UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
PÓS-GRADUAÇÃO EM QUÍMICA

Raphaela Cristina Cancela Marques

Estudo do metaboloma global de amostras de urina analisadas por espectrometria de massas para investigação de biomarcadores associados à COVID-19

Juiz de Fora
2026

Raphaela Cristina Cancela Marques

Estudo do metaboloma global de amostras de urina analisadas por espectrometria de massas para investigação de biomarcadores associados à COVID-19

Dissertação apresentada ao Programa de Pós-Graduação em Química, da Universidade Federal de Juiz de Fora como requisito parcial a obtenção do grau de Mestre em Química. Área de concentração: Química Analítica.

Orientador: Prof. Dr. Marcone Augusto Leal de Oliveira

Coorientadora: Profa. Dra. Adriana Nori de Macedo

Juiz de Fora

2026

Ficha catalográfica elaborada através do programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Marques, Raphaela Cristina Cancela.
Estudo do metaboloma global de amostras de urina analisadas por espectrometria de massas para investigação de biomarcadores associados à COVID-19 / Raphaela Cristina Cancela Marques. -- 2026.
76 p. : il.

Orientador: Marcone Augusto Leal de Oliveira
Coorientadora: Adriana Nori de Macedo
Dissertação (mestrado acadêmico) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas. Programa de Pós-Graduação em Química, 2026.

1. COVID-19. 2. Metabolômica global. 3. Urina. 4. Espectrometria de massas . 5. Quimiometria. I. Oliveira, Marcone Augusto Leal de, orient. II. Macedo, Adriana Nori de, coorient. III. Título.

Raphaela Cristina Cancela Marques

Estudo do metaboloma global de amostras de urina analisadas por espectrometria de massas para a investigação de biomarcadores associados aà Covid-19

Dissertação apresentada ao Programa de Pós-Graduação em Química da Universidade Federal de Juiz de Fora como requisito parcial à obtenção do título de Mestre em Química. Área de concentração: Química.

Aprovada em 20 de fevereiro de 2026.

BANCA EXAMINADORA

Prof. Dr. Marcone Augusto Leal de Oliveira - Orientador
Universidade Federal de Juiz de Fora

Profa. Dra. Adriana Nori de Macedo - Coorientadora
Universidade Federal de Minas Gerais

Dra. Andrea Tedesco Faccio
Grupo Fleury

Profa. Dra. Daniela Aparecida Chagas de Paula
Universidade Federal de Juiz de Fora

Juiz de Fora, 13/02/2026.

Documento assinado eletronicamente por **Andréa Tedesco Faccio**, **Usuário Externo**, em 20/02/2026, às 16:14, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Daniela Aparecida Chagas de Paula**, **Professor(a)**, em 20/02/2026, às 16:15, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcone Augusto Leal de Oliveira**, **Professor(a)**, em 23/02/2026, às 15:07, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Adriana Nori de Macedo**, **Usuário Externo**, em 05/03/2026, às 12:37, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no Portal do SEI-Ufjf (www2.ufjf.br/SEI) através do ícone Conferência de Documentos, informando o código verificador **2875910** e o código CRC **B400B1F8**.

AGRADECIMENTOS

Agradeço, antes de mais nada, a todos os guias espirituais que cuidaram dos meus caminhos e me deram força e clareza para seguir neles.

Aos meus pais, Cida e Cidinho, que dedicaram a vida a tentar me oferecer todas as oportunidades que lhes foram negadas. Se hoje estou aqui e desejo ir mais longe, é pelo amor da minha mãe e pela coragem de meu pai. Vocês são meus professores mais sábios e meus exemplos mais inspiradores.

Ao meu irmão, por suportar minha ausência.

Aos meus amigos do Grupo de Química Analítica e Quimiometria, especialmente ao Luiz Henrique, Patrícia, Bruna, e Fernanda, que me receberam com carinho e tornaram este processo muito mais leve, por tudo compartilhado dentro e fora do laboratório.

Ao Manu, que tem sido minha maior alegria.

À Dona Iolanda, pelo imenso carinho e amizade.

Ao meu orientador, Marcone, por todas as oportunidades, ensinamentos e por balancear tão bem meu pessimismo.

À minha coorientadora, Adriana, uma das pessoas mais didáticas e gentis que conheço, que com muita competência me ensinou e se dedicou ao meu trabalho.

A todos os membros que aceitaram compor minha banca examinadora por todas as contribuições.

À Universidade Federal de Juiz de Fora, ao programa de Pós-Graduação em Química e às agências de fomento, CNPq, CAPES e FAPEMIG pelo apoio financeiro e institucional.

RESUMO

A pandemia de COVID-19, causada pelo vírus SARS-CoV-2, representou um dos maiores desafios à saúde pública global, demandando estratégias de diagnóstico eficazes para o controle da transmissão. Embora o teste RT-PCR seja o padrão-ouro, este possui uma amostragem invasiva, o que abre espaço para a busca de matrizes alternativas. A urina destaca-se como uma amostra biológica promissora dada sua coleta não invasiva, facilidade de manuseio e composição representativa do estado de saúde do indivíduo. Nesse contexto, este trabalho aplicou a abordagem metabolômica global para investigar metabólitos associados à COVID-19 em amostras utilizando cromatografia líquida de alta eficiência acoplada à espectrometria de massas (HPLC-MS). Foram utilizadas 100 amostras de voluntários, divididas entre Grupo Teste (38 positivos para COVID-19, confirmado por RT-PCR) e Grupo Controle (62 negativos). As amostras passaram por inativação viral e precipitação proteica antes de serem analisadas nos modos de ionização positivo e negativo. O processamento de dados foi realizado nas plataformas XCMS Online e MetaboAnalyst 6.0. Após a filtragem de dados provenientes de brancos analíticos e de *molecular features* com baixa reprodutibilidade (RSD > 30%), os dados foram normalizados pela creatinina e pela mediana, transformados logaritmicamente e escalonados por Pareto. Foram aplicados os modelos PCA e PLS-DA. Na análise por PCA, observou-se considerável sobreposição entre os grupos, indicando ausência de separação natural nos dados. Já o PLS-DA apresentou indícios de sobreajuste, sugerindo capacidade preditiva limitada. A análise de *Fold Change* (FC > 2) identificou pares *m/z* e RT com variações de intensidade relevantes entre os grupos, entretanto, ao integrar o critério de significância estatística (p-valor < 0.05), não foi observada diferenciação significativa entre os grupos, sugerindo uma limitação na distinção estatística sob os protocolos testados. Análises de agrupamento reforçaram a ausência de separação clara, indicando que a similaridade entre as amostras pode ser resultado de fatores individuais que fugiram do escopo deste trabalho. Conclui-se que, embora a urina seja uma matriz viável, a complexidade metabólica e as variações individuais representam desafios para a definição de metabólitos úteis para diagnóstico, destacando a necessidade de estudos futuros que considerem variáveis clínicas adicionais para uma visão mais abrangente da resposta metabólica à COVID-19.

Palavras-chave: COVID-19, metabolômica global, urina, espectrometria de massas quimiometria.

ABSTRACT

The COVID-19 pandemic, caused by the SARS-CoV-2 virus, represented one of the greatest challenges to global public health, requiring effective diagnostic strategies to control transmission. Although the RT-PCR test is the gold standard, it involves invasive sampling, which leaves room for the search for alternative matrices. Urine stands out as a promising biological sample due to its non-invasive collection, ease of handling, and representative composition of an individual's health status. In this context, this study applied the global metabolomics approach to investigate metabolites associated with COVID-19 in samples using high-performance liquid chromatography coupled with mass spectrometry (HPLC-MS). One hundred samples from volunteers were used, divided into a Test Group (38 positive for COVID-19, confirmed by RT-PCR) and a Control Group (62 negative). The samples underwent viral inactivation and protein precipitation prior to analysis in both positive and negative ionization modes. Data processing was performed on the XCMS Online and MetaboAnalyst 6.0 platforms. After filtering signals from analytical blanks and molecular features with low reproducibility (RSD > 30%), the data were normalized by creatinine and median, logarithmically transformed, and Pareto scaled. PCA and PLS-DA models were applied. PCA analysis showed considerable overlap between groups, indicating the absence of natural separation in the data. PLS-DA exhibited signs of overfitting, suggesting limited predictive performance. Fold Change analysis (FC > 2) identified m/z and retention time pairs with relevant intensity variations between groups; however, when combined with statistical significance criteria (p-value < 0.05), no significant differences were observed, indicating limited statistical discrimination under the tested protocols. Cluster analyses reinforced the absence of clear separation, suggesting that the similarity between samples may be the result of individual factors that were beyond the scope of this work. It is concluded that, although urine presents a viable matrix, metabolic complexity and individual variability pose challenges for the definition of metabolites suitable for diagnostic purposes, highlighting the need for future studies that incorporate additional clinical variables to achieve a more comprehensive understanding of the metabolic response to COVID-19.

Keywords: COVID-19, global metabolomics, urine, mass spectrometry, chemometrics.

LISTA DE ILUSTRAÇÕES

Figura 1 - Exemplo da organização da estrutura da tabela necessária para o processamento de dados subsequente……25

Figura 2 - Sobreposição dos cromatogramas de íons totais das amostras de controle de qualidade nos modos de ionização negativo e positivo, respectivamente……26

Figura 3 - Pressão do sistema cromatográfico em cada uma das corridas no modo de ionização negativo e positivo, respectivamente……...27

Figura 4 - Sobreposição de Cromatogramas de Íons Totais dos conjuntos N1 e N3, antes e depois alinhamento de tempos de retenção gerado pelo XCMS Online……..28

Figura 5 – Sobreposição de Cromatogramas de Íons Totais dos conjuntos P1 e P3, antes e depois alinhamento de tempos de retenção gerado pelo XCMS Online……..29

Figura 6 – Alinhamento de tempos de retenção gerado pelo XCMS Online para os conjuntos N1, N3, P1 e P3. Os pontos representam a diferença entre os tempos de retenção antes e após a correção…….30

Quadro 1 - Estratégias de processamento aplicadas aos dados.……..35

Figura 7 - Distribuição de dados antes e após normalização, transformação e escalonamento……..37

Figura 8 - Representação tridimensional (PC1 vs. PC2 vs. PC3) dos escores dos componentes principais do conjunto N2 demonstrando a clusterização das amostras de controle de qualidade……...39

Figura 9 - Gráfico de variância explicada pelos oito primeiros componentes principais (PCs). Dados extraídos do processamento de N2: PC1 (18.4%), PC2 (11.8%), PC3 (7.3%), PC4 (6.3%), PC5 (5.4%), PC6 (4.8%), PC7 (4.2%) e PC8 (3.3%).....40

Figura 10 - Representação bidimensional (PC1 vs. PC2) dos escores dos componentes principais……..42

Figura 11 - Gráfico de Fold Change com limiar de diferenciação igual a 2……...45

Figura 12 – Análise de agrupamento por K-means (k=2) das amostras de urina. Disposição dos clusters no espaço multivariado (PC1 vs. PC2).....47

Figura 13 - Dendrograma gerado por análise hierárquica do conjunto N1…….50

Figura 14 - Dendrograma gerado por análise hierárquica do conjunto N2…….51

Figura 15 - Dendrograma gerado por análise hierárquica do conjunto N3…….52

Figura 16 - Dendrograma gerado por análise hierárquica do conjunto P1……..53

Figura 17 - Dendrograma gerado por análise hierárquica do conjunto P2……..54

Figura 18 - Dendrograma gerado por análise hierárquica do conjunto P3……..55

Figura 19 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 1 e 2, respectivamente……..62

Figura 20 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 3 e 4, respectivamente……..62

Figura 21 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 5 e 6, respectivamente……..63

Figura 22 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 7 e 8, respectivamente……..63

Figura 23 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 9 e 10, respectivamente……64

Figura 24 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 11 e 12, respectivamente…..64

Figura 25 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 13 e 14, respectivamente…..65

Figura 26 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 15 e 16, respectivamente…..65

Figura 27 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 17 e 18, respectivamente…..66

Figura 28 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 19 e 20, respectivamente…..66

Figura 29 - PLS-DA aplicado ao conjunto N1……68

Figura 30 - PLS-DA aplicado ao conjunto N2……68

Figura 31 - PLS-DA aplicado ao conjunto N3……69

Figura 32 - PLS-DA aplicado ao conjunto P1……69

Figura 33 - PLS-DA aplicado ao conjunto P2……70

Figura 34 - PLS-DA aplicado ao conjunto P3……70

LISTA DE TABELAS

Tabela 1 - Número de features detectadas nos modos de ionização positivo e negativo utilizando parâmetros otimizados e os parâmetros padrão do XCMS Online.……31

Tabela 2 - Impacto da filtragem no número de features e valores faltantes.……...……33

Tabela 3 - Percentual de variância explicada pelas componentes principais.…….41

Tabela 4 - Métricas de desempenho dos modelos PLS-DA……43

Tabela 5 - Características do conjunto amostral para estudos relacionados à COVID-19…..49

LISTA DE ABREVIATURAS E SIGLAS

ANVISA	Agência Nacional de Vigilância Sanitária
cDNA	DNA complementar
CEP	Comitê de Ética em Pesquisa
COVID-19	Doença infecciosa causada pelo coronavírus SARS-CoV-2
DNA	Ácido desoxirribonucleico
DOE	Planejamento experimental (*Design of Experiments*)
Dual AJS-ESI	Tecnologia *Dual Jet Stream* aplicada à ionização por eletrospray
ESI (+/-)	Ionização por eletrospray nos modos positivo e negativo
FC	*Fold Change*
FDR	Taxa de falsa descoberta (*False Discovery Rate*)
GQAQ	Grupo de Química Analítica e Quimiometria
HCA	Análise Hierárquica de Agrupamentos (*Hierarchical Cluster Analysis*)
HCoV	Coronavírus humanos associados principalmente a infecções respiratórias leves
HPLC-ESI-QTOF	Cromatografia líquida de alta eficiência acoplada à espectrometria de massas com ionização por eletrospray e analisador por tempo de voo (*High performance liquid chromatography combined to mass spectrometry with electrospray ionization and time-of-flight analyzer*)
HU-UFJF	Hospital Universitário da Universidade Federal de Juiz de Fora
IPO	*Isotopologue Parameter Optimization*
LC-MS	Cromatografia líquida acoplada à espectrometria de massas
m/z	Razão massa/carga
MERS-CoV	Coronavírus causador da síndrome respiratória do Oriente Médio
ML	Aprendizado de Máquina (*Machine Learning*)
OMS (WHO)	Organização Mundial da Saúde (*World Health Organization*)
OPAS	Organização Pan-Americana da Saúde

PCA	Análise de Componentes Principais (*Principal Component Analysis*)
PCs	Componentes Principais (*Principal Components)*
PLS-DA	Análise Discriminante por Mínimos Quadrados Parciais (*Partial Least Squares – Discriminant Analysis*)
ppm	Partes por milhão
Q^2	Parâmetro de validação cruzada que expressa a capacidade preditiva de um modelo
QC	Amostra de controle de qualidade (*Quality Control*)
R^2	Coeficiente de determinação
RNA	Ácido ribonucleico
RSD	Desvio padrão relativo (*Relative Standard Deviation*)
RT-PCR	Reação em Cadeia da Polimerase precedida de Transcrição Reversa (*Reverse Transcription Polymerase Chain Reaction*)
SARS-CoV	Cornovírua causador da síndrome respiratória aguda grave (SARS)
SARS-CoV-2	Coronavírus 2 da síndrome respiratória aguda grave, agente etiológico da COVID-19
s/n	Relação sinal/ruído (*signal-to-noise*)
TCLE	Termo de Consentimento Livre e Esclarecido
TIC	Cromatograma de íons totais (*Total Ion Chromatogram*)
UFJF	Universidade Federal de Juiz de Fora

SUMÁRIO

1 INTRODUÇÃO....13
1.1 COVID-19....13
1.2 DIAGNÓSTICO....14
1.3 METABOLÔMICA....16
1.4 CROMATOGRAFIA LÍQUIDA ACOPLADA À ESPECTROMETRIA DE MASSAS APLICADA A METABOLÔMICA....17
1.5 TRATAMENTO DE DADOS....18
2 OBJETIVOS....20
2.1 OBJETIVO GERAL....20
2.2 OBJETIVOS ESPECÍFICOS....20
3 PROCEDIMENTO EXPERIMENTAL....21
3.1 CONJUNTO DE AMOSTRAS....21
3.2 ANÁLISE INSTRUMENTAL....22
3.3 PRÉ-PROCESSAMENTO....23
3.3.1 SELEÇÃO DOS PARÂMETROS DE PROCESSAMENTO DOS DADOS....23
3.3.2 XCMS Online....24
4 RESULTADOS E DISCUSSÕES....27
4.1 ANÁLISE VISUAL....27
4.2 DETECÇÃO E ALINHAMENTO DE FEATURES....29
4.3 FILTRAGEM....34
4.4 NORMALIZAÇÃO....36
4.5 ESTATÍSTICA....39
4.5.1 ANÁLISE DE COMPONENTES PRINCIPAIS....40
4.5.2 PARTIAL LEAST SQUARES - DISCRIMINANT ANALYSIS....44
4.5.3 FOLD CHANGE....46
4.5.4 ANÁLISE DE AGRUPAMENTO....48
5 CONSIDERAÇÕES FINAIS....58
REFERÊNCIAS....59
APÊNDICE A - Comparação de representações gráficas dos procedimentos de processamento testado....64
APÊNDICE B - Modelos PLS-DA....70
ANEXOS....74
ANEXO I – Créditos de disciplinas necessárias para integralização do currículo....74
ANEXO II – Carga horária dedicada a Atividade Prática Docente (Tutoria)....74

1 INTRODUÇÃO

Em 2019, foi registrado o primeiro caso de COVID-19, uma enfermidade altamente transmissível causada pelo novo coronavírus que originou um dos maiores desafios à saúde pública do século XXI. A rápida difusão do vírus gerou efeitos globais e fez com que a OMS declarasse estado de emergência de saúde pública global em janeiro de 2020 – status que se manteve até maio de 2023. Uma das principais formas de enfrentamento, foi o diagnóstico e isolamento de sujeitos infectados, entretanto, a testagem para esta enfermidade parte da utilização de métodos invasivos de coleta, que podem gerar desconforto ao paciente e comprometer a qualidade da amostra. Além disso, os testes possuem eficácia variável de acordo com o estágio da infecção, influenciada por fatores como: carga viral, especificidade e acurácia dos métodos (Santos, 2021). Por essa razão, torna-se necessário a busca por novas matrizes e métodos que permitam contornar estes entraves.

