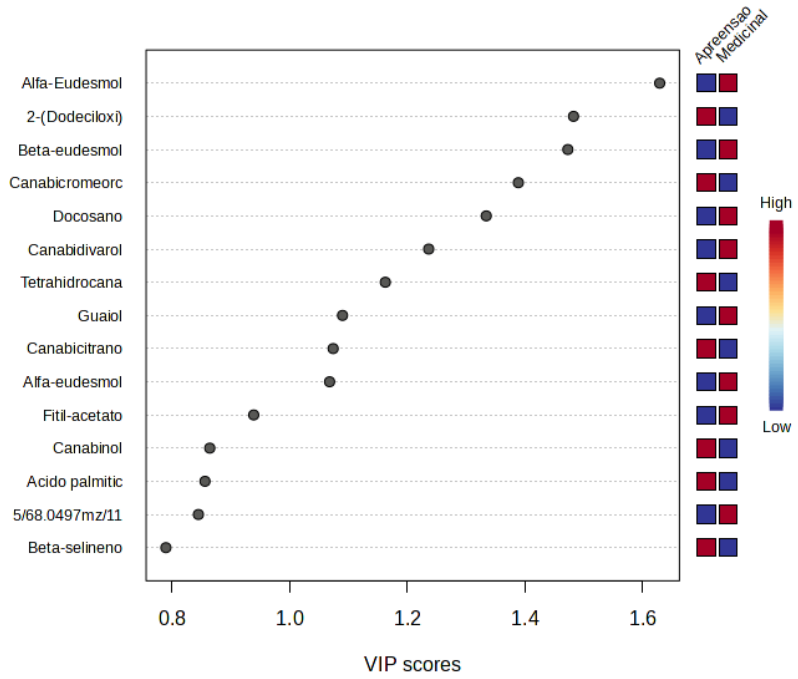


Figura 16: Variáveis de maior importância para composição do modelo de PLS-DA.



O gráfico de *volcano plot* apresentou 12 *features* de abundâncias diferentes discriminando os dois grupos analisados (Figura 17), dentre os significativos, seis foram mais abundantes nas amostras de cannabis medicinal (canabidivarol, beta-eudesmol, alfa-eudesmol, docosano, guaiol, fitil-acetato) e seis mais abundantes nas de cannabis de apreensão (2-(dedociloxi)etanol, canabicromeorcina, tetrahidrocanabivarina, canabicitrano, canabinol, verbenona). Os *features* relacionados à separação dos dois grupos evidenciados na análise de *volcano plot* são apresentados na Figura 18.

Figura 17: Gráfico de volcano plot dos grupos de cannabis medicinal e de apreensão: à esquerda são os features elevados no grupo medicinal enquanto à direita estão o do grupo de apreensão.

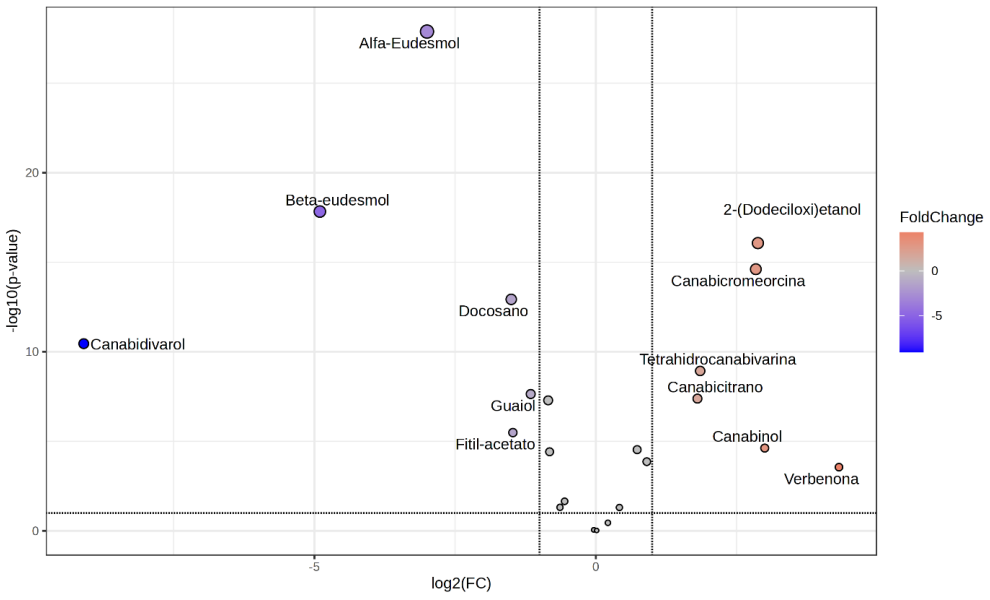
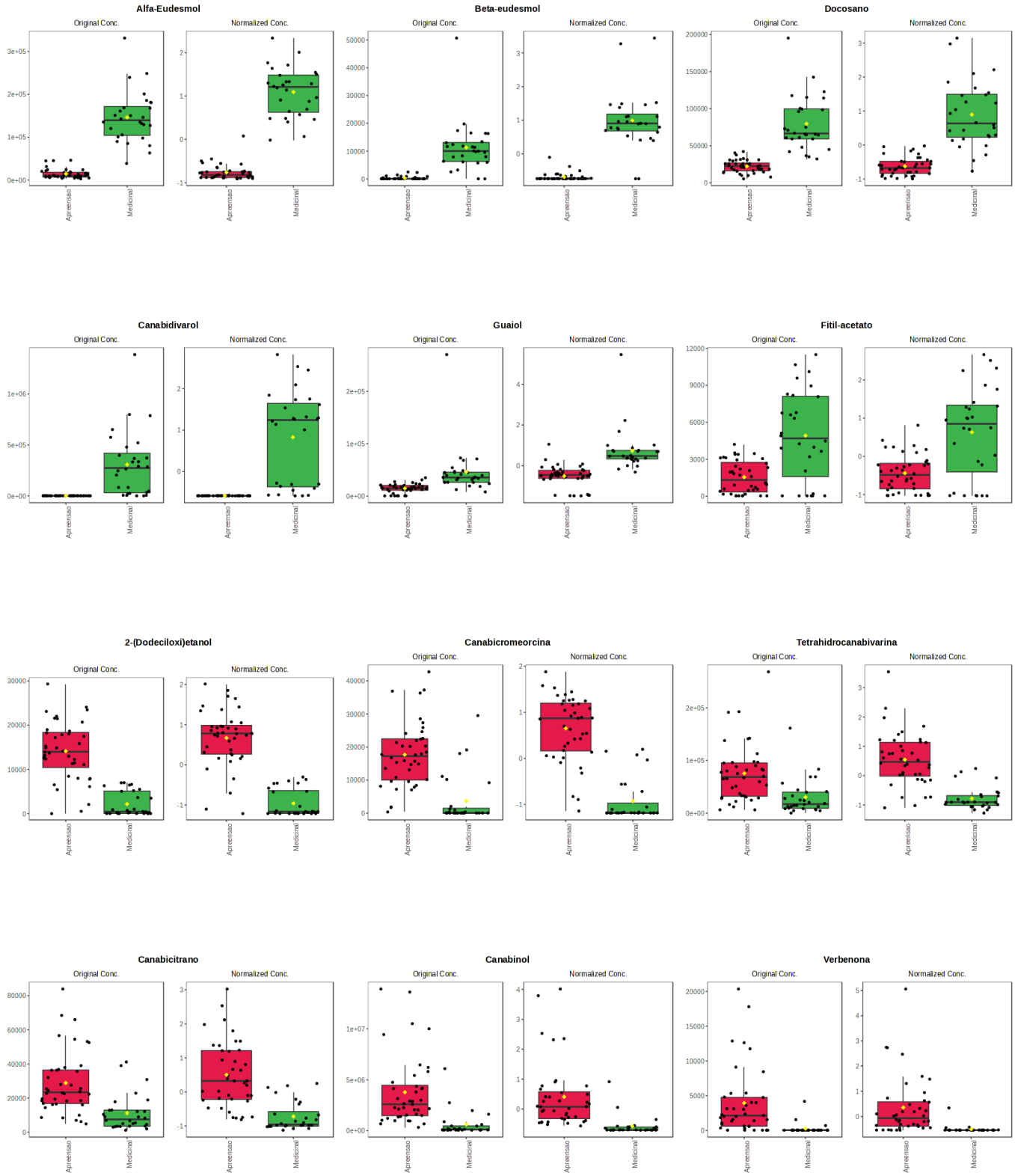