1.1 *COVID-19*

A nomenclatura estabelecida pela Organização Mundial de Saúde (OMS), define a denominação do vírus como coronavírus-2 da síndrome respiratória aguda grave (SARS-CoV-2) e a doença infecciosa como COVID-19 (WHO, 2020). O termo "novo" é empregado por se tratar da variação de um vírus descrito pela primeira vez na década de 1960 (Souza *et al.*, 2021). No total, além do SARS-CoV-2, foram identificados seis tipos de coronavírus que afetam os seres humanos: HCoV-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU, o SARS-CoV e (responsável pela síndrome respiratória aguda grave) MERS-CoV (que causa a síndrome respiratória do Oriente Médio) (OPAS, 2024).

Embora declarado o fim do estado emergencial causado pela doença, é reconhecido historicamente que o fim de uma pandemia raramente é abrupto, havendo sempre o risco de ressurgimento ou aparecimento de novas variantes. Conforme destacado por Meumann e Robson (2023), desde dezembro de 2019 foram identificadas cinco variantes: Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2) e Omicron (B.1.1.529). Esta última, em particular, mostrou-se mais transmissível e evasiva ao sistema imunológico. Em 2022, múltiplas ondas foram associadas ao surgimento de sublinhagens da Omicron.

As principais manifestações clínicas da COVID-19 são associadas ao trato respiratório inferior, sendo os sintomas mais comuns a febre, tosse, fadiga, falta de ar, dor de cabeça, dor muscular, congestão nasal, síncope e conjuntivite. Em alguns casos, no entanto, a infecção pode ocorrer de forma assintomática (Ebrahim *et a.*, 2020). Além disso, certas condições podem ser consideradas como comorbidades para a COVID-19, como cardiopatias, tabagismo, doenças pulmonares crônicas, diabetes e hipertensão arterial (Feitosa *et al*., 2020). Tais condições, aliadas a fatores como o mecanismo e letalidade da infecção viral, culminaram em um alarmante índice de mortalidade, tornando essenciais a exploração e aprofundamento do conhecimento sobre os aspectos do vírus. Essa necessidade pode ser observada numericamente por estudos como o de Makoana *et al.* (2025), que evidencia que até maio de 2024, mais de setecentos milhões de pessoas em todo o mundo foram contaminadas, resultando em mais de 6,8 milhões de óbitos.

Neste cenário, muitas ações preventivas desempenharam um papel fundamental no enfrentamento da pandemia. Entre elas, períodos de *lockdown* e o isolamento social destacaram-se pela eficácia na contenção da disseminação viral, visto que um dos maiores meios de transmissão é o contato com indivíduos infectados pela inalação direta de gotículas expelidas por tosse, espirro e por meio de secreções nasais, orais, oculares e mucosas (Makoana *et al.*, 2025). No entanto, embora indispensáveis sob tais circunstâncias, essas medidas impactaram significativamente vários outros aspectos na esfera social, como a política, economia, educação e a relações sociais. Além das consequências diretas na saúde física, houve também um grande efeito psicológico na população, intensificando estresse, ansiedade, depressão e medo, sobretudo em grupos socialmente vulneráveis. Diante disso, mobilizaram-se esforços globais para desenvolvimento de tratamentos e vacinas, sendo o diagnóstico um ponto fundamental até os dias atuais.

1.2 *DIAGNÓSTICO*

O correto diagnóstico de qualquer enfermidade permite a elaboração de condutas adequadas frente ao quadro e, a partir daí, determinação de quais serão os próximos passos. No caso da COVID-19, dada a dimensão tomada, a adaptação de testes diagnósticos ocorreu em tempo recorde. De acordo com Souza *et al.*, 2021, países como a Coreia do Sul adotaram estratégias de testagem em massa, o que possibilitou maior controle da contaminação e

redução na taxa de letalidade. Em contrapartida, no Brasil, a testagem se restringiu apenas aos casos mais graves, o que pode, inclusive, ter causado uma subnotificação dos casos, já que muitos não foram testados e/ou apresentaram infecção assintomática.

Verotti *et al*. (2020) compilaram uma lista de testes diagnósticos registrados pela Agência Nacional de Vigilância Sanitária (ANVISA). Até o momento do estudo, havia um total de 217 testes de diferentes marcas, divididos principalmente em quatro tipos: 135 testes de imunocromatografia; 34 testes de Transcrição Reversa seguida de Reação em Cadeia da Polimerase (RT-PCR); 15 testes de anticorpos; e 11 imunoensaios fluorescentes. Além disso, Pereira *et al.* (2021) destacam que a depender da amostra biológica analisada, a sensibilidade do teste pode variar amplamente, indo desde 93% partindo do lavado broncoalveolar até 29% em fezes. O RT-PCR é reconhecido como o "padrão-ouro". O método de amostragem mais comum para sua realização é o *swab* nasofaríngeo, cujo objetivo é obter células superficiais do epitélio respiratório (Menezes; Lima; Martinelo, 2020; Morales-Angulo *et al*., 2020). Para isso, é feita a inserção do *swab* pela narina dos pacientes, até que este chegue à parte posterior da garganta, onde é feita a coleta de secreções.

Essa técnica identifica sequências específicas do genoma do vírus por meio da extração e conversão do RNA genômico do vírus em DNA complementar (cDNA) por meio da enzima transcriptase reversa. Com isso, a amplificação do DNA é detectada e quantificada em tempo real conforme a reação em cadeia de polimerase (PCR) progride. Em geral, é recomendado detectar, de início, uma região menos específica, como o gene do envelope ou o gene E, para triagem, seguida por uma região mais específica, como o gene da polimerase RdRp RNA, para confirmação. Dessa forma, é possível que cada laboratório utilize combinações de sequências variadas, resultando em diferentes taxas de sensibilidade e especificidade (Souza *et al.*, 2021). Baseado nessa extração de RNA viral do material coletado do paciente, a contabilização final de genes na amostra, que requer o tempo de aproximadamente 2 a 5 horas, representa um resultado classificado como positivo ou negativo (Dutta *et al*., 2022; Oliveira *et al*., 2022; Yuce; Filiztekin; Özkaya, 2021; Valdés, 2020).

Por sua vez, os testes de antígeno se destacam pela rapidez, fornecendo resultados de 15 a 30 minutos, além de apresentarem custo reduzido em relação ao RT-PCR. Sua realização ocorre por ensaios de fluxo lateral em que a amostra de material da nasofaringe reage com uma solução para que o vírus libere seus antígenos e, logo depois, é colocada sobre um dispositivo de papel no qual os antígenos migram ao longo da superfície e interagem com os anticorpos específicos do Sars-Cov-2, obtendo uma resposta qualitativa.

De modo geral, uma comparação entre os métodos diagnósticos como a discutida por Karki *et al.* (2024) revelam a necessidade de aprimoramento nas estratégias de testagem, já que mesmo o padrão-ouro pode apresentar falsos-negativos, especialmente em casos de coleta inadequada, falha na extração de RNA viral ou baixa quantidade de material celular na amostra (Oliveira; Matos; Morais, 2020).

Ressalta-se, portanto, a importância da investigação de métodos alternativos de testagem, sobretudo perante às limitações associadas às técnicas amplamente utilizadas. Diante destes obstáculos, a urina se apresenta como uma potencial amostra biológica alternativa, posto que a coleta não é invasiva, não representa um provável ponto de contágio, trata-se um material de fácil manuseio e armazenamento, além de ter composição bastante representativa no que se refere a saúde do sujeito. Como demonstrado por Tristán *et al.* (2024), poucos estudos publicados exploram essa matriz para investigação desta síndrome respiratória. Tal lacuna, aliada à relevância direta para saúde pública, evidencia a importância de explorar essa promissora abordagem.

1.3 *METABOLÔMICA*

As ciências ômicas integram um conjunto amplo de abordagens analíticas voltadas à caracterização, identificação e quantificação de moléculas em sistemas biológicos, para que, a partir delas, se compreenda processos que ocorrem nestes sistemas. Nesse contexto, destacam-se áreas como genômica, transcriptômica, proteômica, lipidômica e metabolômica, cada uma focada em diferentes níveis da organização biológica. De forma geral, essas abordagens compartilham um caráter interdisciplinar e integrativo, sendo algumas de suas aplicações mais frequentes a investigação de biomarcadores associados a diagnósticos e acompanhamento e progressão de doenças.

Nesse âmbito, a abordagem empregada neste estudo, a metabolômica, tem como foco investigar alteração em organismos por meio da análise de seus metabólitos, que são moléculas de baixa massa molar, como açúcares, aminoácidos, ácidos orgânicos, nucleotídeos, entre outros. Isso porque tais componentes são responsáveis pela homeostase orgânica e retratam o estado funcional das células, podendo sofrer modificações na presença de patologias (Padovani, 2025). Em particular, a metabolômica global adota uma análise que

compreende o maior número de metabólitos possível, incluindo diversas classes, que são medidas através de uma comparação entre os grupos de amostras para que se possa determinar uma diferenciação entre eles (Canuto *et al*., 2018).

Com avanço contínuo de técnicas analíticas, essa ciência emergente evolui cada vez mais. A literatura documenta várias aplicações neste campo, tanto no estudo de fluídos biológicos como plasma (Da Silva; Amaral; Martins, 2017) e urina (Araújo, 2020), quanto fungos (Souza, 2020) e plantas (Leite, 2024), entre outros. Essa versatilidade metodológica demonstra o vasto potencial de sua aplicabilidade.

1.4 *CROMATOGRAFIA LÍQUIDA ACOPLADA À ESPECTROMETRIA DE MASSAS APLICADA A METABOLÔMICA*

Diversos métodos de separação e detecção podem ser utilizados no estudo da metabolômica global. No entanto, a cromatografia líquida acoplada à espectrometria de massas (LC-MS) destaca-se como uma técnica amplamente empregada por permitir a medição simultânea de um grande número de metabólitos em uma única análise. A espectrometria de massas pode também ser aplicada de forma isolada por meio da análise direta, já que possibilita a identificação de moléculas com elevado grau de confiabilidade (Borges *et al.*, 2022). No entanto, a associação a um método de separação como a cromatografia, evita supressão de sinais e efeitos de matriz pois permite que a detecção leve em consideração também os tempos de retenção das moléculas de acordo com suas respectivas características físico-químicas.

A diferenciação de dois ou mais grupos em estudos metabolômicos é realizada a partir dos *features* obtidos durante a etapa experimental. Esses *features*, definidos como íons caracterizados por razão massa/carga (*m/z*) e tempo de retenção únicos, apresentam sinais associados a possíveis metabólitos detectados na amostra. Sendo assim, a capacidade de detectar simultaneamente um grande número desses sinais constitui uma vantagem importante desta técnica. No entanto, o tempo de retenção de um composto é suscetível a variação entre diferentes corridas analíticas devido a fatores experimentais como degradação da coluna, flutuações de temperatura, alterações no pH da fase móvel, entre outros.

Neste estudo, o fluxo de trabalho será executado a partir do processamento de dados, uma vez que as etapas experimentais foram previamente executadas por outros membros do Grupo de Química Analítica e Quimiometria (GQAQ/UFJF). Embora a aquisição de dados não integre o escopo deste trabalho, destaca-se a relevância da técnica empregada para este tipo de análise, cujo protagonismo se deve à capacidade de medida rápida e rotineira de grande número de metabólitos com excelente precisão em uma única análise. Contudo, as variações aleatórias citadas e o volume extenso de dados a serem analisados podem dificultar a interpretação dos dados, tornando imprescindível o aporte de ferramentas estatísticas para determinar as variáveis mais relevantes e estabelecer seus potenciais como biomarcadores diagnósticos de forma precisa e reprodutível (Tautenhahn *et al.*, 2012; Clish, 2015).

1.5 *TRATAMENTO DE DADOS*

O enorme volume de dados em estudos metabolômicos, caracterizado por informações tridimensionais de elevada complexidade, são compostos por sinais ao longo do tempo de retenção e valores de razão *m/z*, frequentemente acompanhados por alterações decorrentes de fatores experimentais descritos anteriormente. Tais variações podem ocasionar desvios e dificultar a comparação direta entre amostras (Pilon *et al.*, 2020).

Diante disso, a primeira medida a ser tomada após a aquisição dos dados é transformar os dados brutos para torná-los comparáveis, o que se dá por meio da detecção de picos, alinhamento de tempos de retenção e agrupamento de sinais, permitindo a identificação de *features* que são compilados em uma matriz estruturada adequadamente para análise estatística. Tal matriz reúne as informações que caracterizam os candidatos a metabólitos, contendo as intensidades relativas associadas a cada variável, bem como o tempo e massa que o caracterizam. Além disso, são também listadas a classificação de cada amostra em seu grupo amostral, permitindo a comparação entre condições distintas.

Mesmo após a organização dos dados em formato matricial, a interpretação manual ainda é inviável. Por essa razão, torna-se necessária a aplicação de ferramentas que possibilitem a análise e interpretação desses dados de forma sistemática. A quimiometria, entendida como uma área interdisciplinar que combina métodos estatísticos, matemáticos e químicos, destaca-se como uma abordagem fundamental nesse processo, sendo aplicada com o propósito de analisar e interpretar os dados. Por meio dela é possível explorar tanto análises multivariadas, que avaliam padrões globais e relações entre múltiplas variáveis

simultaneamente, quanto análises univariadas, voltadas à investigação de variações individuais em *features* específicas (Lovatti, 2019; Veras *et al*., 2022).

Adicionalmente, técnicas de aprendizado de máquina (*machine learning,* ML) podem ser empregadas como parte das abordagens quimiométricas, consistindo em métodos capazes de identificar padrões nos dados e construir modelos preditivos sem necessariamente serem programados para cada situação, utilizando conjuntos de dados e instruções iniciais para construir, progressivamente, seu próprio conhecimento com base nos resultados obtidos (Schiaffino, 2020; Alzubi; Nayyar; Kumar, 2018; Alba *et al*., 2022).

De acordo com Souza *et al*. (2021), o uso de técnicas de ML tem o potencial de reduzir substancialmente o esforço envolvido na programação direcionada à mineração de dados e à resolução de determinados problemas. Em termos práticos, simplifica a busca por um modelo simples a partir do tratamento de dados extensos, transformando um volume expressivo de informações em algo interpretável por meio da sua classificação, agrupamento e associação.

Uma das principais barreiras associadas ao tratamento de dados metabolômicos está na dificuldade das ferramentas disponíveis, que frequentemente exigem conhecimento técnico específico para sua instalação e operação. Essa demanda por expertise, embora compreensível diante da complexidade dos dados, acaba restringindo significativamente o uso dessas ferramentas. Como condição de contorno, surgem softwares alternativos como o XCMS Online (Tautenhahn *et al*., 2012), uma plataforma derivada do pacote XCMS do R (Smith, C. A. *et al*., 2015), operado por meio de uma interface intuitiva que dispensa conhecimentos aprofundados em programação, enquanto concentra as funcionalidades do XCMS original, um programa operado por meio de linhas de comando e scripts. Nesse sentido, o presente trabalho utilizou essa plataforma para o processamento dos dados combinada ao software MetaboAnalyst 6.0 (Pang *et al*., 2024), empregado por concentrar uma variedade de tratamentos estatísticos.

Diante deste contexto, é proposto o uso combinado de Quimiometria e da ciência de dados, em associação com o fluxograma clássico de processamento metabolômico para investigar as respostas metabólicas do corpo humano à infecção com o vírus SARS-CoV-2. Serão avaliados os resultados gerados a partir da análise de amostras de urina por HPLC-MS, que compõem um banco de 600 cromatogramas associados com espectros de massas de alta resolução nos modos de ionização positiva (ESI +) e negativa (ESI -).

2 OBJETIVOS

Nesta seção são apresentados os objetivos estabelecidos para o desenvolvimento deste trabalho.

2.1 *OBJETIVO GERAL*

Realizar o processamento de dados de amostras de urina analisadas por meio de HPLC-ESI-QTOF, aplicando a metabolômica global para a seleção de metabólitos diferenciais entre Grupo Teste (RT-PCR positivo) e Grupo Controle (RT-PCR negativo) da COVID-19.

2.2 *OBJETIVOS ESPECÍFICOS*

- Testar diferentes parâmetros de processamento para analisar os dados metabolômicos obtidos por HPLC-ESI-QTOF;
- Aplicar a abordagem metabolômica global para determinar metabólitos diferenciais presentes nas amostras de urina;
- Utilizar principais metabólitos diferenciadores para desenvolver um modelo preditivo que possa distinguir entre amostras de urina de indivíduos com COVID-19 (RT-PCR positivo) e sem (RT-PCR negativo).

3 PROCEDIMENTO EXPERIMENTAL

Nesta seção serão apresentados inicialmente os protocolos experimentais adotados para a obtenção dos dados brutos analisados neste projeto. Em seguida, serão descritas as etapas de tratamento de dados empregadas.