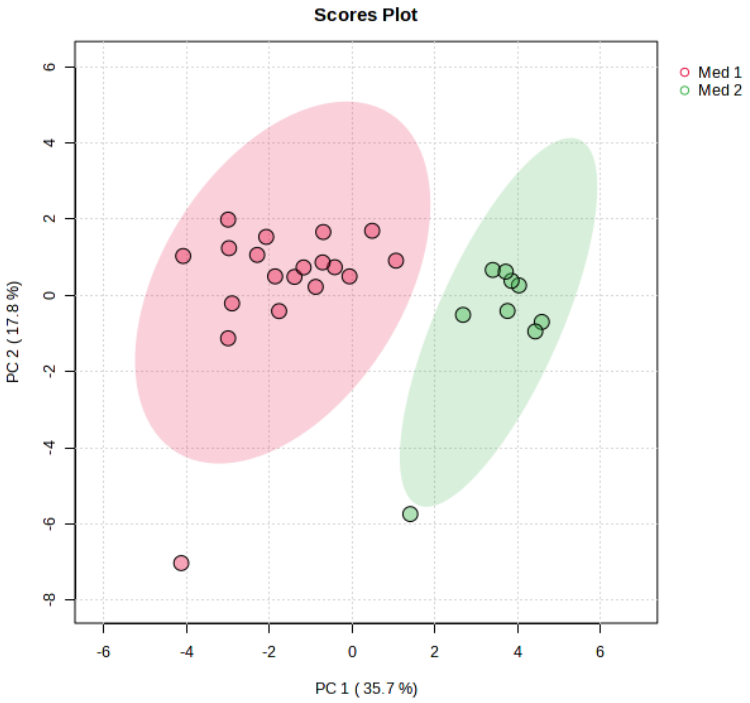
Figura 18: Os *features* diferenciais entre os dois grupos medicinal e de apreensão analisados com significância estatística.



5.4) Análises posteriores

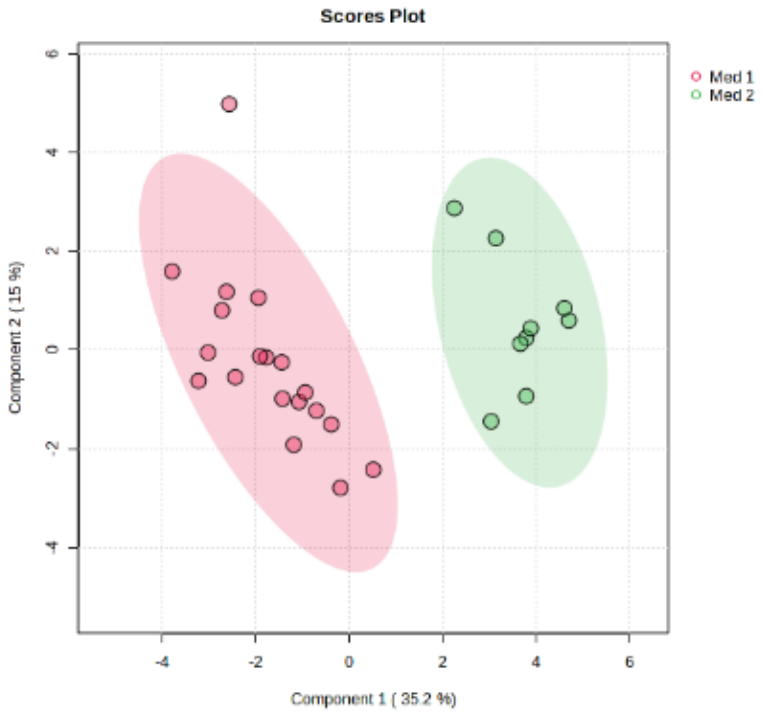
Após a análise entre grupos, foi realizada a análise dos dois subgrupos formados após a análise de *K-Means*, observados dentro do grupo de cannabis medicinal, separando-os em subgrupos com os códigos Med1 e Med2. Para tal, repetiu-se o processamento de normalização pela mediana, centralização pela média e escalonamento pelo desvio padrão. Após isso, realizou-se a análise de componentes principais para a exploração inicial desse subconjunto de dados, observando uma separação completa entre eles ao longo do componente principal 1 (Figura 19).

Figura 19: Análise de componentes principais dos subgrupos Med1 e Med2.



A PLS-DA foi utilizada como uma das análises supervisionadas, a fim de verificar quais *features* contribuíram mais para a separação dos subgrupos do grupo medicinal (Figura 20). Verificou-se forte tendência de separação dos dois grupos ao longo do componente 1 e o modelo criado para a análise apresentou boa capacidade preditiva (dados não mostrados).

Figura 20: Análise por PLS-DA dos subgrupos Med1 e Med2, apresentando o *score plot*.



Por fim, analisou-se os subgrupos no gráfico *Volcano Plot*, onde se pôde observar os *features* com maior variação relativa entre subgrupos e com diferença estatística significativa (Figura 21). No total, 4 *features* apresentaram maior abundância no subgrupo Med1 e 5 *features* apresentaram maior abundância no subgrupo Med2. Os features relacionados à separação dos dois grupos evidenciados no *volcano plot* são apresentados na Figura 22.

Figura 21: *Volcano plot* dos subgrupos Med1 e Med2, discriminando *features* distintivos.

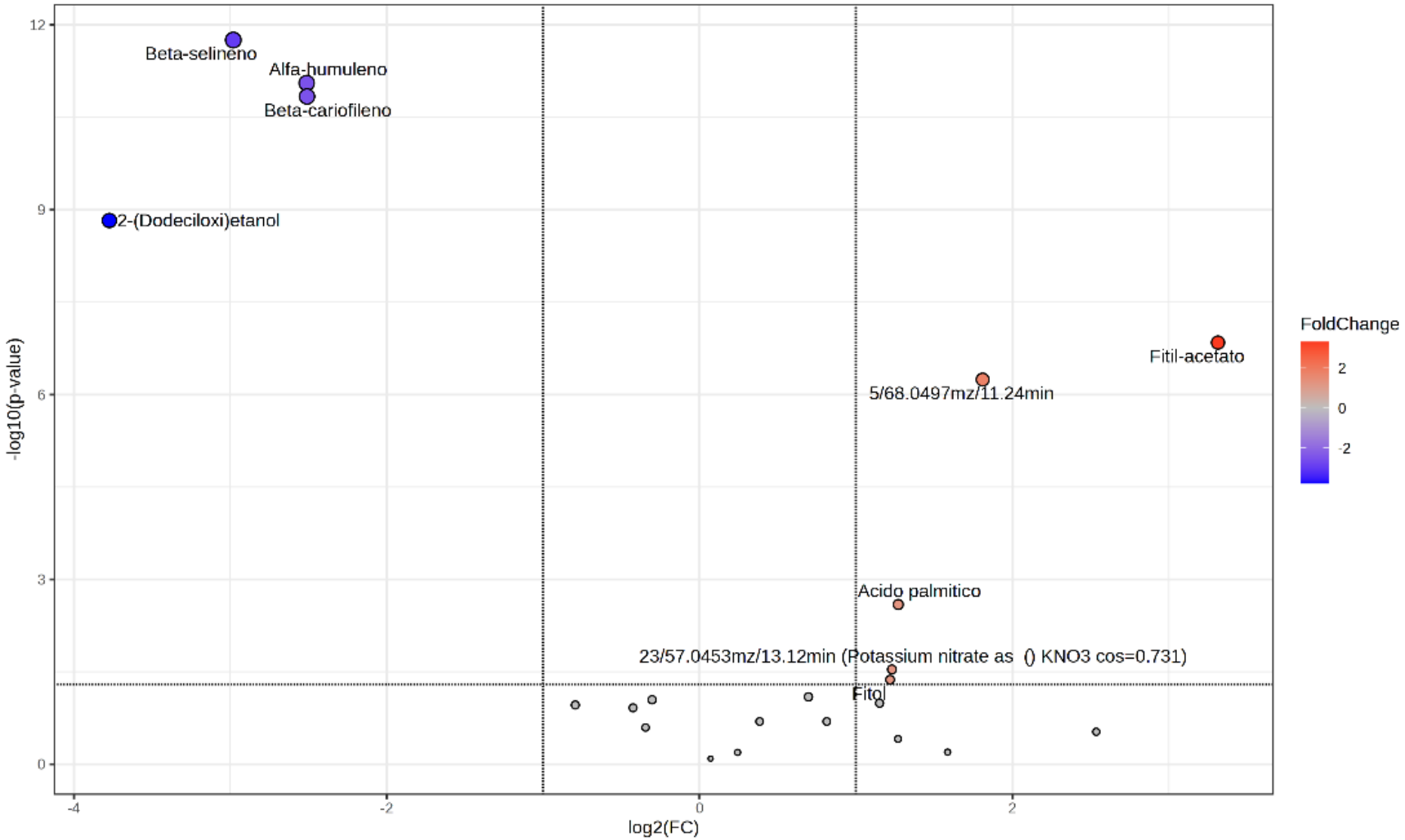
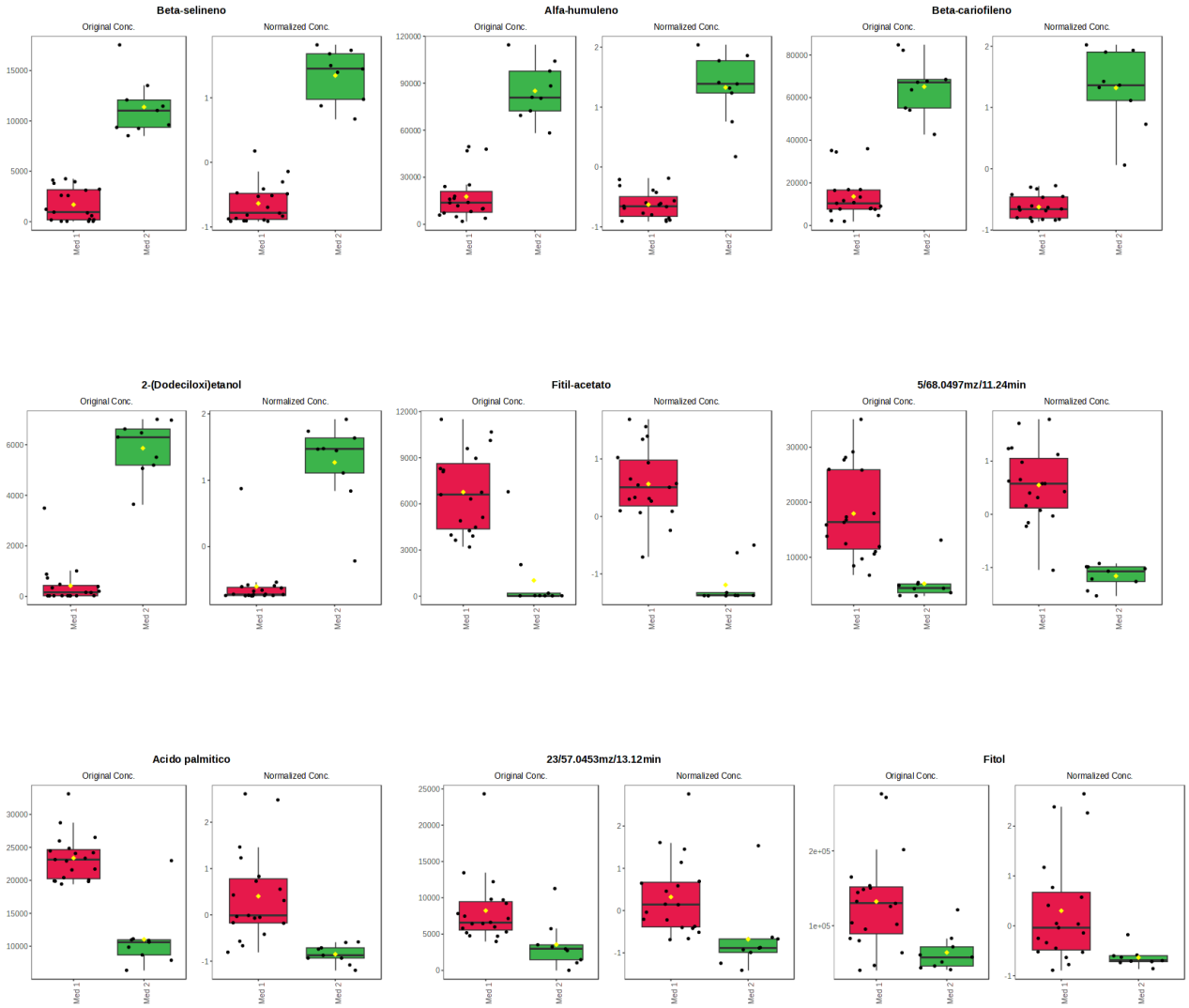