3.1 *CONJUNTO DE AMOSTRAS*

Os dados brutos utilizados como base para a busca por biomarcadores foram obtidos anteriormente ao início deste projeto, a partir de procedimentos experimentais realizados por outros membros do grupo de pesquisa (Moreira, 2024). A coleta amostral foi realizada após aprovação do Comitê de Ética em Pesquisa (CEP) do Hospital Universitário da Universidade Federal de Juiz de Fora (HU-UFJF), sob os protocolos 4.473.404; 4.566.092; 5.039.371, que permitiam a coleta de fluidos biológicos humanos voltados à pesquisa acadêmica. Participaram deste estudo voluntários maiores de 18 anos com suspeita de infecção por SARS-CoV-2, mediante assinatura do Termo de Consentimento Livre e Esclarecido (TCLE), preenchimento de ficha de anamnese e coleta de 50 mL de urina. O recrutamento e as coletas foram realizados em parceria com o Lemos Laboratórios de Análises Clínicas, em Juiz de Fora (MG).

A confirmação diagnóstica foi feita por RT-PCR utilizando o kit TaqPath™ COVID-19 CE-IVD (Thermo Fisher Scientific), com interpretação automatizada via software da própria fabricante. Após coleta, as amostras foram adicionadas de 10% (v/v) com solução de acetona/metanol (60:40, v/v) para inativação viral, congeladas e enviadas ao laboratório. Lá, cada amostra original foi subdividida em cinco alíquotas (“amostras-filhas”), congeladas a –50 °C para evitar ciclos de descongelamento e preservar a integridade da matriz biológica.

As amostras foram posteriormente descongeladas em temperatura ambiente, homogeneizadas e submetidas a precipitação proteica com 100 μL de acetonitrila gelada para cada mL de urina. O sobrenadante obtido após centrifugação foi utilizado para análise em espectrometria de massas. Foram preparadas três réplicas autênticas de cada amostra-filha (“R1”, “R2”, “R3”), submetidas independentemente ao processo completo do preparo.

O conjunto final é composto por 100 amostras de urina, sendo 38 positivas e 62 negativas para SARS-CoV-2. Todas as análises foram conduzidas em regime duplo-cego, ou seja, o resultado dos testes de RT-PCR e qualquer outra informação relacionada ao paciente só

foram acessados após conclusão da fase analítica. É importante destacar que o critério de diferenciação entre os grupos de estudo foi apenas o resultado de RT-PCR.

3.2 *ANÁLISE INSTRUMENTAL*

A otimização dos parâmetros instrumentais foi realizada com base em planejamento experimental (*Design of Experiments*, DOE) do tipo fatorial parcial 3^3 (Box-Behnken), totalizando 15 experimentos, incluindo triplicata no ponto central, conforme descrito por Moreira *et al.* (2023). A condição ótima foi definida com base na obtenção de perfil cromatográfico de íons totais (*Total Ion Chromatogram*, TIC) de maior qualidade e um maior número de *molecular features*. Entre os fatores avaliados (*nozzle voltage* – x_1, nebulizador – x_2 e fragmentação - x_3), a voltagem de fragmentação foi o fator que mais influenciou no resultado final (valor $p < 0,05$), seguido pelo primeiro fator, *nozzle voltage*. Os valores selecionados para os parâmetros avaliados foram $x_1 = 0$ V, $x_2 = 50$ psig e $x_3 = 175$ V.

As análises foram realizadas em HPLC modelo 1260 Infinity II equipado com uma bomba quaternária e acoplado ao espectrômetro de massas de alta resolução modelo 6530 Accurate-Mass QTOF, operando com fonte de ionização por Eletrospray Dual Jet Stream Technology (Dual AJS-ESI), todos os módulos da marca Agilent Technologies (Palo Alto, CA, EUA), e software de aquisição de dados MassHunter LC/MS B.08.00 para controle do sistema descrito.

Para a separação cromatográfica, empregou-se uma coluna de fase reversa C18 Infinity Lab Poroshell 120 EC-C18 (4,6 x 100 mm x 2,7 μm), mantida a 30 °C enquanto o amostrador permaneceu a 10 °C. O volume de injeção adotado foi de 5 μL. O gradiente de eluição iniciou com 0% de solvente B de 0 a 0,5 min, aumentando para 9% até 2 min, 20% em 5 min, 45% em 8 min e atingindo 100% em 9,5 min, onde foi mantida até 11 min e, após isso, a etapa de reequilíbrio se deu até os 20 minutos para retorno à condição inicial. Sendo o solvente A composto de água e ácido fórmico (99,9:0,1, v/v) e B de acetonitrila e ácido fórmico (99,9:0,1, v/v) a 0,5 mL min^{-1}.

A aquisição dos dados foi realizada nos modos de ionização positiva e negativa, sendo que em ambos a operação ocorreu em varredura com faixa de massa de 50 a 1500 a uma taxa de 1,02 espectros/s. O gás de secagem foi mantido a 325 °C com vazão de 11 L min^{-1} e o gás de revestimento a 350 °C com vazão de 11 L min^{-1}. A voltagem capilar foi ajustada a 4000 V. As massas de referência habilitadas foram 121,0509 (Purina, $C_5H_4N_4$) e 922,0098

(Hexakis(1H, 1H, 3Htetrafluoropropoxi)fosfazina, $C_{18}H_{18}O_6N_3P_3F_{24}$) para ESI (+). 119,0363 (Purina, $C_5H_3N_4$) e 301,9981 (Tris (2,4 ,6-trifluorometil)-1,3,5-triazina, $C_6N_3F_9OH$) considerando ESI (-).

Os experimentos foram conduzidos de forma aleatória, utilizando uma amostra controle de qualidade (Quality Control, QC) preparada a partir de um *pool* representativo das amostras de urina analisadas. Para sua obtenção, alíquotas de 20 µl de cada bloco de réplicas autênticas (R1, R2 e R3) foram combinadas, resultando em soluções estoque de QC capazes de refletir a variabilidade metabólica do conjunto amostral. O QC foi empregado durante os processos de otimização do método e inserido de forma intercalada ao longo das sequências analíticas. Paralelamente, amostras de branco analítico, constituídas por água ultrapura, foram organizadas de forma randômica ao longo das sequências de análise, intercaladas com as amostras biológicas e QCs, permitindo o monitoramento de possíveis contaminações, efeitos de *carryover* e interferências provenientes do sistema analítico.

Antes de cada experimento, após o reequilíbrio da coluna cromatográfica, o sistema foi submetido à injeção exclusiva de fase móvel com o objetivo de promover a limpeza e estabilização dos parâmetros variáveis. Adicionalmente, a parte externa da fonte de ionização foi limpa com isopropanol de grau HPLC-MS. O software MassHunter foi utilizado para inspeção visual dos dados.

3.3 *PRÉ-PROCESSAMENTO*

Neste trabalho, esta etapa compreende os procedimentos de organização, inspeção e refinamento inicial de dados brutos, o que inclui a exploração visual dos dados e elaboração e filtragem dos conjuntos finais a serem submetidos às etapas subsequentes. Já o processamento em si, corresponde à etapa em que os dados previamente extraídos são sistematicamente tratados e interpretados por meio da aplicação de métodos estatísticos.

3.3.1 SELEÇÃO DOS PARÂMETROS DE PROCESSAMENTO DOS DADOS

Para operação, o XCMS Online requer a definição de diversos parâmetros para realizar o processamento, que inclui a detecção de picos, alinhamento de tempo de retenção e o agrupamento de sinais. Embora sejam fornecidas configurações *default* para diferentes

equipamentos, elas podem ser customizadas e ajustadas com base nas características específicas de cada conjunto de dados brutos.

Uma das maneiras de realizar essa customização dos parâmetros de entrada é o pacote do R *Isotoprologue Parameter Optimization* (IPO). Ele automatiza a otimização aplicando uma abordagem sistemática de avaliação para encontrar os melhores parâmetros, baseando-se em análises sucessivas com diferentes configurações em uma lógica similar à de um planejamento experimental (Tautenhahn *et al.*, 2012).

Para tal, foram realizados dois testes. No primeiro, foram selecionados três QCs aleatórios, posicionados estrategicamente no início, meio e fim da sequência analítica. Já no segundo, como condição de contorno ao número de *features* muito baixos encontrados no primeiro caso, utilizou-se 10 QCs. O IPO foi executado no ambiente RStudio (versão 4.3.1), gerando superfícies de resposta dos parâmetros testados, bem como um arquivo de saída contendo os valores otimizados.

Apesar de demandar maior tempo de processamento computacional, já que a ferramenta realiza iterações até identificar os valores que resultam no melhor desempenho do processamento, o seu uso é uma alternativa à otimização manual, pois dessa maneira, confere maior robustez e reprodutibilidade ao tratamento dos dados, conforme destacado em Dos Santos e Canuto (2023).

3.3.2 XCMS Online

Como já citado, a primeira etapa realizada pelo XCMS Online é a detecção de *features*, nela, é feita a identificação de potenciais metabólitos ao longo da análise, processo que se dá a partir da busca de sinais consistentes ao longo das consecutivas varreduras. São definidos parâmetros que definem certo limites associados a essa detecção, como o desvio máximo tolerado em ppm entre uma varredura e outra para determinar se se trata de um mesmo *feature* (*ppm*). Essa avaliação evita a associação indevida de sinais que possuem diferenças significativas. Além disso, um limite também é estabelecido para que haja diferenciação entre sinais reais e ruídos (*s/n threshold*), abaixo do qual, um pico pode ser invalidado, contribuindo com a redução de falsos positivos decorrentes de flutuações aleatórias do sinal instrumental.

Ocorre também o estabelecimento de largura mínima e máxima para o pico cromatográfico (*minimum peak width* e *maximum peak width*), além da diferença em razão

m/z para que picos sejam considerados diferentes (*mzdiff*). De acordo com as particularidades de cada *dataset*, é possível a aceitação de picos sobrepostos por meio de valores negativos para o *mzdiff*. No geral, a etapa de *Peak Picking*, define sinais com determinado *m/z* e seu tempo de retenção, indicando a sua altura, que é a máxima intensidade dos pontos, e sua área, definida como a soma das intensidades dos pontos (Pilon *et al*., 2020) por meio dos parâmetros citados.

Em seguida, ocorre o Alinhamento de Tempo de Retenção (*Retention Time Correction*) por meio de critérios que controlam a formação de grupos de correção, como o número (*minsamp*) ou fração mínima (*minfrac*) de amostras necessário para que seja feito um agrupamento. Gerada essa junção, ocorre a minimização da variação dos tempos de retenção dos mesmos *features* identificados em diferentes amostras, que pode ser realizada de maneira linear ou não linear.

É viável a realização de testes estatísticos e etapas posteriores no XCMS Online, entretanto, não foi aplicado. Quando finalizados os processamentos, as tabelas resultantes foram exportadas sem filtragem ou tratamento adicional, apenas as médias dos *features* de cada triplicata foram calculadas manualmente no Microsoft Excel e os valores faltantes foram substituídos por células vazias. O formato ajustado para o presente tipo de estudo (Figura 1) deve conter o identificador das amostras (*Sample*) e o grupo experimental a qual pertencem (*Label*) a fim de garantir a compatibilidade com o carregamento no MetaboAnalyst. Na primeira coluna é possível observar os *features* determinados pelo XCMS, que segue um padrão MXTY, onde X indica a massa daquele íon e Y o tempo em que ela foi detectada (em minutos). No restante da tabela, são apresentadas as intensidades de picos dos *molecular features* em cada uma das amostras.

Figura 1 - Exemplo da organização da estrutura da tabela necessária para o processamento de dados subsequente.

Sample	R10	R11	R12	R16	R18	R1	R21	R22
Label	CONTROLE	CONTROLE	CONTROLE	CONTROLE	CONTROLE	CONTROLE	CONTROLE	CONTROLE
M395T11	28472,796	25239,958	67359,080	124567,111	14642,510	121791,553	13902,203	26069,820
M369T10_3	56292,842	83639,669	103601,561	179004,063	18588,795	409686,544	20883,175	51999,200
M370T10	18349,203	20090,686	31418,587	40409,694	10785,135	89646,933	10381,517	14321,085
M320T10	32314,798	49372,209	125927,048	24560,267	50669,137	38097,703	52060,121	18825,713
M429T11	149464,318	62700,906	433541,935	144962,565	90441,010	131463,394	96514,849	22304,000
M417T2_2	5236,737	5754,377	7696,254	2425,118	2034,601	7601,998	3192,816	4160,646
M396T11	17794,537	19810,882	21307,567	33581,236	9728,245	40906,509	9847,070	22298,465
M514T11	29971,272	21647,040	14086,657	3137,205	5615,668	18451,564	10395,954	6058,879
M134T9	55729,927	146316,351	180897,738	101471,538	37438,533	237331,686	52799,419	93573,716
M513T11	107310,070	80137,006	46896,391	7204,724	8465,606	71201,716	16485,033	12996,529
M276T10	36023,355	35089,015	18726,011	27024,824	18043,801	64816,867	21435,512	50501,373
M291T11	102968,328	248034,094	332458,260	103403,533	129445,997	342124,255	80864,402	112963,413
M395T2	10306,810	8044,978	8622,131	4040,535	2244,292	11889,801	4018,184	4518,811
M379T2	8811,234	17190,829	22621,720	18045,648	6197,903	22017,721	13638,551	11220,379
M534T10_2	181127,810	13660,895	308902,312	119046,511	10776,658	105967,874	75240,466	56002,134
M477T9	159981,794	8364,909	138841,954	52791,097	7742,488	22151,740	96410,325	23366,774

Fonte: Elaborado pela autora (2025).

Ao final do processo de otimização, os parâmetros otimizados partindo de 3 QCs no modo de ionização negativo foram: ppm = 39; *minimum peak width* = 11,2; *maximum peak width* = 69,5; *mzdiff* = -0,0026; s/n = 10; *prefilter peaks* = 3; *prefilter intensity* = 500; *noise filter* = 0; *bw* = 1; *minfrac* = 1; *mzwid* = 0,022; *minsamp* = 1; e *max* = 50. Já quando se partiu de 10 QCs no modo de ionização negativo, os parâmetros obtidos foram: *ppm* = 37; *minimum peak width* = 11,2; *maximum peak width* = 72,5; *mzdiff* = -0,0032 ; *s/n* = 10; *prefilter peaks* = 3; *prefilter intensity* = 100; *noise filter* = 0; *bw* = 1; *minfrac* = 1; *mzwid* = 0,023; *minsamp* = 1; e *max* = 50.

Para o modo de ionização positivo, os parâmetros otimizados com 3 QCs foram: ppm = 36; *minimum peak width* = 11,2; *maximum peak width* = 56; *mzdiff* = -0,0032; s/n = 10; *prefilter peaks* = 3; *prefilter intensity* = 500; *noise filter* = 0; *bw* = 1; *minfrac* = 1; *mzwid* = 0,021; *minsamp* = 1; e *max* = 50. Para a otimização com 10 QCs, foram obtidos: ppm = 37; *minimum peak width* = 12; *maximum peak width* = 47; *mzdiff* = -0,0026; s/n = 10; *prefilter peaks* = 3; *prefilter intensity* = 100; *noise filter* = 0; *bw* = 1; *minfrac* = 1; *mzwid* = 0,032; *minsamp* = 1; e *max* = 50.

Os valores *default* do software, também foram utilizados em ambos os modos, são eles: *ppm* = 30; *minimum peak width* = 10; *maximum peak width* = 60; *mzdiff* = 0,01; *s/n* = 6; *prefilter peaks* = 3; *prefilter intensity* = 500; *noise filter* = 0; *bw* = 5; *minfrac* = 0,5; *mzwid* = 0,025; *minsamp* = 1; e *max* = 100.

4 RESULTADOS E DISCUSSÕES

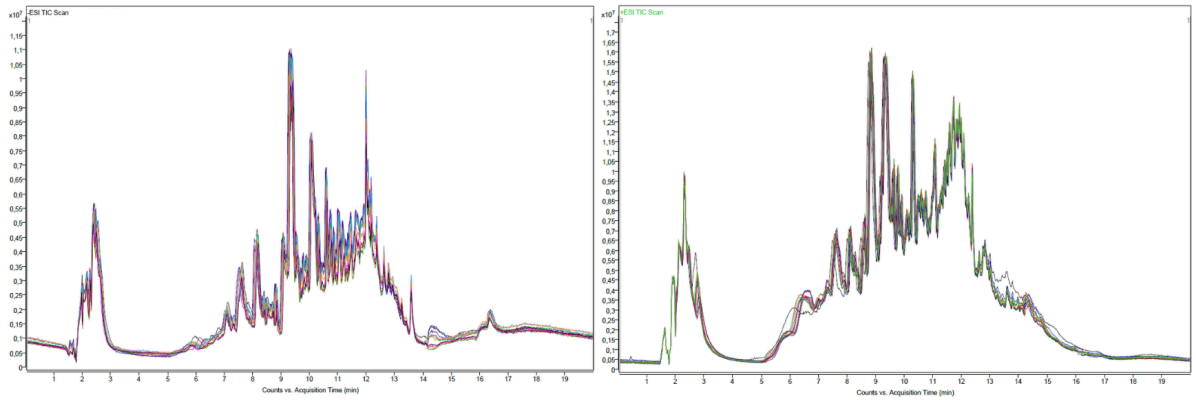
Nesta seção serão apresentados os resultados obtidos e as discussões sobre o trabalho desenvolvido.

4.1 *ANÁLISE VISUAL*

É possível realizar a exploração inicial dos dados por meio da avaliação visual dos cromatogramas, com o intuito de verificar anomalias ou tendências sistemáticas que podem comprometer os resultados, como, por exemplo, um aumento ou queda progressiva da intensidade do sinal ao longo do tempo ou alguma amostra pontual que teve resultados discrepantes das demais e que poderia causar desvios indevidos durante o processamento dos dados.