Figura 22: Os *features* discriminados entre os subgrupos Med1 e Med2 a partir do gráfico de *volcano plot*.



5.5) Perfil químico

A partir das abundâncias dos *features*, determinadas pela metodologia proposta, foi possível observar perfis químicos que discriminaram a proveniência das amostras de cannabis (Tabela 3). A primeira característica foi a presença de três sesquiterpenos nas amostras de cannabis medicinais, enquanto houve muito pouca abundância ou ausência desses compostos na maconha. O mesmo se observou para o diterpenoide acetato de fitila. Tal perfil foi observado em trabalho anterior com análises quantitativas de terpenos em cannabis ricas em CBD (CARVALHO *et al*., 2022b). Considerando que as amostras do grupo medicinal foram previamente caracterizadas como ricas em CBD por cromatografia líquida acoplada a detector de ultravioleta/visível, observa-se uma correlação possível entre os perfis fitocanabinóides e terpênicos em cannabis cultivadas em contexto medicinal.

Também se observou o fitocanabinóide canabidivarol presente nas amostras de cannabis medicinal e ausente nas amostras de maconha. Já outros fitocanabinóides foram predominantes no perfil químico da maconha. Um fitocanabinóide em questão, o canabinol, é característico de amostras ricas em THC, pois aquele é produto de degradação direta (UNODC, 2022).

A interpretação da comparação do perfil químico das duas apresentações da cannabis deste trabalho é complexa. A procedência duvidosa da maconha, o que inclui o processamento da cannabis com outras partes da planta fora a inflorescência, a possível presença de adulterantes, as condições de conservação, a exposição a intempéries diversos durante seu transporte clandestino até o ponto de venda; todos esses fatores são variáveis que não se encontram em plantas cultivadas para fins medicinais. Essas últimas apresentam um controle maior do processo de cultivo, colheita, secagem e armazenamento das inflorescências.

De forma contraintuitiva, observou-se menor variação química entre as amostras de maconha do que na de cannabis medicinal. Uma possível razão para esse achado é a condição de armazenamento dos dois grupos de cannabis, de forma que compostos mais voláteis, como os terpenos, são bem conservados nas condições controladas de produção da cannabis medicinal, enquanto há maior perda dessa classe de compostos sob as condições diversas e inadequadas de conservação e transporte da maconha. É bem descrito o impacto da conservação da cannabis no conteúdo de terpenos mais voláteis (BUENO et al., 2020).

Tabela 3: *Features* discriminados pela análise de *Volcano Plot* das amostras de cannabis e suas classificações químicas.

ID	m/z[1]	TR[2]	Anotação Final	Classe	Relação AP/SP
2	149.121	10.98	β-eudesmol	sesquiterpeno	SP
3	63.585	8.90	α -eudesmol	sesquiterpeno	SP
6	116.452	8.58	guaiol	sesquiterpeno	SP
8	82.074	11.69	acetato de fitila	diterpenoide	SP
11	84.488	18.95	docosano	hidrocarboneto	SP
18	203.065	15.17	canabidivarol	fitocanabinóide	SP
1	275.162	18.32	canabinol	fitocanabinóide	AP
13	241.154	16.06	tetrahidrocanabivarina	fitocanabinóide	AP
15	150.1	5.56	verbenona	monoterpeno	AP
16	175.069	14.46	canabicromeorcina	fitocanabinóide	AP
17	83.055	10.12	2-dodeciloxietanol	detergente	AP
21	231.108	15.76	canabicitrano	fitocanabinóide	AP

Nota: [1]m/z de referência. [2]Tempo de retenção em minutos.

Como discriminado pelo gráfico de *K-means*, foi observado o agrupamento de amostras de cannabis com dois perfis químicos distintos, cujos features encontram discriminados na Tabela 4. Uma diferença significativa foi encontrada nos compostos sesquiterpenoídes detectados (beta-selileno, beta-humuleno e alfa-cariofileno), estando todos esses mais abundantes em um grupo (Med2) do que no outro (Med1). De forma oposta, compostos diterpenóides apresentaram-se mais abundantes do grupo Med1 em relação ao Med2. Tal fato é um achado relevante, considerado a atividade sinérgica que os terpenos apresentam em conjunto com os fitocanabinóides em diferentes tratamentos (RAZ et al., 2023). Cabe avaliar as informações referentes às variedades de cannabis utilizadas e fatores relacionados ao plantio, colheita e armazenamento da inflorescência a fim de avaliar possíveis correlações das variáveis mencionadas com o perfil químico observado.

Tabela 4:Anotações de identidade dos *features* detectados na análise de amostras de cannabis.

ID	m/z[1]	TR[2]	Anotação Final	Classe	Relação Med1/Med2
5	68.05	11.24	sem notação	desconhecida	Med1
8	82.074	11.69	acetato de fitila	diterpenoide	Med1
19	71.055	14.05	fitol	diterpenoide	Med1
20	73.035	12.49	ácido palmítico	ácido graxo	Med1
22	57.045	13.12	sem notação	desconhecida	Med1
7	93.082	7.47	β-selineno	sesquiterpeno	Med2
9	93.084	6.56	β-cariofileno	sesquiterpeno	Med2
10	93.085	6.96	α -humuleno	sesquiterpeno	Med2
17	83.055	10.12	2-Dodeciloxietanol	Detergente	Med2

Nota: [1]m/z de referência. [2]Tempo de retenção em minutos. [3]Valor de cosseno da busca de similaridade espectral.

6) CONCLUSÃO

A partir do trabalho executado, foi possível aplicar uma metodologia de processamento de dados brutos e análises estatísticas em dados oriundos de cromatografia gasosa acoplada a espectrometria de massas. Essa metodologia foi proposta com uso de *softwares* abertos e gratuitos, em específico: o Mzmine, o MetaboAnalyst e a linguagem R. Essa proposta de análise foi feita com a finalidade de facilitar os pesquisadores iniciantes na análise metabolômica, sendo considerada exitosa nesse objetivo, exceto para o uso da base de dados de acesso gratuito MetaboAnalyst, que não se mostrou equivalente à base de dados NIST 11 na anotação dos *features*. A descrição dos parâmetros de cada etapa do processamento vem a servir como uma referência para análises de dados adquiridos de plataformas analíticas similares.

A verificação da conformidade dos *outputs* de cada processamento também foi usada em etapas como a de assinalamento de *features*, de anotação de identidade e das estratégias de normalização. Devido a limitações naturais dos algoritmos presentes nos *softwares*, informações errôneas e ruídos computacionais podem ser incluídas no resultado final e prejudicar a sua qualidade e a confiança. Em específico à normalização, a avaliação crítica que tal transformação do dado se faz necessária para evitar distorções das características reais dos grupos experimentais. Somado a isso, o uso da normalização mais adequada junto com dados provenientes da análise de componentes principais permitiu a identificação de amostras aberrantes que poderiam alterar a qualidade do resultado de análises estatísticas posteriores.

Avaliados e otimizados os fatores supracitados referentes ao processamento dos dados, este foi então aplicado a dados de um estudo de perfil químico de duas fontes distintas de *Cannabis sativa*. Observou-se um perfil químico característico de uma planta cultivada sob um controle necessário para o uso medicinal de qualidade e se comparou tal perfil com o de cannabis de procedência ilícita e para uso recreativo. As diferenças observadas são justificáveis frente ao conhecimento existente dos quimiotipos da cannabis, além da biossíntese dos metabólitos secundários principais da planta. Por fim, a natureza exploratória da análise metabolômica global permitiu a descrição de dois subgrupos do grupo amostral de cannabis medicinal com base na diferença química observada por ferramentas estatísticas não-supervisionadas. Tal achado possui importância em vista de evidências de que o perfil qualitativo e quantitativo de compostos terpênicos possuem efeitos sinérgicos na atividade neuronal de fitocanabinóides, além de ser usado para a classificação de diferentes cultivares de *Cannabis sativa*.

7) PERSPECTIVAS

O trabalho desenvolvido abre possibilidades para consolidar fluxos de trabalho relacionados ao processamento de dados em outras plataformas analíticas, como os já citados *softwares* XCMS e MSDIAL. Os resultados dos processamentos feitos nessas plataformas com o mesmo conjunto de dados utilizados nesse projeto podem ser comparados a fim de avaliar a qualidade daqueles, bem como as vantagens e limitações de cada plataforma. A possibilidade de usar mais de um *software* para processamento dos dados a fins de corroboração das diferentes saídas de dados destes (por exemplo, uma tabela contendo os *features* e suas abundâncias entre amostras) fornece maior robustez à análise estatística e interpretação biológica posteriores.

Outro aspecto a explorar é o desenho experimental referente ao protocolo de extração de metabólitos e de análise instrumental. Nos casos onde isso é possível, o uso de amostras contendo padrões conhecidos interlotes, misturas de amostras como um padrão intralote e amostras contendo somente o solvente inicial de extração (como um controle negativo) auxiliam na avaliação do controle de qualidade do experimento e fornece mais alternativas no tratamento dos dados, como critérios de inclusão ou exclusão de *features*. Além disso, a aleatorização das amostras, tanto na etapa de extração quanto na etapa de injeção, reduz os erros estatísticos relacionados à ordem de análise ou tempos de injeção diferentes.

No que tange à análise do perfil metabolômico global do experimento realizado, o desenvolvimento desse fluxo de trabalho pode servir de base para a análise de dados submetidos em repositórios de experimentos similares ao nosso. Tal prática traz benefícios como a comparação de dados de outros laboratórios para corroborar ou avaliar interpretações acerca do estudo e a comparação de qualidade dos dados, possibilitando a otimização da própria metodologia analítica.

Um ponto importante a se lembrar é o caráter exploratório e formador de hipóteses da metabolômica global. Enquanto tal estudo traz uma grande quantidade de informações sobre o perfil de diferentes compostos dos grupos estudados, é prudente a confirmação dos achados científicos por uma abordagem alvo, utilizando padrões para corroborar o comportamento de dos compostos com uma quantificação absoluta e para corroborar a identidade proposta pela anotação advinda da abordagem global.

8) REFERÊNCIAS BIBLIOGRÁFICAS

ABU-FARHA, M. et al. Editorial: Metabolomics in genetic and endocrinological diseases. **Frontiers in Molecular Biosciences**, v. 11, n. January, p. 1–2, 2024.

ALHARBI, Y. N. Current legal status of medical marijuana and cannabidiol in the United States. **Epilepsy and Behavior**, v. 112, p. 107452, 2020.

ALIFERIS, K. A.; BERNARD-PERRON, D. Cannabinomics: Application of Metabolomics in Cannabis (Cannabis sativa L.) Research and Development. **Frontiers in Plant Science**, v. 11, n. May, p. 1–20, 2020.

ALSEEKH, S. et al. Mass spectrometry-based metabolomics: a guide for annotation, quantification and best reporting practices. **Nature Methods**, v. 18, n. 7, p. 747–756, 2021.

ANVISA. **Registrado primeiro medicamento à base de Cannabis sativa**. Disponível em: <https://www.gov.br/anvisa/pt-br/assuntos/noticias-anvisa/2018/registrado-primeiro-medicamento-a-base-de-cannabis-sativa>. Acesso em: 7 maio. 2024.

ANVISA. **Nota técnica 35/2023: Importação de Cannabis in natura e partes da planta não será permitida**. Brasília: [s.n.].

ANVISA. **Lista de substâncias sujeitas a controle especial no Brasil**. Disponível em: <https://www.gov.br/anvisa/pt-br/assuntos/medicamentos/controlados/lista-substancias>. Acesso em: 7 maio. 2024.

ARON, A. T. et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. **Nature Protocols**, v. 15, n. 6, p. 1954–1991, 2020.