Utilizou-se o software MassHunter (Agilent) com o intuito de sobrepor os TICs das amostras de controle de qualidade (Figura 2), o que evidenciaram boa consistência entre os cromatogramas, com variabilidade visual pouco expressivas entre os QCs, indicando de forma geral, boa estabilidade no sistema ao longo das análises.

Figura 2 - Sobreposição dos cromatogramas de íons totais das amostras de controle de qualidade nos modos de ionização negativo e positivo, respectivamente.



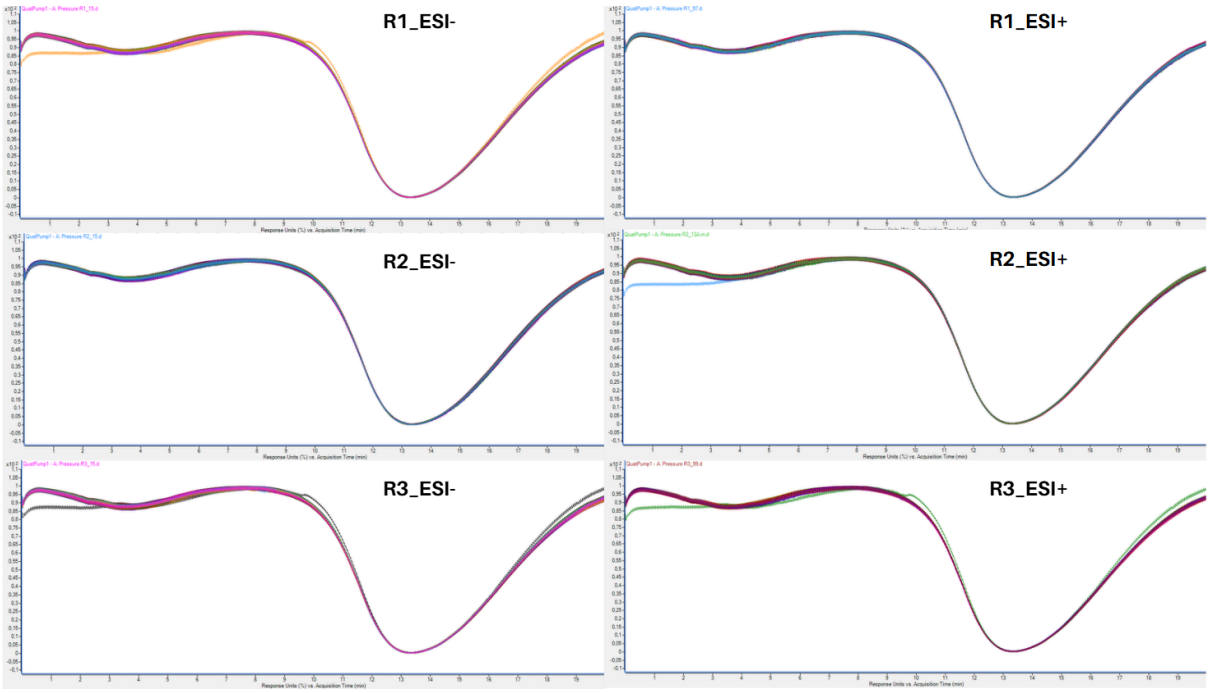
Fonte: MassHunter (2024), dados processados pela autora.

Optou-se por utilizar os QCs nessa análise, pois é esperado que as amostras apresentem variações nos TICs, uma vez que o objetivo do estudo é justamente identificar

diferenças entre grupos experimentais. Enquanto os QCs, além de terem sido submetidos ao mesmo protocolo analítico, como um *pool* de todas as amostras, possuem uma composição idêntica, fazendo com que não sejam observadas as mesmas alteridades.

Adicionalmente, foi avaliada a pressão do sistema cromatográfico (*backpressure*) das corridas (Figura 3). Variações na resistência encontrada pela fase móvel à medida que esta flui pelo sistema são comuns, já que devido ao gradiente e a composição, a viscosidade dessa fase é consequentemente alterada, afetando diretamente a pressão exercida sobre a coluna cromatográfica. Contudo, é esperado que isso ocorra de forma semelhante em todas as corridas, visto que as condições são as mesmas. Considerando a similaridade nos perfis obtidos, foi constatada a estabilidade do sistema ao longo de todas as corridas, não indicando nenhuma anormalidade.

Figura 3 - Pressão do sistema cromatográfico em cada uma das corridas no modo de ionização negativo e positivo, respectivamente.



Fonte: MassHunter (2024), dados processados pela autora.

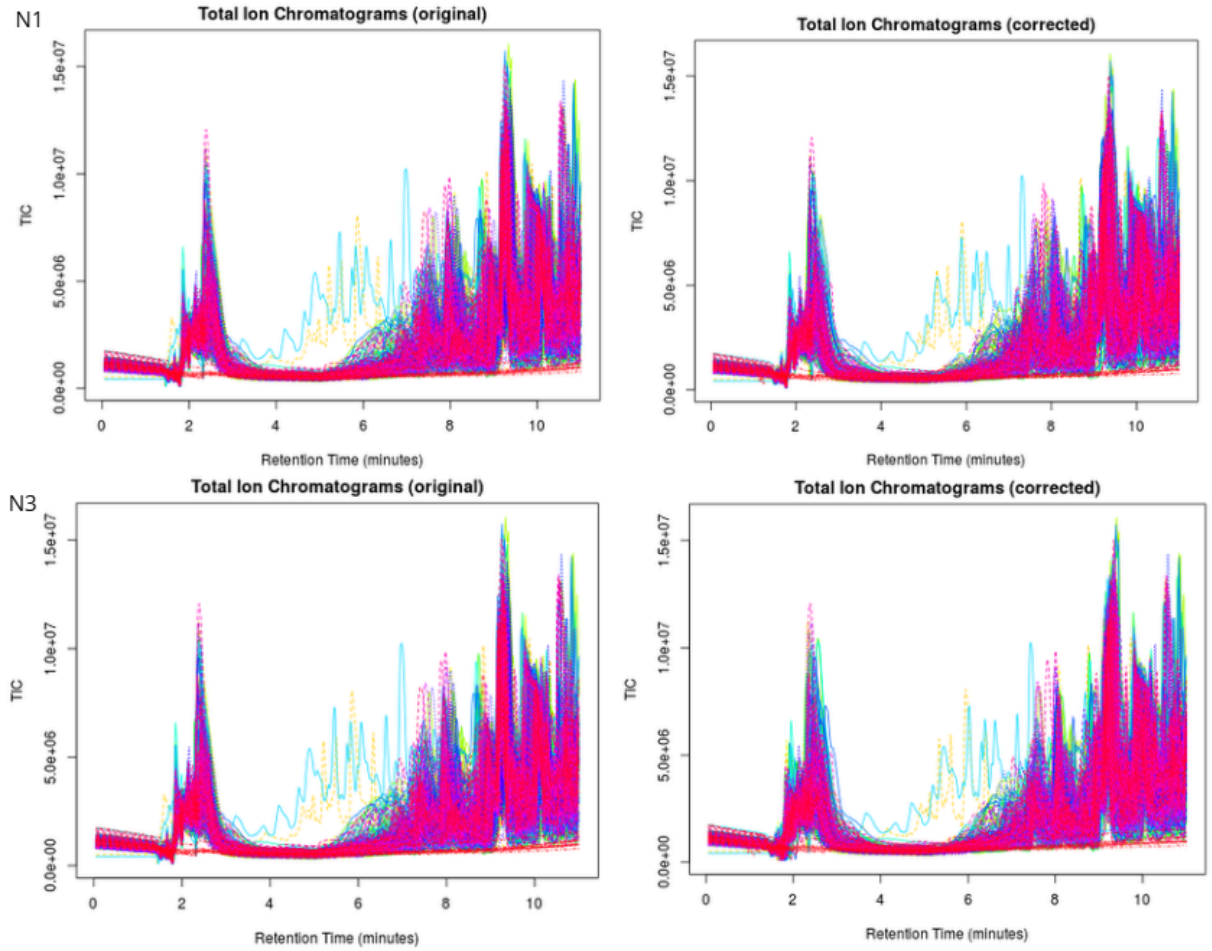
Os dados brutos, a princípio, gerados em formato, "*.d*" – próprio da fabricante Agilent, que nem sempre é compatível com os softwares de processamento, foram convertidos para o formato "*.mzML*" utilizando o software MSConvert (*ProteoWizard*, versão 3) (CHAMBERS et al., 2012). Esse formato, além de ser amplamente aceito por diferentes plataformas, é mais leve e mantém todas as informações necessárias.

Durante o processo de conversão, também foi realizado o corte dos arquivos em 11 minutos. Este intervalo é determinado pelo gradiente cromatográfico estabelecido durante os experimentos para aquisição dos dados, uma vez que, a partir deste ponto, ocorre apenas a limpeza e reequilíbrio da coluna.

4.2 *DETECÇÃO E ALINHAMENTO DE FEATURES*

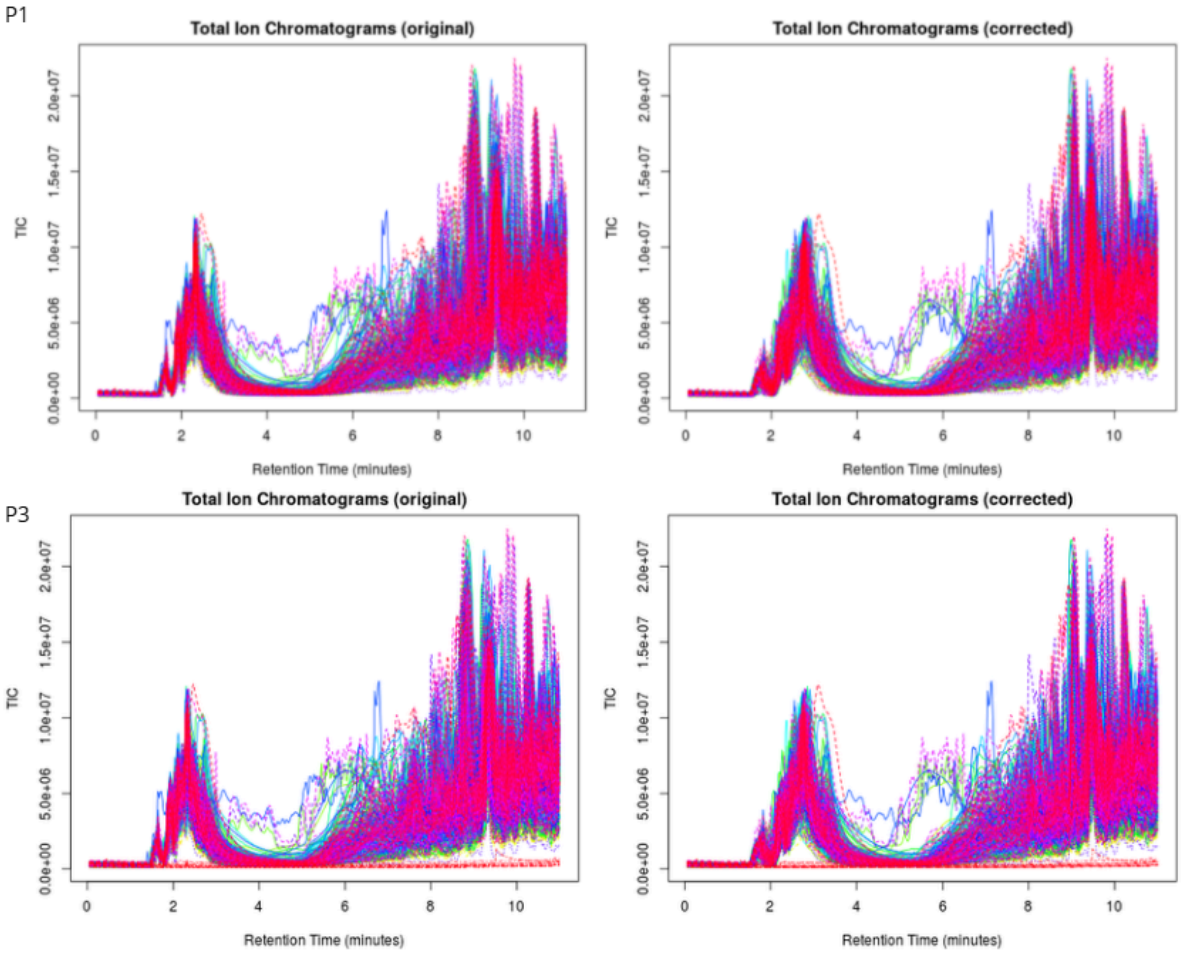
Ao observar os Cromatogramas de Íons Totais sobreposto antes e depois da correção (Figura 4 e 5), é possível perceber que a correção do tempo de retenção realizada pela plataforma, utilizando o algoritmo *obiwarp*, não impactou de maneira radical em nenhum conjunto de dados. Apesar das diferenças na definição de parâmetros em cada caso, é possível observar inegável semelhança entre os cromatogramas para o mesmo modo de ionização depois da correção realizada pelo software.

Figura 4 – Sobreposição de Cromatogramas de Íons Totais dos conjuntos N1 e N3, antes e depois alinhamento de tempos de retenção gerado pelo XCMS Online.



Fonte: XCMS Online (2025), dados processados pela autora.

Figura 5 – Sobreposição de Cromatogramas de Íons Totais dos conjuntos P1 e P3, antes e depois alinhamento de tempos de retenção gerado pelo XCMS Online.



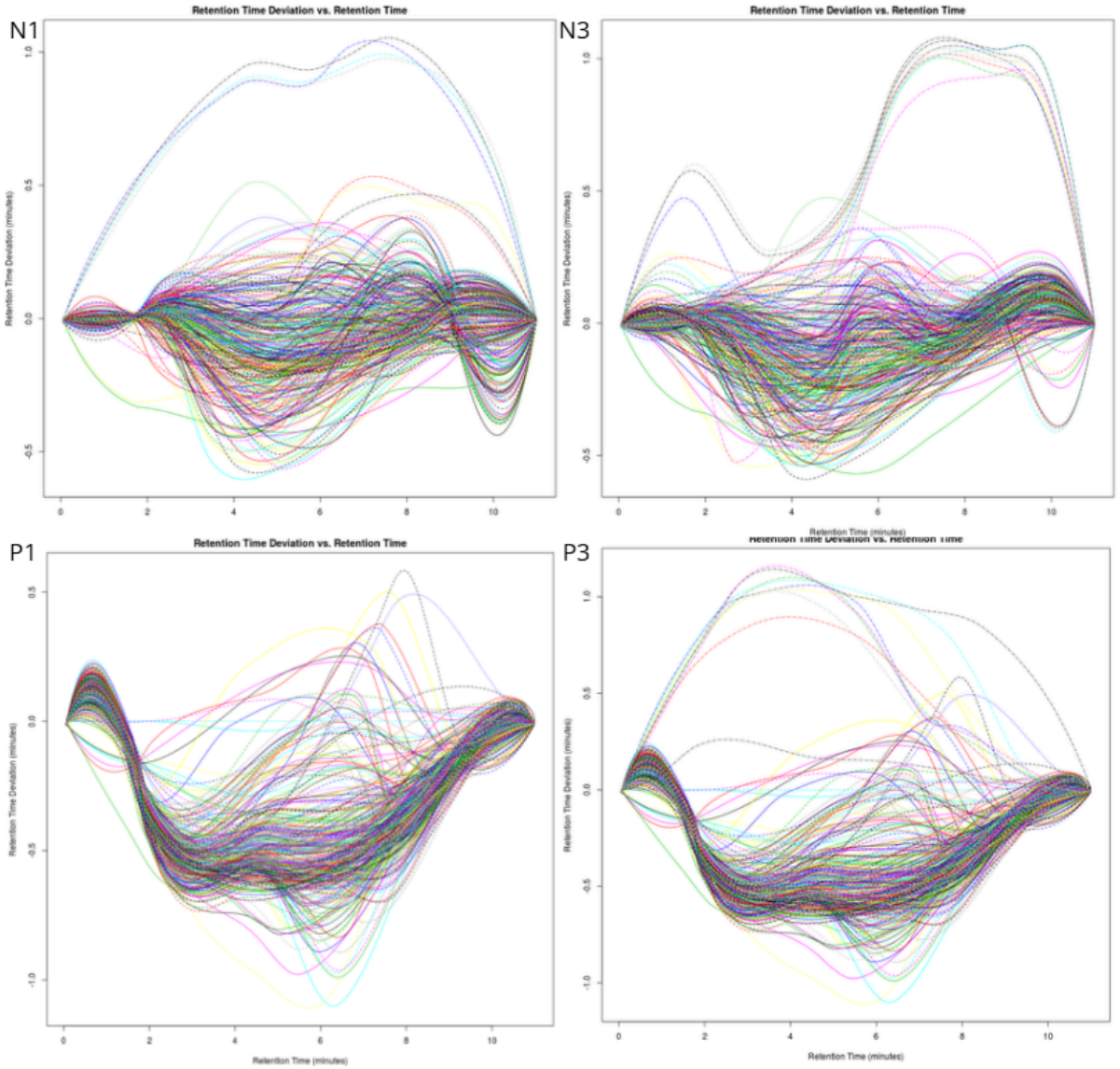
Fonte: XCMS Online (2025), dados processados pela autora.

Foram observadas variações no perfil das amostras no intervalo aproximado de 3 e 7 minutos, entretanto, é possível observar que, conforme o gráfico ilustrado na Figura 6, mesmo este sendo o intervalo em que o algoritmo de correção realizou mais ajustes, durante todo o tempo houveram modificações, algumas mais significativas, no sentido de alterar mais o tempo de retenção de uma corrida, outras menos. A dificuldade de observar essas mudanças em imagens como a figura 4 e 5, se dá por conta da sobreposição de um volume muito grande de informações.