BECKER, S. et al. LC-MS-based metabolomics in the clinical laboratory. **Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences**, v. 883–884, p. 68–75, 2012.

BLAISE, B. J. et al. Statistical analysis in metabolic phenotyping. **Nature Protocols**, v. 16, n. 9, p. 4299–4326, 2021.

BORGES, R. M. et al. Guide for Chromatography Coupled To Mass Spectrometry Data Processing. **Quimica Nova**, v. 45, n. 5, p. 608–620, 2022.

BUENO, J. et al. The preservation and augmentation of volatile terpenes in cannabis inflorescence. **Journal of Cannabis Research**, v. 2, n. 1, 2020.

CAMPBELL, K.; XIA, J.; NIELSEN, J. The Impact of Systems Biology on Bioprocessing. **Trends in Biotechnology**, v. 35, n. 12, p. 1156–1168, 2017.

CANUTO, G. A. B. et al. Metabolômica: Definições, Estado-Da-Arte E Aplicações Representativas. **Quimica Nova**, v. 41, n. 1, p. 75–91, 2018.

CARVALHO, V. M. et al. Facing the Forensic Challenge of Cannabis Regulation: A Methodology for the Differentiation between Hemp and Marijuana Samples Presumptive and confirmatory methods for hemp and marijuana analysis. **Brazilian Journal of Analytical Chemistry**, v. 9, n. 34, p. 162–176, 2022a.

CARVALHO, V. M. et al. Chemical profiling of Cannabis varieties cultivated for medical purposes in southeastern Brazil. **Forensic Science International**, v. 335, p. 111309, 2022b.

CHACKO, S.; HASEEB, Y. B.; HASEEB, S. Metabolomics Work Flow and Analytics in Systems Biology. **Current Molecular Medicine**, v. 22, n. 10, p. 870–881, 2021.

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

CHANG, H. Y. et al. A Practical Guide to Metabolomics Software Development. **Analytical Chemistry**, v. 93, n. 4, p. 1912–1923, 2021.

CHEN, Y.; LI, E. M.; XU, L. Y. Guide to Metabolomics Analysis: A Bioinformatics Workflow. **Metabolites**, v. 12, n. 4, 2022.

CONNOR, J. P. et al. Cannabis use and cannabis use disorder. **Nature Reviews Disease Primers**, v. 7, n. 1, p. 1–24, 2021.

DUNN, W. et al. **Metabolomics conference workshop 12: metabolite identification and annotation, Metabolomics Conf (2017 work)**. Disponível em: <http://metabolomicssociety.org/site-map/articles/88-videos/262-2017-conference-workshop-videos-public>. Acesso em: 6 dez. 2017.

DUNN, W. B.; ELLIS, D. I. Metabolomics: Current analytical platforms and methodologies. **TrAC - Trends in Analytical Chemistry**, v. 24, n. 4, p. 285–294, 2005.

EUROPEAN MEDICINES AGENCY ICH. Q2 (R1): Validation of analytical procedures: text and methodology. **International Conference on Harmonization**, v. 2, n. May 1997, p. 1–15, 2005.

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

FISCHER, B. et al. Lower-Risk Cannabis Use Guidelines (LRCUG) for reducing health harms from non-medical cannabis use: A comprehensive evidence and recommendations update. **International Journal of Drug Policy**, v. 99, p. 103381, 2022.

FUHRER, T.; ZAMBONI, N. High-throughput discovery metabolomics. **Current Opinion in Biotechnology**, v. 31, p. 73–78, 2015.

GROSS, J. H. **Mass Spectrometry: A Textbook.** [s.l: s.n.].

HILLIG, K. W. A chemotaxonomic analysis of terpenoid variation in Cannabis. **Biochemical Systematics and Ecology**, v. 32, n. 10, p. 875–891, 2004.

HILLIG, K. W.; MAHLBERG, P. G. A chemotaxonomic analysis of cannabinoid variation in Cannabis (Cannabaceae). **American Journal of Botany**, v. 91, n. 6, p. 966–975, 2004.

HOLMES, E.; WILSON, I. D.; NICHOLSON, J. K. Metabolic Phenotyping in Health and Disease. **Cell**, v. 134, n. 5, p. 714–717, 2008.

INMETRO. Orientação sobre Validação de Métodos Analíticos: Documento de caráter orientativo (DOQ-CGCRE-008). **Coordenação Geral de Acreditação**, p. 30, 2020.

JOHNSON, C. H.; IVANISEVIC, J.; SIUZDAK, G. Metabolomics: Beyond biomarkers and towards mechanisms. **Nature Reviews Molecular Cell Biology**, v. 17, n. 7, p. 451–459, 2016.

KANO, M. Control of synaptic function by endocannabinoid-mediated retrograde signaling. **Proceedings of the Japan Academy Series B: Physical and Biological Sciences**, v. 90, n. 7, p. 235–250, 2014.

KATAJAMAA, M.; OREŠIČ, M. Data processing for mass spectrometry-based metabolomics. **Journal of Chromatography A**, v. 1158, n. 1–2, p. 318–328, 2007.

KIND, T.; FIEHN, O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. **BMC Bioinformatics**, v. 8, p. 1–20, 2007.

KIRWAN, J. A. et al. Direct infusion mass spectrometry metabolomics dataset: A benchmark for data processing and quality control. **Scientific Data**, v. 1, p. 1–13, 2014.

KOWALCZYK, T. et al. Mass spectrometry based proteomics and metabolomics in personalized oncology. **Biochimica et Biophysica Acta - Molecular Basis of Disease**, v. 1866, n. 5, p. 165690, 2020.

LIOTTA, L. A. et al. Importance of communication between producers and consumers of publicly available experimental data. **Journal of the National Cancer Institute**, v. 97, n. 4, p. 310–314, 2005.

LOCCI, E. et al. Forensic NMR metabolomics: one more arrow in the quiver. **Metabolomics**, v. 16, n. 11, p. 1–16, 2020.

LOWE, H. et al. The endocannabinoid system: A potential target for the treatment of various diseases. **International Journal of Molecular Sciences**, v. 22, n. 17, 2021.

MANDOLINO, G.; CARBONI, A. Potential of marker-assisted selection in hemp genetic improvement. **Euphytica**, v. 140, n. 1–2, p. 107–120, 2004.

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

MARTENS, L. et al. mzML - A community standard for mass spectrometry data. **Molecular and Cellular Proteomics**, v. 10, n. 1, p. R110.000133, 2011.

MEDIANI, A.; BAHARUM, S. N. Metabolomics: Challenges and Opportunities in Systems Biology Studies. **Methods in Molecular Biology**, v. 2745, p. 77–90, 2024.

MOSLEY, J. D. et al. Establishing a framework for best practices for quality assurance and quality control in untargeted metabolomics. **Metabolomics**, v. 20, n. 2, 2024.

MUSSAP, M. et al. Slotting metabolomics into routine precision medicine. **Expert Review of Precision Medicine and Drug Development**, v. 6, n. 3, p. 173–187, 2021.

NEW FRONTIER DATA. **Cannabis Capital: 2023 Industry Market Report**. Disponível em: <https://info.newfrontierdata.com/2023-cannabis-capital-report>. Acesso em: 6 maio. 2024.

NISHIUMI, S. et al. Comparative Evaluation of Plasma Metabolomic Data from Multiple Laboratories. **Metabolites**, v. 12, n. 2, 2022.

O GLOBO. **Cannabis medicinal: quais são as formas de acesso aos medicamentos no Brasil?** Disponível em: <https://oglobo.globo.com/saude/medicina/noticia/2023/12/14/cannabis-medicinal-quais-sao-as-formas-de-acesso-aos-medicamentos-no-brasil.ghtml>. Acesso em: 7 maio. 2024.

PANG, Z. et al. MetaboAnalyst 5.0: Narrowing the gap between raw spectra and functional insights. **Nucleic Acids Research**, v. 49, n. W1, p. W388–W396, 2021.

PEDRIOLI, P. G. A. et al. A common open representation of mass spectrometry data and its application to proteomics research. **Nature Biotechnology**, v. 22, n. 11, p. 1459–1466, 2004.

PLUSKAL, T. et al. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. **BMC Bioinformatics**, v. 11, 2010.

POLITI, M. et al. Direct NMR analysis of cannabis water extracts and tinctures and semi-quantitative data on Δ9-THC and Δ9-THC-acid. **Phytochemistry**, v. 69, n. 2, p. 562–570, 2008.

PROCOPIO, A. et al. Combined mechanistic modeling and machine-learning approaches in systems biology – A systematic literature review. **Computer Methods and Programs in Biomedicine**, v. 240, p. 107681, 2023.

RADWAN, M. M. et al. Cannabinoids, phenolics, terpenes and alkaloids of cannabis. **Molecules**, v. 26, n. 9, 2021.

RAZ, N. et al. Selected cannabis terpenes synergize with THC to produce increased CB1 receptor activation. **Biochemical Pharmacology**, v. 212, n. January, p. 115548, 2023.

REN, G. et al. Large-scale whole-genome resequencing unravels the domestication history of Cannabis sativa. **Science Advances**, v. 7, n. 29, p. 1–12, 2021.

REZENDE, B. et al. Endocannabinoid System: Chemical Characteristics and Biological Activity. **Pharmaceuticals**, v. 16, n. 2, 2023.

ROCHA, E. D. et al. Qualitative terpene profiling of Cannabis varieties cultivated for medical purposes [Perfil de terpenos de variedades de Cannabis cultivadas para uso medicinal]. **Rodriguesia**, v. 7, 2020.

SCHEIER, L. M.; GRIFFIN, K. W. Youth marijuana use: a review of causes and consequences. **Current Opinion in Psychology**, v. 38, p. 11–18, 2021.

SCHYMANSKI, E. L. et al. Identifying small molecules via high resolution mass spectrometry: Communicating confidence. **Environmental Science and Technology**, v. 48, n. 4, p. 2097–2098, 2014.

SHLENS, J. A tutorial on principal component analysis: derivation, discussion and singular value decomposition. **Online Note httpwww snl salk edushlenspubnotespca pdf**, v. 2, p. 1–16, 2003.

SMALL, E. .; BECKSTEAD, H. Common cannabinoid phenotypes in 350 stocks of Cannabis. **Lloydia**, v. 36, p. 144–165, 1973.

TAHIR, M. N. et al. The biosynthesis of the cannabinoids. **Journal of Cannabis Research**, v. 3, n. 1, 2021.

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

UNODC. Identification and Analysis of Cannabis and Cannabis Products National Drug Analysis Laboratories. 2022.

VAN DEN BERG, R. A. et al. Centering, scaling, and transformations: Improving the biological information content of metabolomics data. **BMC Genomics**, v. 7, p. 1–15, 2006.

VIANT, M. R. et al. Use cases, best practice and reporting standards for metabolomics in regulatory toxicology. **Nature Communications**, v. 10, n. 1, 2019.

WATSON, C. W. M. et al. A longitudinal study of cannabis use and risk for cognitive and functional decline among older adults with HIV. **AIDS and Behavior**, v. 27, n. 10, p. 3401–3413, 2023.

WILLIAMS, J.; KIRBY, H. Paper Chromatography Using Capillary Ascent. **Science**, v. 107, n. 4, p. 7–9, 1948.

WISHART, D. S. Metabolomics for investigating physiological and pathophysiological processes. **Physiological Reviews**, v. 99, n. 4, p. 1819–1875, 2019.

WISHART, D. S. et al. HMDB 5.0: The Human Metabolome Database for 2022. **Nucleic Acids**

Research, v. 50, n. D1, p. D622–D631, 2022.

XIE, H. et al. Circulating metabolic signatures of heart failure in precision cardiology. **Precision Clinical Medicine**, v. 6, n. 1, 2023.

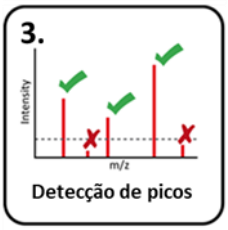
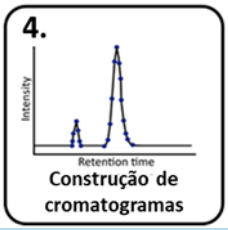
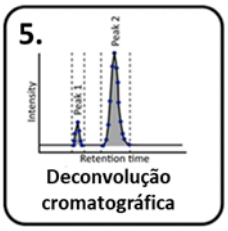
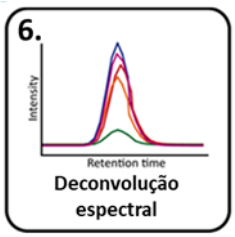
XU, M. et al. Systems biology guided by XCMS Online metabolomics. **Nat Methods**, v. 24, n. 8, p. 1246–1256, 2019.

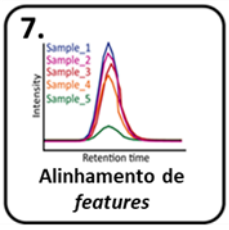
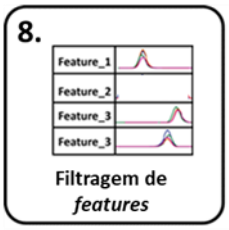
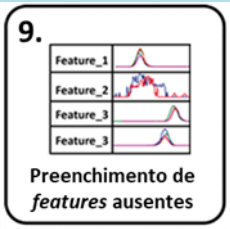
YU, H.; CHEN, Y.; HUAN, T. Computational variation: An underinvestigated quantitative variability caused by automated data processing in untargeted metabolomics. **Analytical Chemistry**, 2021.