De maneira geral, a baixa amplitude da oscilação indica pouca interferência, com variação de cerca de ±0,5 minutos e algumas poucas amostras com variações mais amplas, chegando ±1,0 minutos. No entanto, para os conjuntos N1 e N3, nota-se que determinadas amostras exibem desvios mais pronunciados em relação aos demais. As representações gráficas para N2 e P2 não foram apresentadas, pois em ambos os casos o XCMS falhou em realizar a correção do tempo de retenção. Como a correção aplicada para os demais casos não

foi severa, os conjuntos foram mantidos e seguiram para as etapas seguintes mesmo sem que este procedimento fosse realizado.

Figura 6 – Alinhamento de tempos de retenção gerado pelo XCMS Online para os conjuntos N1, N3, P1 e P3. Os pontos representam a diferença entre os tempos de retenção antes e após a correção.



Fonte: XCMS Online (2025), dados processados pela autora.

Ao finalizar o processamento no XCMS Online, com os parâmetros otimizados utilizando 3 QCs listados anteriormente, foi observada uma quantidade de *features* muito reduzida em relação ao que se observa em trabalhos similares, como demonstrado na Tabela 1. Já na otimização com 10 QCs, o número de *features* observado foi superior ao primeiro

processamento, mas ainda muito distante do encontrado com valores *default*. Para fins de padronização e clareza, a partir deste ponto, os conjuntos de dados provenientes de cada uma das otimizações aqui citadas, que serviram como dados de entrada e resultaram em seis tabelas distintas, passarão a ser identificados por códigos, também demonstrados na tabela.

Tabela 1 – Número de *features* detectadas nos modos de ionização positivo e negativo utilizando parâmetros otimizados e os parâmetros padrão do XCMS Online.

	ESI-			ESI+		
	Otimizados (3 QCs)	Otimizados (10 QCs)	*Default*	Otimizados (3 QCs)	Otimizados (10 QCs)	*Default*
n° de *features*	190	271	4512	752	824	6134
Código	N1	N2	N3	P1	P2	P3

Fonte: Elaborado pela autora.

A diferença observada na comparação entre o número de *features* obtidos para os diferentes conjuntos de parâmetros evidencia uma relação direta com as configurações adotadas no XCMS e pode ser explicada, principalmente, pelos parâmetros relacionados ao agrupamento e a filtragem dos sinais. Nos valores *default*, o *bw* foi definido como 5 e o *minfrac* como 0,5, permitindo maior tolerância na variação dos tempos de retenção e exigindo que uma *feature* estivesse presente em apenas 50% das amostras de um grupo para ser considerada válida. Em contraste, nas condições otimizadas, *bw* foi reduzido para 1 e *minfrac* elevado para 1, tornando os critérios mais restritivos, uma vez que apenas *features* presentes em todas as amostras passaram a ser consideradas. Essa mudança impacta diretamente na redução do número de *features*, favorecendo a seleção de sinais mais consistentes, porém potencialmente excluindo variáveis relevantes presentes de forma mais esporádica.

Além disso, o aumento da razão sinal/ruído (de 6 nos valores *default* para 10 nas condições otimizadas) contribui para uma filtragem mais rigorosa dos sinais, reduzindo a inclusão de ruídos, mas também podendo eliminar picos de baixa intensidade. De forma semelhante, a variação no parâmetro *prefilter intensity*, especialmente nos casos em que foi reduzido de 500 para 100, altera o limiar mínimo para detecção de picos, influenciando a sensibilidade de detecção.

No que se refere aos parâmetros de detecção de picos, as variações em ppm e *mzdiff* indicam ajustes mais refinados na tolerância de massa, enquanto as mudanças na largura máxima de pico afetam diretamente a definição dos limites cromatográficos dos sinais.

Embora essas variações sejam menos expressivas do que aquelas observadas nos parâmetros de agrupamento, elas também contribuem para as diferenças no número final de *features*.

De maneira geral, observa-se que as condições *default* favorecem a detecção de um maior número de *features*, pois possivelmente são incluídos sinais redundantes ou ruídos, enquanto os parâmetros otimizados resultam em um conjunto mais restrito e conservador de variáveis, priorizando reprodutibilidade e consistência entre as amostras. Esse comportamento evidencia a sensibilidade do processamento dos dados à escolha dos parâmetros e destaca a importância de um balanço entre abrangência e confiabilidade na definição das *features* utilizadas nas análises subsequentes.

4.3 *FILTRAGEM*

O primeiro critério de exclusão de *features* se baseou na remoção de sinais detectados em amostras de branco. Os brancos são processados com o intuito de representar as contribuições de reagentes, solventes e até mesmo do sistema no qual foi analisado, e portanto, a presença de *features* com intensidades pronunciadas nessas amostras, podem indicar que este não é proveniente de alteração fisiológica, e sim do protocolo ao qual foi submetido.

Para tal filtragem, foi realizado o cálculo da média das intensidades do sinal de cada *features* nas amostras de branco, bem como nas amostras reais. As intensidades foram comparadas e quando os valores das médias nas amostras de urina foram inferiores a 3 vezes a média dos brancos, o *feature* passou a ser desconsiderado para as etapas posteriores.

O segundo critério se baseou nas amostras de controle de qualidade, onde foi calculado o desvio padrão relativo (RSD) para cada *feature* nessas amostras, uma medida de dispersão para compreender o grau de variação dos dados, para o qual o limite de aceitação estabelecido foi até 30%. Valores acima deste indicam variabilidade elevada durante as análises, o que pode comprometer a modelagem ou gerar resultados não confiáveis caso mantidos. Feitas estas exclusões, os *features* restantes são apenas os que apresentam sinais reprodutíveis e consistentes, a Tabela 2 representa o número de *features* e valores faltantes antes e depois da filtragem. Após aplicados os filtros, os QCs e brancos foram excluídos manualmente e os dados carregados novamente no MetaboAnalyst.

Tabela 2 - Impacto da filtragem no número de *features* e valores faltantes.

	Pré filtragem		Pós filtragem	
	n° de *features*	Valores faltantes (%)	n° de *features*	Valores faltantes (%)
N1	190	659 (2.9)	189	7 (0)
N2	271	919 (2.8)	267	5 (0)
N3	4512	25196 (4.6)	3899	1123 (0.3)
P1	752	4016 (4.5)	735	358 (0.5)
P2	824	4476 (4.5)	803	376 (0.5)
P3	6134	35876 (4.9)	5572	2384 (0.4)

Fonte: Elaborada pela autora (2025).

Foi observada uma redução bastante pronunciada no número de valores faltantes, indicando que a maioria deles estava associada a *features* que não passaram pela triagem inicial ou a amostras de branco. Todavia, mesmo que a porcentagem de valores faltantes em relação ao número total de *features* tenha sido significativamente menor após os procedimentos citados, para evitar impactos indesejados nas análises estatísticas ainda se faz necessário lidar com os *missing values*.

Cabe destacar que o número total de valores faltantes é superior ao número de *features*, uma vez que a matriz de dados é composta por todas as combinações de entre *features* e amostras, ou seja, o número total de valores corresponde ao produto entre o número de *features* e o número de amostras. A título de exemplo, em um conjunto de dados com 752 *features* e 100 amostras, a matriz conterá 75.200 valores. Assim, mesmo que o número absoluto de *missing values* seja elevado, sua proporção em relação ao total de dados pode não ser tão expressiva.

Enquanto a exclusão de valores foi realizada no Excel, a imputação dos valores faltantes foi realizada diretamente no MetaboAnalyst. O software oferece 11 diferentes formas para isso, e, no caso deste estudo, foi selecionada a substituição das células vazias por ⅕ do menor valor de intensidade registrada para o íon em questão, o que faz com que o valor seja estimado como equivalente a um ruído, assumindo que se encontra abaixo do limite de detecção, mas não seja zerado.

4.4 *NORMALIZAÇÃO*

A urina representa uma das primeiras matrizes biológicas utilizadas para fins de diagnóstico. Sua aplicação em análises laboratoriais tem acompanhado o desenvolvimento de novas metodologias e inovações tecnológicas com o passar do tempo, tornando os procedimentos para sua avaliação mais sofisticados e robustos (Ataíde *et al.*, 2024). No entanto, é substancial a aplicação de condições de contorno para as variações significativas relacionadas a efeitos de diluição deste fluido biológico. Com essa finalidade, foi levado em consideração o valor de creatinina de cada amostra para realizar a correção das intensidades dos *features* por meio da divisão pelo respectivo valor de creatinina associado. Além disso, foi averiguado diferentes estratégias de normalização, transformação e escalonamento dos dados, como descrito no Quadro 1. O propósito da investigação foi verificar o impacto que essas operações teriam na distribuição dos dados e na quantidade de *outliers.*

Quadro 1 - Estratégias de normalização aplicadas aos dados.

	Normalização	Transformação	Escalonamento
Processamento 1	Mediana	Log_{10}	Centralização na média
Processamento 2	Mediana	Log_{10}	Autoescalonamento
Processamento 3	Mediana	Log_{10}	Pareto
Processamento 4	Mediana	Log_{10}	Por intervalo
Processamento 5	Mediana	Log_2	Centralização na média
Processamento 6	Mediana	Log_2	Autoescalonamento
Processamento 7	Mediana	Log_2	Pareto
Processamento 8	Mediana	Log_2	Por intervalo

Processamento 9	Mediana	Raiz quadrada	Centralização na média
Processamento 10	Mediana	Raiz quadrada	Autoescalonamento
Processamento 11	Mediana	Raiz quadrada	Pareto
Processamento 12	Mediana	Raiz quadrada	Por intervalo
Processamento 13	Mediana	Raiz cúbica	Centralização na média
Processamento 14	Mediana	Raiz cúbica	Autoescalonamento
Processamento 15	Mediana	Raiz cúbica	Pareto
Processamento 16	Mediana	Raiz cúbica	Por intervalo
Processamento 17	Mediana	Normalização estabilizadora de variância	Centralização na média
Processamento 18	Mediana	Normalização estabilizadora de variância	Autoescalonamento
Processamento 19	Mediana	Normalização estabilizadora de variância	Pareto
Processamento 20	Mediana	Normalização estabilizadora de variância	Por intervalo

Fonte: Elaborado pela autora (2026).

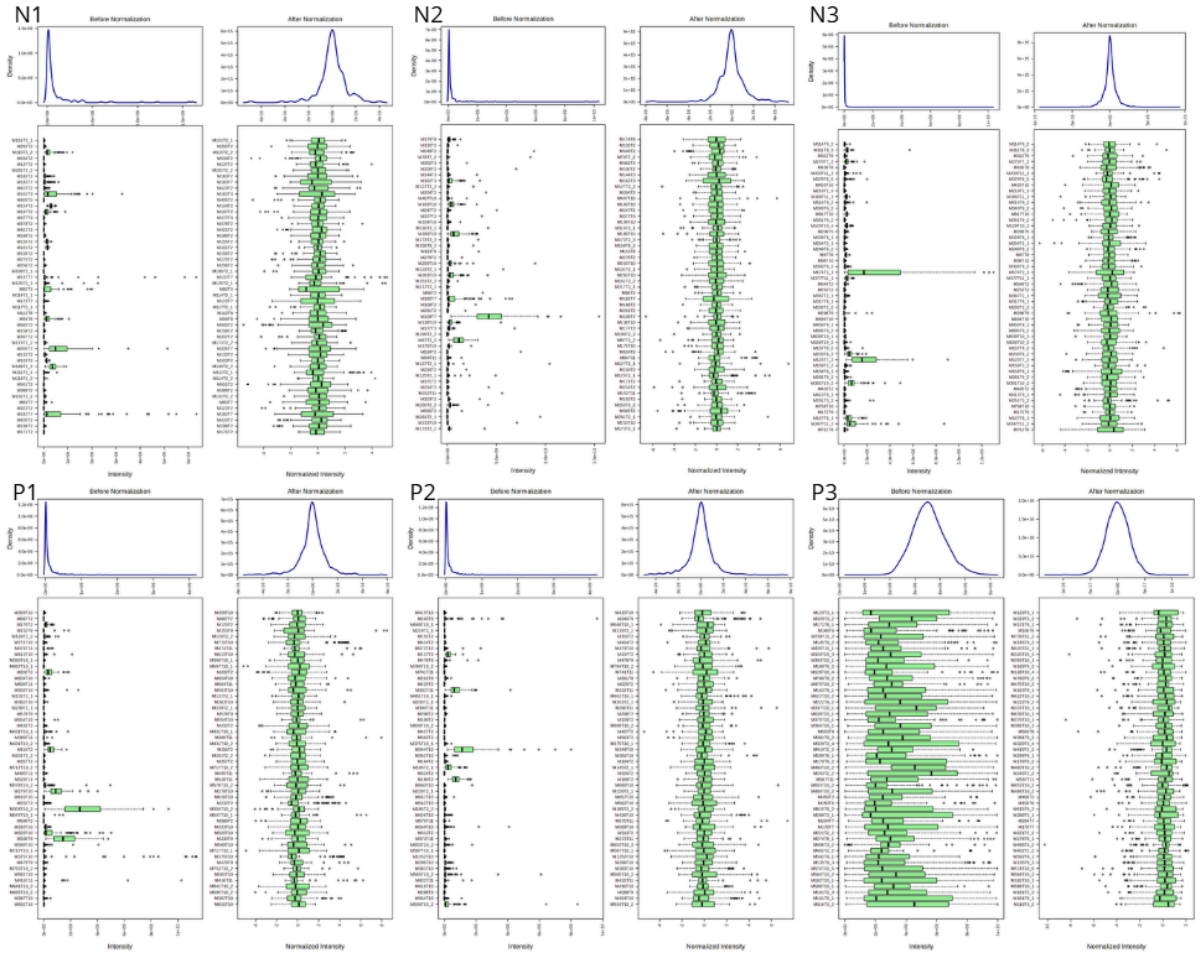
As representações gráficas referentes à distribuição dos dados e boxplots dos *features* após cada um dos processamentos encontram-se no Apêndice A. A seleção da condição mais adequada foi realizada a partir da comparação visual desses gráficos, onde buscou-se averiguar a proximidade das médias, que indicam maior uniformidade na distribuição dos dados, além do número de *outliers* em relação aos dados não tratados. Observados estes

pontos, foi selecionada a normalização pela mediana, transformação logarítmica na base 2 e o escalonamento de Pareto.

A transformação dos dados auxilia em situações onde a distribuição dos dados está inclinada à esquerda ou direita (Miot, 2017), como no caso dos conjuntos sob análise. Já após a transformação, a distribuição dos dados tornou-se mais semelhante à uma curva normal e, além disso, a aplicação de escala logarítmica auxilia bastante em questões de heterocedasticidade, condição na qual a variância não se mantém constante, mas aumenta conforme os valores médios das intensidades dos sinais. Dessa forma, esse tipo de transformação contribui para reduzir a influência desproporcional de metabólitos de alta intensidade (Pino, 2014).

Para o escalonamento de Pareto, cada variável tem sua média calculada e então, tem seu valor subtraído dessa média, o que faz com que a nova média torne-se 0 e, após essa centralização, cada variável é dividida pela raíz quadrada do seu desvio padrão. A figura 7 ilustra como cada um dos conjuntos de dados se comportou quando submetidos a esses procedimentos.

Figura 7 - Distribuição de dados antes e após normalização, transformação e escalonamento.



Fonte: MetaboAnalyst (2026), dados processados pela autora.

4.5 *ESTATÍSTICA*

Quando partimos de um conjunto de dados muito extenso, torna-se essencial a extração de informações relevantes. Entretanto, a análise isolada desses dados geralmente não é suficiente para obter conclusões significativas, surgindo a necessidade da aplicação de ferramentas que permitam um exame conjunto e sistemático das informações disponíveis.

O aprendizado de máquina pode ser distinguido em duas categorias principais: supervisionada e não supervisionada. No aprendizado não supervisionado, não há rótulos ou categorias previamente conhecidas. Apenas os dados de entrada são utilizados, e o objetivo é descobrir padrões, estruturas ou regularidades dentro do conjunto. Nesse tipo de abordagem, formam-se agrupamentos naturais de dados, com base nas similaridades observadas entre os exemplos. A análise de agrupamento, ou *clusterização*, é a principal tarefa nesse contexto.

Por outro lado, no aprendizado supervisionado trabalha-se com problemas de classificação e regressão, nos quais os dados de entrada são acompanhados de suas

respectivas saídas. O processo é composto por uma fase inicial de treinamento, onde um conjunto de dados rotulados é fornecido ao sistema e o algoritmo busca estabelecer uma função que relacione entradas e saídas. Com essa generalização da relação entre dado de entrada e a respectiva resposta, torna-se possível a previsão do resultado e ajuste de parâmetros para outros conjuntos além do utilizado no treinamento (Schiaffino, 2020). Em essência, esse tipo de aprendizado aprende o mapeamento entre entrada e saída com base em exemplos colocados a seu dispor.

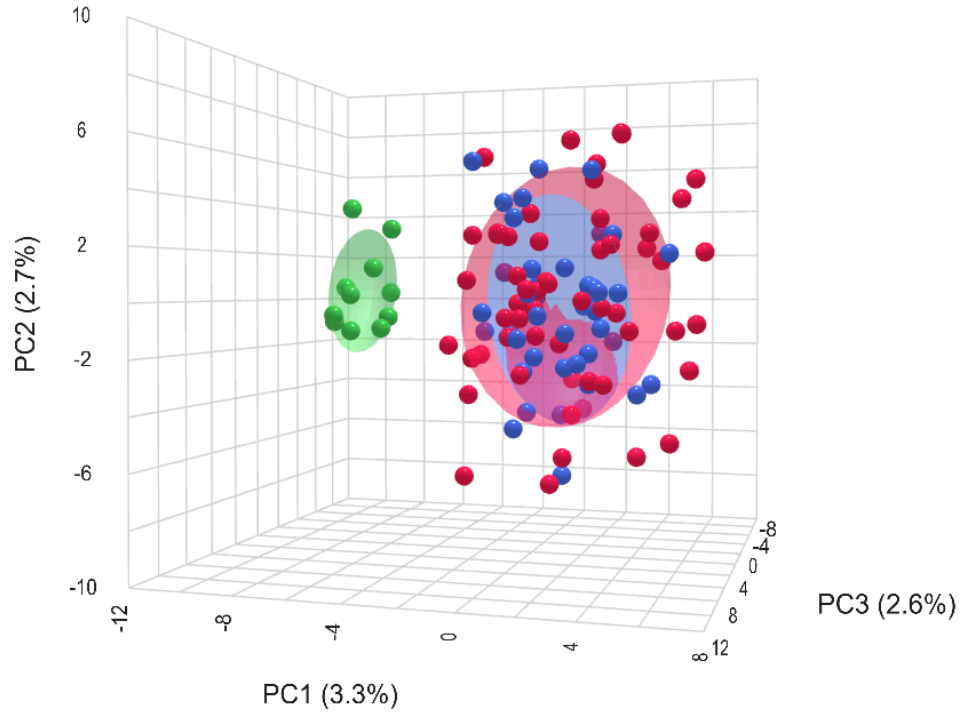
4.5.1 ANÁLISE DE COMPONENTES PRINCIPAIS

A Análise de Componentes Principais (PCA) é um método exploratório clássico que se baseia no conceito de redimensionamento de dados, parte da projeção de um conjunto de dados multivariados em um subespaço de menor dimensão, mantendo intactas as relações entre as amostras (Wold *et al.*, 1987). Em outras palavras, uma grande matriz de dados é transformada em um número reduzido de variáveis, chamadas de componentes principais (PCs), que são novos eixos que buscam reter a maior parte da informação original (Souza; Poppi, 2012).

Essa transformação permite uma visualização da estrutura de dados de maneira que facilita a identificação de padrões e tendências por similaridade entre as amostras, que passam a ser pontos localizados, com base na sua proximidade no espaço das componentes principais. Além disso, identifica a dimensionalidade intrínseca do conjunto de dados e a representa de forma simplificada.

Alternativamente, essa abordagem pode ser utilizada para avaliar a qualidade do tratamento de dados, e não necessariamente na busca pela separação dos grupos experimentais. Para isso, consideram-se os dados antes da remoção dos controles de qualidade. Como resultado, espera-se que os QCs apresentem agrupamento consistente, enquanto as replicatas devem se posicionar próximas entre si. Tal comportamento foi avaliado para todos os conjuntos e, a título de ilustração, a Figura 8 representa tal análise realizada para o conjunto N2. O resultado indicou que o tratamento de dados foi adequado e não gerou distorções significativas na estrutura original dos dados.

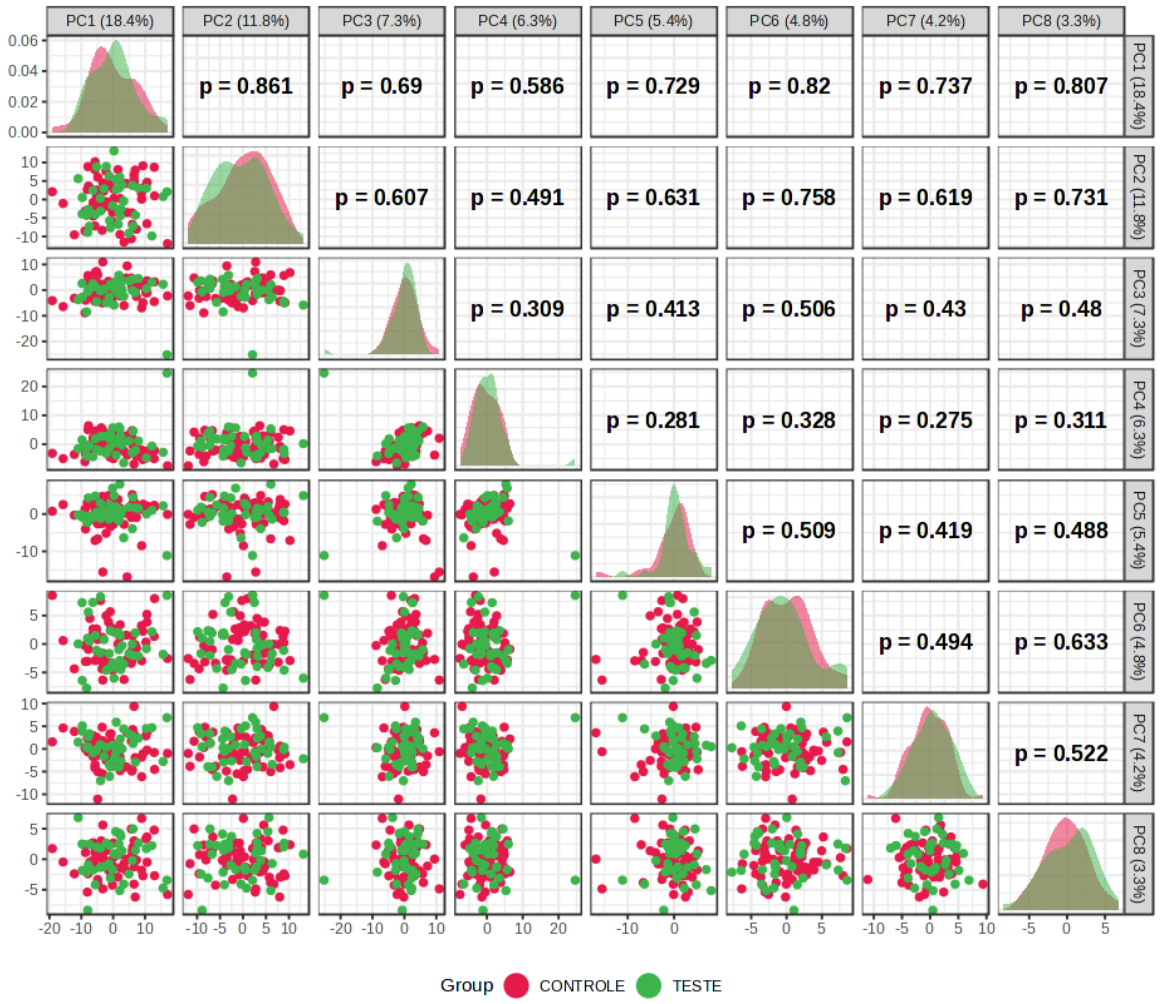
Figura 8 - Representação tridimensional (PC1 vs. PC2 vs. PC3) dos escores dos componentes principais do conjunto N2 demonstrando a clusterização das amostras de controle de qualidade.



Fonte: MetaboAnalyst (2026), dados processados pela autora.

As PCs são organizadas em ordem decrescente de variância explicada, ou seja, a primeira componente é traçada no sentido da maior variação no conjunto de dados, já a segunda, é projetada de maneira ortogonal à primeira e assim por diante. Essa variância explicada se refere exatamente à quantidade de informações capaz de ser descrita pela componente (Ferreira, 2022). A Figura 9, apresenta a porcentagem dos dados representada por cada componente. Neste caso foi escolhido um exemplo aleatório entre os conjuntos de dados processados até então, e suas 8 primeiras PCs.

Figura 9 - Gráfico de variância explicada pelos oito primeiros componentes principais (PCs). Dados extraídos do processamento de N2: PC1 (18.4%), PC2 (11.8%), PC3 (7.3%), PC4 (6.3%), PC5 (5.4%), PC6 (4.8%), PC7 (4.2%) e PC8 (3.3%).



Fonte: MetaboAnalyst (2025), dados processados pela autora.

Além da porcentagem explicada, é possível observar as distribuições de densidade para cada componente principal, posicionados na diagonal, nas quais cada cor representa um grupo experimental. A sobreposição dessas distribuições indica similaridade entre os grupos naquela componente específica, enquanto separações visuais sugeriram diferenças nos perfis metabólicos. Adicionalmente, os valores de p apresentados na porção superior correspondem a testes aplicados aos escores das componentes principais, também com o objetivo de avaliar discrepâncias. Valores inferiores a 0,05 indicam alterações estatisticamente significativas, enquanto valores superiores a esse limiar sugerem ausência de distinção.

Mesmo com o elevado número de componentes, apenas 61,5% seria explicado para este conjunto. Comportamento semelhante foi observado também para P1, P2, P3, N1 e N3, sendo os dados de variância expostos na Tabela 3. De acordo com Lovatti (2019), é desejado que as primeiras duas ou três componentes sejam suficientes para capturar a maior parte da

variação presente nos dados, de maneira que uma visualização em gráficos bidimensionais ou tridimensionais seja possível.

Tabela 3 - Percentual de variância explicada pelas componentes principais.

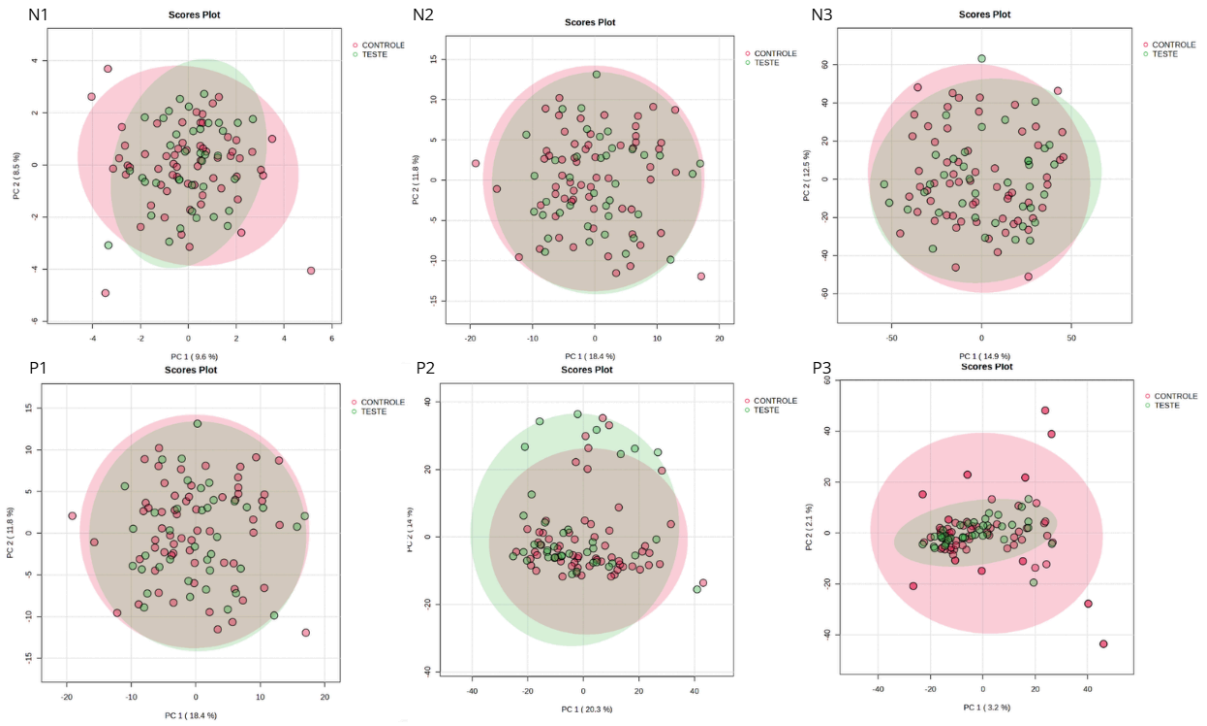
	Variância explicada para cada conjunto de dados					
n° PCs	**N1 (%)**	**N2 (%)**	**N3 (%)**	**P1 (%)**	**P2 (%)**	**P3 (%)**
1	9,6	18,4	14,9	19,6	20,3	3,2
2	8,5	11,8	12,5	14,8	14,0	2,1
3	8,4	7,3	4,8	8,4	8,6	2,0
4	7,5	6,3	3,9	6,6	6,6	1,9
5	6,8	5,4	3,4	4,9	5,3	1,9
6	6,2	4,8	2,9	3,6	3,7	1,9
7	6,1	4,2	2,8	3,0	3,2	1,9
8	5,6	3,3	2,4	2,7	3,0	1,8
Variância explicada acumulada	58,7	61,5	47,6	63,6	64,7	16,7

Fonte: Elaborada pela autora (2026).

Ao observar os valores de variância explicada total para este número de componentes, é possível realizar comparações diretas entre conjuntos que passaram pelo mesmo método de otimização de parâmetros de processamento, ou seja, N1 vs. P1 e N2 vs. P2. A diferença sugere que amostras submetidas à ionização ESI+ tem sua variabilidade metabólica melhor capturadas. Entretanto, o mesmo não se aplica à comparação N3 vs. P3, já que P3 apresenta um valor inferior e bastante anômalo quando comparado aos demais.

Ainda que uma porcentagem baixa dos dados tenha sido explicada por PC1 e PC2, as representações bidimensionais de todos os conjuntos são apresentadas na Figura 10 para verificar o posicionamento e separação das amostras pertencentes ao grupo Controle e Teste.

Figura 10 - Representação bidimensional (PC1 vs. PC2) dos escores dos componentes principais.



Fonte: MetaboAnalyst (2026), dados processados pela autora.

A proximidade entre as amostras no gráfico de escores reflete similaridade em seus perfis metabólicos. Nesse contexto, não foi observada a formação de agrupamentos distintos entre os grupos avaliados na análise por PCA, o que sugere elevada semelhança global entre as amostras. Esse comportamento pode estar associado tanto à ausência de diferenças metabólicas expressivas quanto à elevada complexidade da matriz urinária, que é influenciada por fatores individuais, como dieta, medicação e variabilidade fisiológica.

Embora não tenham sido identificadas diferenças globais evidentes, não se pode descartar a ocorrência de variações mais sutis em metabólitos específicos. Dessa forma, essas possíveis diferenças foram posteriormente investigadas por meio da aplicação de modelos supervisionados e análises estatísticas univariadas, conforme descrito nas seções seguintes.

Assim, a PCA cumpriu seu papel como ferramenta exploratória inicial, permitindo a visualização da distribuição dos dados, embora não tenha sido suficiente para discriminar de forma clara os grupos experimentais.

4.5.2 PARTIAL LEAST SQUARES - DISCRIMINANT ANALYSIS

Após a análise exploratória dos dados por meio de PCA, foi avaliada a aplicação de um método de classificação supervisionado, o *Partial Least Square - Discriminant Analysis*

(PLS-DA), com o objetivo de explorar a capacidade de discriminação entre os grupos e identificar possíveis metabólitos diferenciais. Nele, a calibração ocorre a partir da busca de uma relação direta entre a resposta instrumental e a propriedade de interesse (Santana *et al.*, 2020), enquanto a avaliação é feita por meio da subdivisão do conjunto de amostras, parte em treinamento do modelo, parte para teste e validação. Com o modelo treinado, é feita a verificação de erros e acertos na classificação das amostras separadas para teste. A partir dos resultados obtidos na etapa de classificação da subdivisão de teste do conjunto, o coeficiente de determinação, expresso por R^2, e a capacidade preditiva, expressa por Q^2, são verificadas.

Segundo da Rosa (2023), para matrizes biológicas, os parâmetros de avaliação do modelo são relativamente permissivos, sendo considerados aceitáveis valores de $R^2 > 0.7$ e $Q^2 > 0.4$. Apesar disso, os valores encontrados para os dados processados no presente trabalho e compilados na Tabela 4, indicam mau desempenho. O conjunto P3 apresentou valores para Q^2 um pouco acima dos demais conjuntos, porém todos abaixo do aceitável. Os valores negativos para estes parâmetros em todos os casos, indicam *overffiting*. Ainda que subajuste e sobreajuste possam ser evitados pela verificação da quantidade adequada de variáveis latentes, foi observado que para qualquer número de componentes a capacidade preditiva do modelo não seria satisfatória, fazendo dele incapaz de ser aplicado para outros dados.

Tabela 4 - Métricas de desempenho dos modelos PLS-DA.

	Componente 1		**Componente 2**		**Componente 3**	
Conjunto	**R^2**	**Q^2**	**R^2***	**Q^2***	**R^2***	**Q^2***
N1	0.261	-0.106	0.300	-0.067	0.304	-0.093
N2	0.197	-0.250	0.307	-0.278	0.441	-0.359
N3	0.245	-0.368	0.605	-0.510	0.744	-0.327
P1	0.123	-0.091	0.311	-0.341	0.473	-0.355
P2	0.124	-0.060	0.293	-0.299	0.459	-0.471
P3	0.951	-0.052	0.991	-0.011	0.993	-0.006

Fonte: Elaborado pela autora (2026).

*Os valores de R^2 e Q^2 são apresentados de forma cumulativa em função do número de componentes latentes.

Em razão do desempenho insatisfatório dos modelos elaborados, evidenciados pelos valores de ajuste e capacidade preditiva abaixo do limite considerado aceitável, optou-se por não incluir as representações gráficas correspondentes ao corpo principal do trabalho. Entretanto, a fim de garantir a documentação completa dos resultados, elas são disponibilizadas no Apêndice B.

4.5.3 FOLD CHANGE

A análise de *Fold Change* (FC) tem como objetivo fornecer uma medida da magnitude da diferença entre grupos, permitindo identificar *features* que apresentem variações relevantes de intensidades nas condições avaliadas pelo estudo. Tradicionalmente, essa abordagem tem sido amplamente empregada em estudos de genômica (Dalman *et al.*, 2012; Vaes; Khan; Mombaerts, 2014), mas, recentemente, outras ciências ômicas têm incorporado esta ferramenta univariada para auxiliar na discriminação da enorme gama de variáveis que tipicamente compõem os conjuntos de dados.