ZHOU, J.; YIN, Y. Strategies for large-scale targeted metabolomics quantification by liquid chromatography-mass spectrometry. **Analyst**, v. 141, n. 23, p. 6362–6373, 2016.

APÊNDICE

ANEXO 1: Parâmetros das etapas de processamento realizadas no *software* Mzmine.

Processamento	Parâmetros	Observações
	Raw data methods - Raw data import	
	Raw data methods - Filtering - Crop filter Retention time: 4.00 - 24.00 min m/z: 49.7000 - 497.4501 (auto range)	Remoção de dados cromatográficos sem importância analítica, para correção de dados e menor uso de memória e de processamento.
	Raw data methods - Feature Detection - Mass Detection Mass detector: centroid Noise level: 1.0 E3	Noise level escolhido com o valor acima da intensidade dos ruídos de espectros de diferentes tempos da corrida
	Raw data methods - Feature Detection - ADAP Chromatogram Builder Min group size in # of scans: 8 Group intensity threshold: 2.0E3 Min highest intensity: 3.0E3 m/z tolerance: 0.2 m/z ou 0.0 ppm	Valores obtidos a partir da avaliação dos cromatogramas de pico base dos m/z de picos cromatográficos de menor intensidade
	Feature list methods - Feature Detection - Chromatogram Deconvolution Algorithm: Wavelets (ADAP) S/N threshold: 1 S/N estimator: Wavelets Coeff. SN min feature height: 3000 coefficient/area threshold: 400 Peak duration range: 0.04 - 0.20 RT wavelet range: 0.02 - 0.20	
	Feature list methods - Spectral Deconvolution - Multivariate Curve Resolution Deconvolution window width (min): 0.1 Retention time tolerance (min): 0.03 Minimum Number of Peaks: 3	Para seleção das listas de picos e de cromatográficos, digitar * seguido do sufixo escolhido, com a opção File list name pattern. Ex: *spec

	Feature list methods - Alignnment- ADAP Aligner (GC) Min confidence: 0.02 Retention time tolerante: 0.05 m/z tolerance: 0.1 mqz ou 0.0 ppm Score threshold: 0.6 Score weight: 0.3 Retention time similarity: Retention Time Difference	
	Feature list methods - Filtering - Feature list rows filter Minimum peaks in a row: 4 Reset the peak number ID: check	
	Feature list methods - Gap Filling - Same RT and m/z gap filler m/z tolerance: 0.2 m/z ou 0.0 ppm	
		Observar se a integração se encontra selecionando todo o pico cromatográfico ou a faixa de tempo correspondente a um pico verdadeiro. Clicar com botão direito sobre célula do feature mal integrado e selecionar Define peak manualy.
	Feature list methods - Identification - Local spectra database search MS level: 1 Minimum ion intensity: 0.0E0 Crop spectra to m/z overlap: checked Spectral m/z tolerance: 0.3 m/z ou 0.0 ppm Minimum matched signals: 20 Similarity: Composite dot-product identity Weights: NIST (GC) Minimum cos similarity: 0.7	
		Observar se espectros experimentais e teóricos são de boa qualidade (presença de picos definidos com diferenças de m/z esperadas) e se esses são similares no gráfico espelhado. Verificar se há diferenças significativas entre os espectros teóricos mais bem alinhados

ANEXO 2: Código de programação utilizado para formatação de banco de dados do NIST, na linguagem R.

```
library(readr)

filepath <-"C:\\DB_cannabis\\dd2020_copied.msp"
output <-"C:\\DB_cannabis\\dd2020_format.msp"


lines <- read_lines(filepath)

for (line in lines){
  #line <- gsub("^  ","^",line)
  line <- gsub(";\n","\n", line)
  line <- gsub(";","\n", line)
    #print (line)
  write_lines(line,output,append=T,sep="\n")
}
```

ANEXO 3: Código de programação utilizado para formatação de tabelas e cálculo de média das abundâncias de cada *feature* das triplicatas técnicas das amostras, na linguagem R.

```
library(readr)
library(dplyr)

table <- read.csv2(file.choose(),sep=",",header=F)
table <- t(table) %>% as.data.frame()

mettable <- table

values <- apply(mettable[c(-1,-2),-1],c(1,2), function (x) as.numeric(x))

headtable <- mettable[c(1,2),c(-1)]
feattable <- mettable[,1]
APvalues <- values[,c(1:126)]

APmeantable <- data.frame()

for (ii in 1:nrow(APvalues)) {
  for (i in 1:ncol(APvalues)) {
    meanvalue <- mean(APvalues[ii,i*3-2],APvalues[ii,i*3-1],APvalues[ii,i*3])
    APmeantable[ii,i] <- meanvalue
    if (i*3 >= ncol(APvalues)) {break}
  }
}

APmeantable

SPvalues <- values[,c(146:241)]

SPmeantable <- data.frame()
```

```
for (ii in 1:nrow(SPvalues)) {
  for (i in 1:ncol(SPvalues)) {
    meanvalue <- mean(SPvalues[ii,i*3-2],SPvalues[ii,i*3-1],SPvalues[ii,i*3])
    SPmeantable[ii,i] <- meanvalue
    if (i*3 >= ncol(SPvalues)) {break}
  }
}

BRValues <- values[,127:145]

headmean <- NULL

for (i in 1:126) {
  if  (i %% 3 == 0) {
    if  (is.null(headmean)) {headmean <- headtable[,i]}
    else {
      headmean <- cbind(headmean,headtable[,i])
    }
  }
}

headmean <- as.data.frame(headmean)

for (i in 127:145) {
  headmean <- cbind(headmean,headtable[,i])
}



for (i in 146:241) {
  if  (i %% 3 == 1) {
    if  (is.null(headmean)) {headmean <- headtable[,i]}
    else {
      headmean <- cbind(headmean,headtable[,i])
```

```
   }
  }
}

headmean[1,c(1:42,62:93)] <- apply(headmean[1,c(1:42,62:93)],2, function (x) strsplit(x,"C")[[1]][1] )

headmean <- as.data.frame(headmean)

headmean <- as.matrix(headmean)

valuestable <- cbind(APmeantable,BRValues,SPmeantable) %>% as.matrix()
sampletable <- rbind(headmean,valuestable)
resulttable <- cbind(feattable,sampletable)

head(resulttable)

write.table(resulttable,"output.csv",row.names = F, sep = ",",col.names = F)
```