Existem diferentes expressões para o cálculo do *Fold Change*, como é evidenciado por Lötsch, Kringel e Ultsch (2024), que constataram em seu estudo que apesar da diversidade de expressões matemáticas encontradas na literatura, os resultados fornecidos por elas são, em geral, semelhantes. Para a realização do cálculo, foram empregados os dados anteriores à etapa de escalonamento e transformação, já que estes procedimentos impactam substancialmente os valores absolutos das médias das variáveis. Dessa forma, foi averiguada a variação relativa entre os grupos utilizando dados em sua escala original, considerando a razão entre as médias das observações dos grupos Controle e Teste, conforme demonstrado na Equação 1.

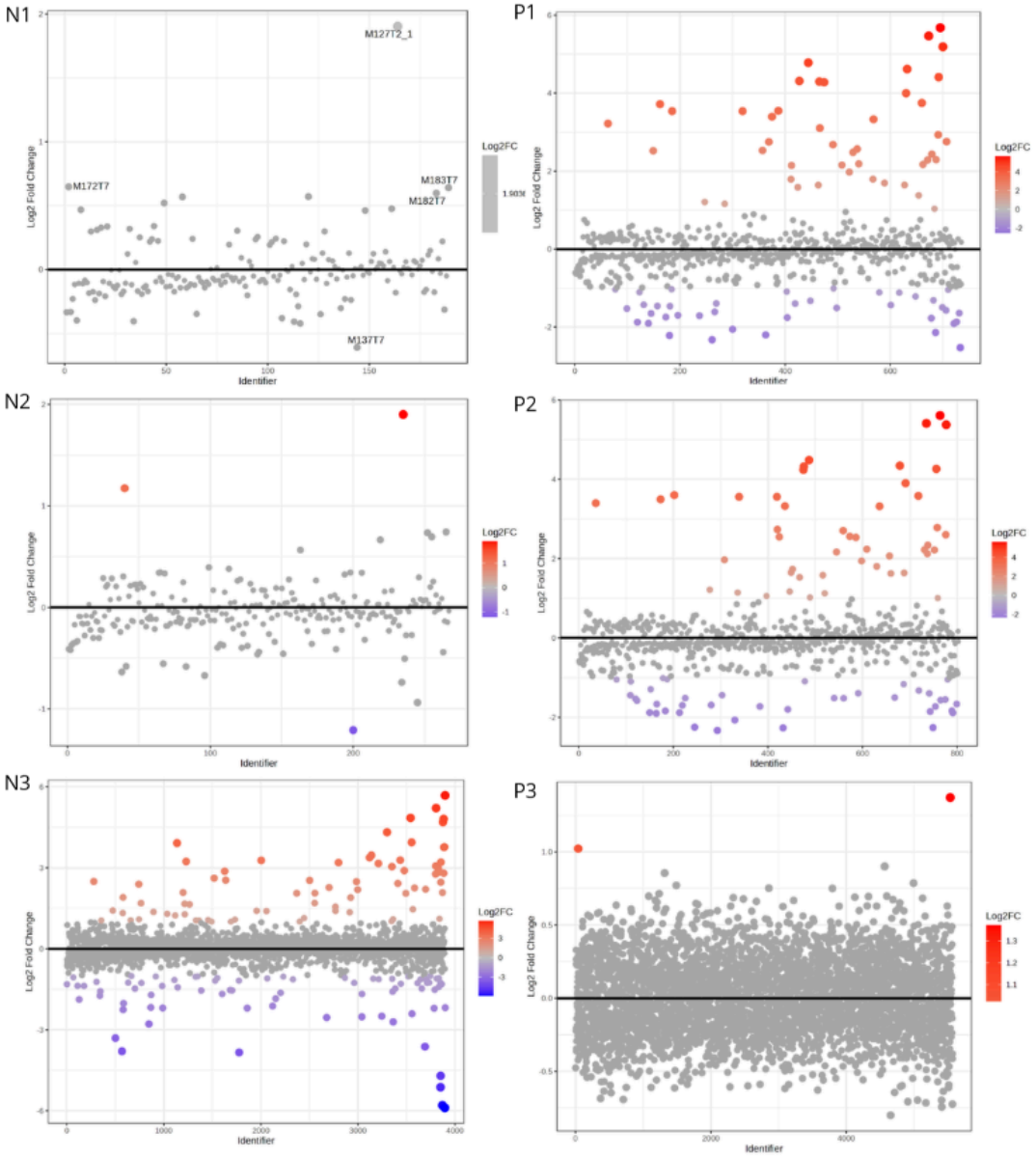
$$FC = X$$

$$\frac{Média_{Controle}}{Média_{Teste}} > X \; ou \; \frac{Média_{Controle}}{Média_{Teste}} < \frac{1}{X} \tag{1}$$

O valor representado por X na fórmula corresponde ao limiar definido como razão mínima para diferenciação de grupos. O valor 1 indica ausência de diferença, sendo, portanto, necessário estabelecer limiares superiores. A Figura 11 apresenta a visualização dos resultados quando o limiar de FC foi definido como 2. Nela, os pontos coloridos se referem a *features* que possuem razões de médias superiores a 2 e, portanto, se diferenciam nos dois

grupos. Embora não representado graficamente, também foi avaliado um limiar mais permissivo (FC ≥ 1,2), para o qual não foram observados *features* estatisticamente significativos entre os grupos ($p < 0,05$).

Figura 11 - Gráfico de *Fold Change* com limiar de diferenciação igual a 2.



Fonte: MetaboAnalyst (2026), dados processados pela autora.

Observa-se que no modo de ionização negativa, a maior parte dos features diferenciais se concentra em N3, enquanto o conjunto N1 não apresenta features com *fold change* superior a 2 e N2 apresenta apenas 3. Em contraste, no modo de ionização positiva, o conjunto P3 possui somente 2 *features* diferenciais, enquanto P1 e P2 uma quantidade mais expressiva. Com o objetivo de identificar quais *features* são representados pelos valores de destaque, o

MetaboAnalyst disponibiliza uma tabela contendo os valores de FC e os respectivos features IDs.

Atrelado à análise de FC, uma abordagem complementar de visualização de diferenças entre grupos é o *Volcano Plot*. No qual, um único gráfico integra o FC, geralmente expresso em $\log_2$, e a significância estatística representada por $-\log_{10}$ do p-valor (sob ajuste de *False Discovery Rate*, FDR). Essa integração facilita a identificação de *features* como candidatos a metabólitos diferenciais. No presente estudo, essa análise foi aplicada aos conjuntos de dados, entretanto, considerando o critério estatístico adotado (p-valor < 0,05), não foi observada diferenciação significativa entre os grupos controle e teste por nenhum *feature*, uma vez que os valores de *p* permaneceram abaixo do limiar estabelecido para discriminação. Por essa razão, mesmo considerando a diferença das médias dos grupos apontadas pelo FC, a falta de distinção estatística apontada pelo p-valor, sugere que não há possibilidade de distinção.

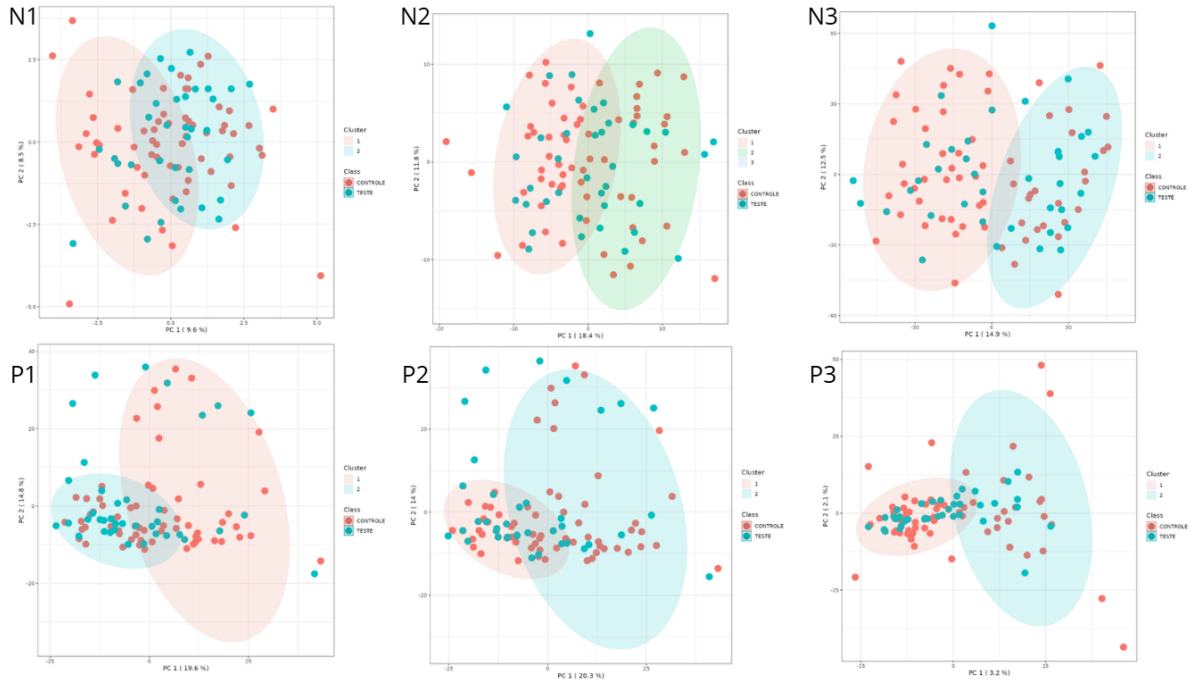
4.5.4 ANÁLISE DE AGRUPAMENTO

A análise de *cluster* surge como uma solução eficaz para lidar com grandes volumes de dados, promovendo sua categorização em pequenos grupos ou subconjuntos com base em semelhanças observadas entre as amostras (Nunes, 2016). A função distância é o principal critério utilizado para quantificar o grau de semelhança ou diferença entre elementos: quanto menor a distância, maior a similaridade (Linden, 2009). Além disso, a similaridade intra-cluster é maximizada e a similaridade inter-cluster minimizada (Fonseca; Beltrame, 2010).

O *K-means*, também conhecido como K-médias, é um dos algoritmos mais empregados na clusterização. Nele, os indivíduos são iterativamente alocados pela similaridade com o ponto médio do grupo e, de acordo com que vão sendo reagrupados, este centróide vai sendo recalculado. Apesar de sua ampla aplicação na análise exploratória de dados, o *K-means* apresenta algumas limitações importantes. Uma delas é a dependência da escolha inicial dos centróides, já que este fator influencia diretamente os resultados obtidos (Souza *et al*., 2021). Então, usualmente o número de *clusters* (k) fornecido ao algoritmo depende quantas categorias de amostras estão contidas nos dados de entrada. Além disso, há uma considerável sensibilidade à presença de ruídos nos dados, visto que pequenos desvios podem impactar significativamente os cálculos das médias dos grupos (Fonseca; Beltrame, 2010).

Na Figura 12, é representada a tentativa de separação em dois *clusters*, referentes às duas classes de amostras – Teste e Controle. Os resultados para k = 2 revelaram que os dados obtidos por diferentes modos de ionização são semelhantes entre si, entretanto, em todos os casos houve uma significativa sobreposição entre os grupos, indicando uma discriminação limitada dos agrupamentos. Considerando que todos os agrupamentos naturais confundem a classificação das amostras de maneira correta de acordo com o critério utilizado neste trabalho, é possível que haja outro fator que melhor explique a similaridade entre as amostras.

Figura 12 – Análise de agrupamento por K-means (k=2) das amostras de urina. Disposição dos clusters no espaço multivariado (PC1 vs. PC2).



Fonte: MetaboAnalyst (2026), dados processados pela autora.

Para uma visualização mais intuitiva da estrutura global dos dados, outras abordagens se mostram mais adequadas. É nesse contexto que a *Hierarchical Cluster Analysis* (HCA) ganha destaque. Este algoritmo categoriza objetos por semelhança, tendo como representação gráfica uma árvore hierárquica multinível que permite identificar, de forma visual, os níveis de semelhança entre os elementos, auxiliando na decisão sobre o número ideal de agrupamentos. Quanto maior a dissimilaridade entre amostras, mais distante ocorre a junção (ou separação) entre elas (Ranjbarzadeh *et al*., 2023).

Como é descrito por Linden (2009), as abordagens hierárquicas podem ser classificadas como aglomerativas ou divisivas. A abordagem aglomerativa inicia o processo

com cada elemento formando um *cluster* individual e vai unindo os elementos de acordo com sua similaridade até que todos estejam em um único grupo. Enquanto a divisiva segue o caminho contrário: tendo início com um único *cluster* contendo todos os elementos e, iterativamente, o divide em subgrupos menores até atingir um critério de parada, como o número desejado de clusters. Independente da escolha, a identificação de subgrupos e a interpretação da similaridade entre elementos é facilitada pelo dendrograma. Mesmo não sendo o caso deste estudo, outra vantagem a ser destacada é que não exige a predeterminação do número de *clusters* a ser gerado, o que confere maior flexibilidade, permitindo que a escolha do número de grupos seja feita posteriormente com base em critérios analíticos ou visuais (Souza *et al.*, 2021).

A eficácia dos métodos de agrupamento está intimamente relacionada à escolha de métricas, parâmetros e estratégias de agrupamento. Ainda que algoritmos diferentes utilizem a mesma métrica e o mesmo método de ligação, resultados equivalentes só serão obtidos se as decisões de agrupamento em cada iteração forem idênticas (Metz; Monard, 2006). Muitos algoritmos atuais incorporam técnicas adicionais para melhorar a eficiência computacional e a qualidade dos agrupamentos, como estruturas otimizadas para o cálculo de distâncias, técnicas de redução de dados e até conceitos da teoria dos grafos. Todas essas estratégias podem contribuir significativamente para a adequação do modelo ao problema em análise.

Apesar disso, a identificação das semelhanças/diferenças entre as amostras a partir das informações que se têm sobre elas e o posicionamento relativo pode dizer qual critério melhor de adequa para classificação, ou não, dos grupos. De acordo com o parâmetro sugerido como diferenciador previamente, esperava-se que os dados se organizassem em duas categorias. No entanto, os dendrogramas resultantes (Figuras 13, 14, 15, 16, 17 e 18) revelaram separação pouco definida entre as categorias, com amostras Teste e Controle distribuídas de maneira desordenada, indo de acordo com o esperado após a análise do *k-means*.

Em busca de estabelecer um critério capaz de explicar as similaridades observadas, buscou-se relacionar a organização das amostras no dendrograma com as informações disponíveis sobre os indivíduos doadores das amostras, captadas durante a anamnese dos voluntários (Tabela 5).

Tabela 5 – Características do conjunto amostral para estudos relacionados à COVID 19.

Variáveis	Grupo Teste (n = 38)	Grupo Controle (n = 62)
Resultado RT-PCR	Positivo	Negativo
Idade (Q_{min} – $Q_{máx}$)	(21 – 72)	(21 – 77)

Sexo masculino, *n*	16	38
Sexo feminino, *n*	22	24
Sintomático, *n (%)*	30 (78.9)	28 (45.6)
Vacinados, *n (%)*	30 (78.9)	60 (96.7)
Comorbidades, *n (%)*	10 (26.3)	10 (26.3)

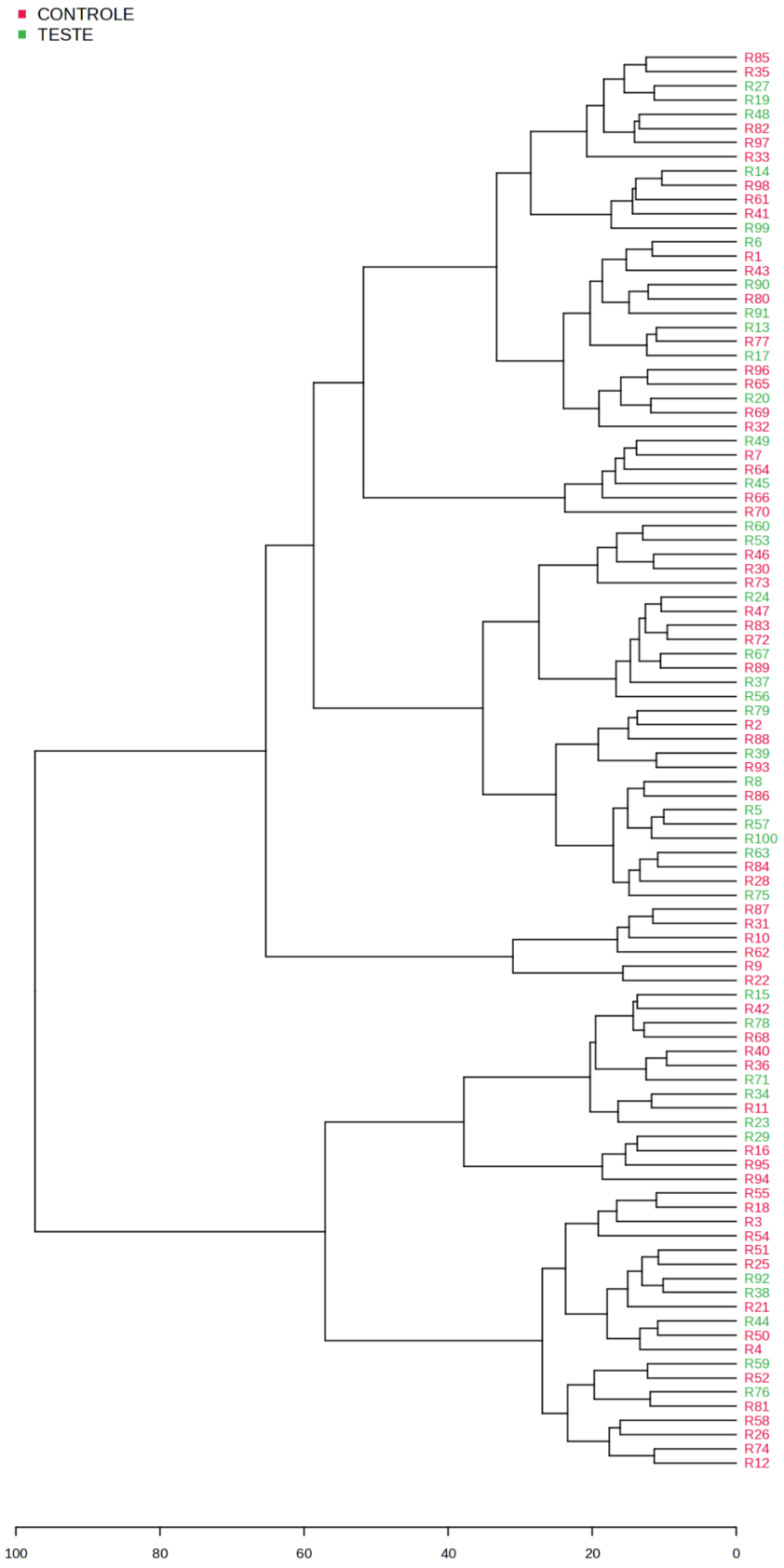
Q_{min} = valor mínimo; Q_{max} = Valor máximo.

Fonte: Adaptado de Moreira, 2024.

Dentre os sintomas relatados pelos indivíduos, tosse, coriza, dor de cabeça e dores no geral foram os mais frequentes, mas febre, perda de olfato, diarreia, náusea, irritação na garganta e desconforto respiratório também foram relatados. Já em relação a comorbidades, haviam pacientes que sofriam de hipertensão, diabetes, asma, anomalias respiratórias e doenças cardiovasculares.

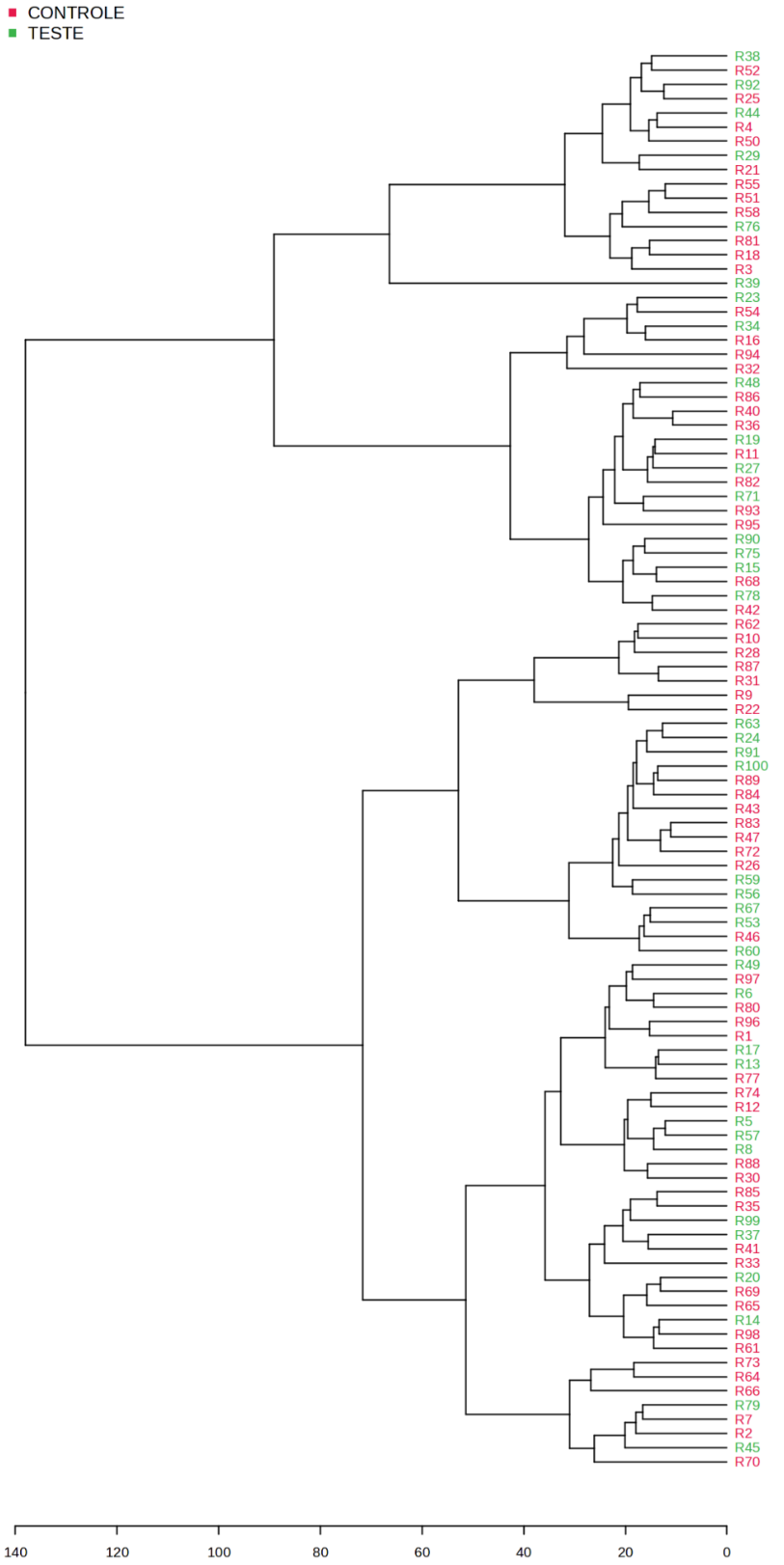
As variáveis como gênero, idade, tipo sanguíneo, fator RH, esquema vacinal para COVID-19, histórico de arbovirose, presença de doenças crônicas, tabagismo e uso de medicamentos foram consideradas, no entanto, a classificação das amostras com base nesses critérios, avaliados individualmente não apresentou resultado satisfatório. De forma semelhante ao observado para o critério de RT-PCR positivo ou negativo, amostras que se esperava estarem próximas no dendrograma mostraram-se distantes, sugerindo que outros fatores refletidos no perfil metabólico e não contemplados pelas informações disponíveis neste estudo, também exercem influência sobre a classificação das amostras.

Figura 13 - Dendrograma gerado por análise hierárquica do conjunto N1.



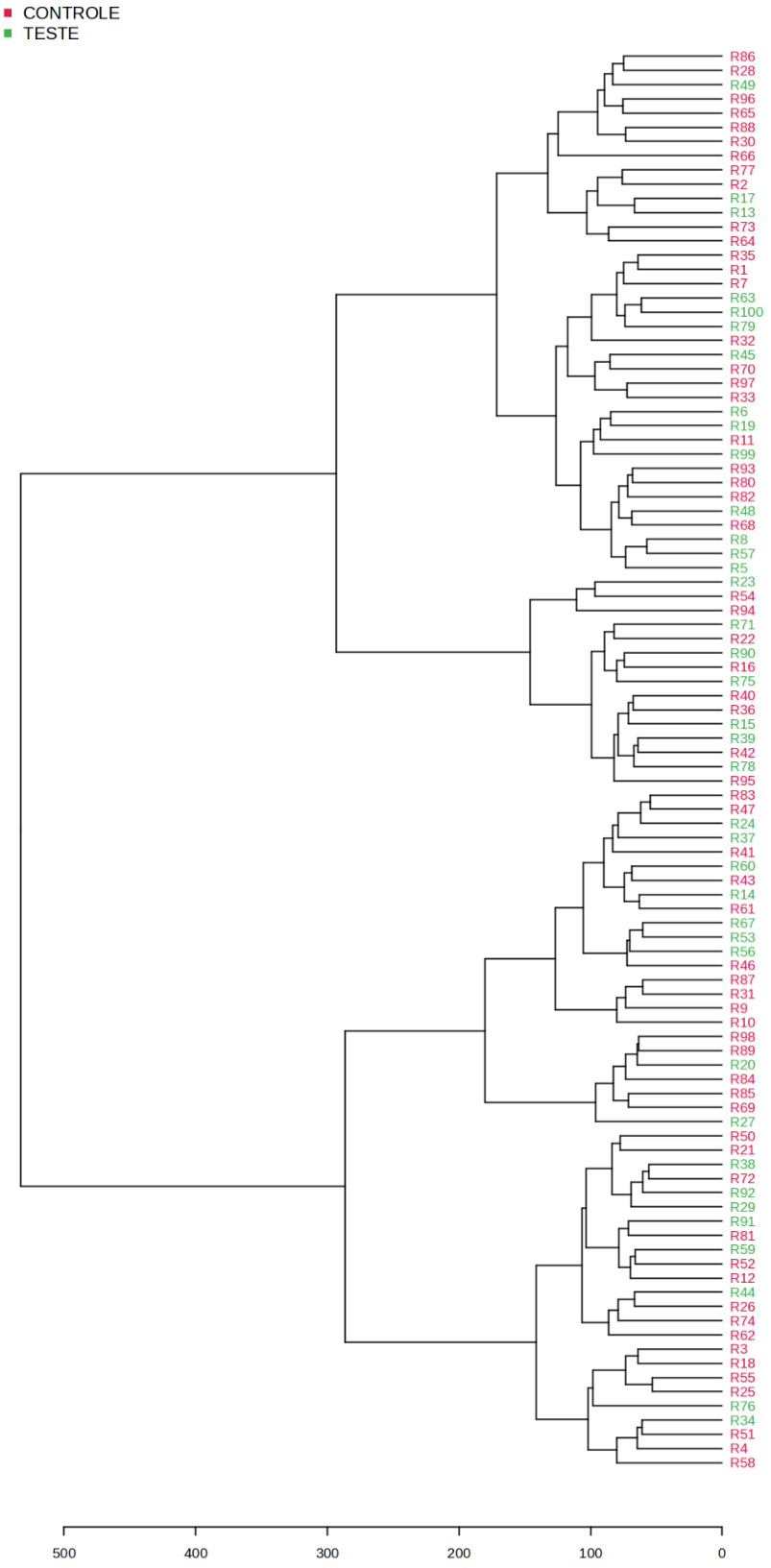
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 14 - Dendrograma gerado por análise hierárquica do conjunto N2.



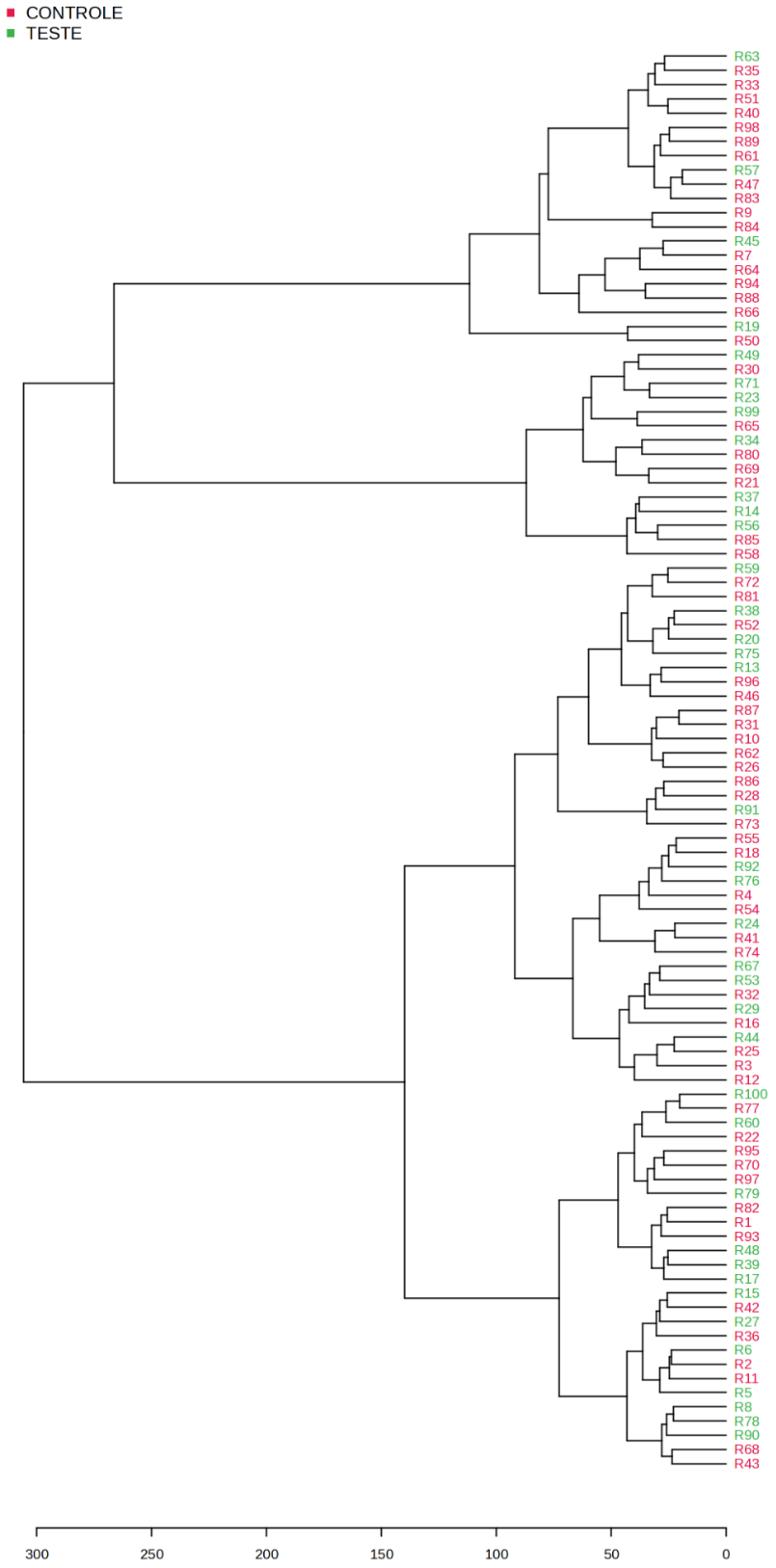
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 15 - Dendrograma gerado por análise hierárquica do conjunto N3.



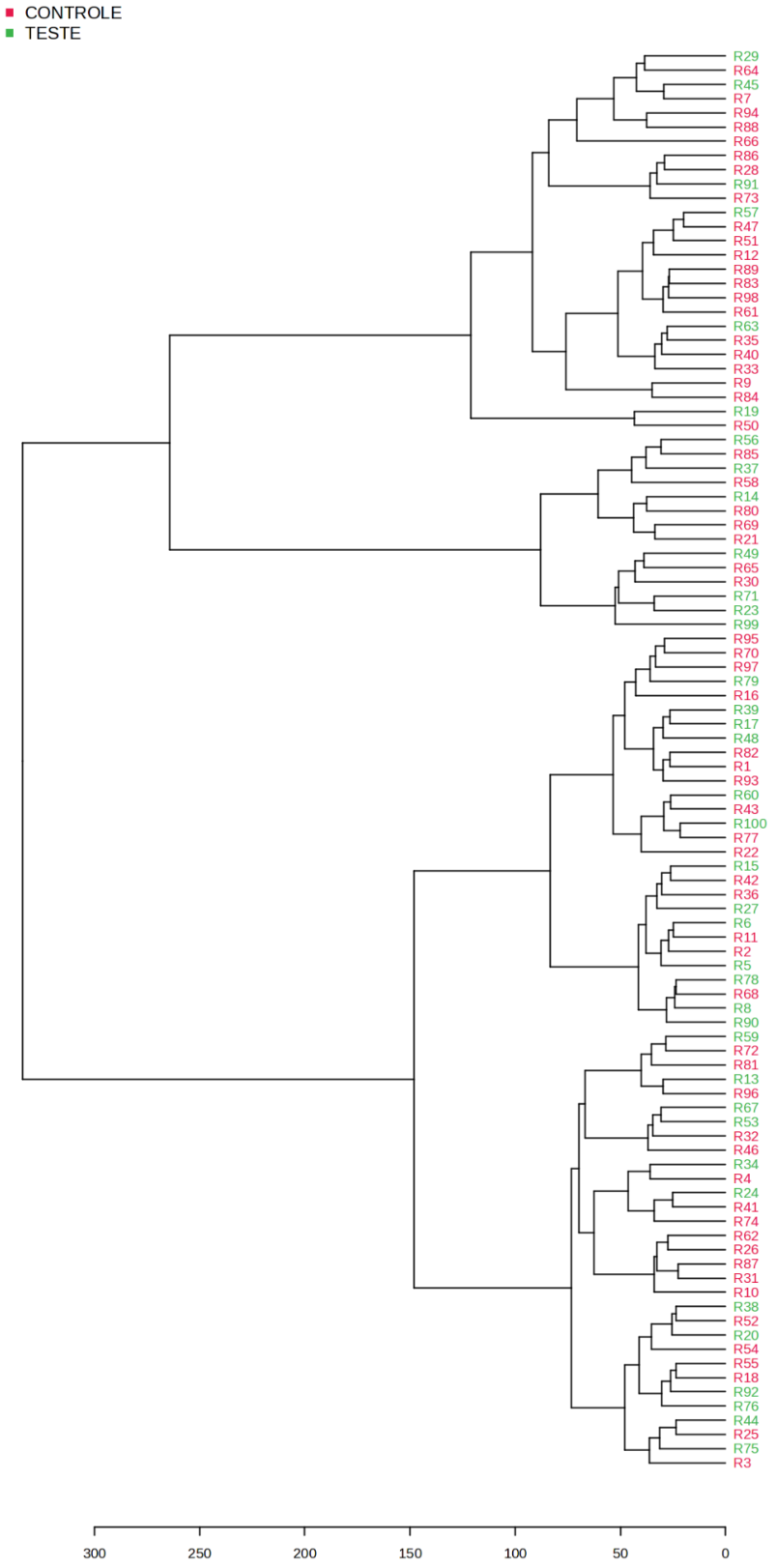
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 16 - Dendrograma gerado por análise hierárquica do conjunto P1.



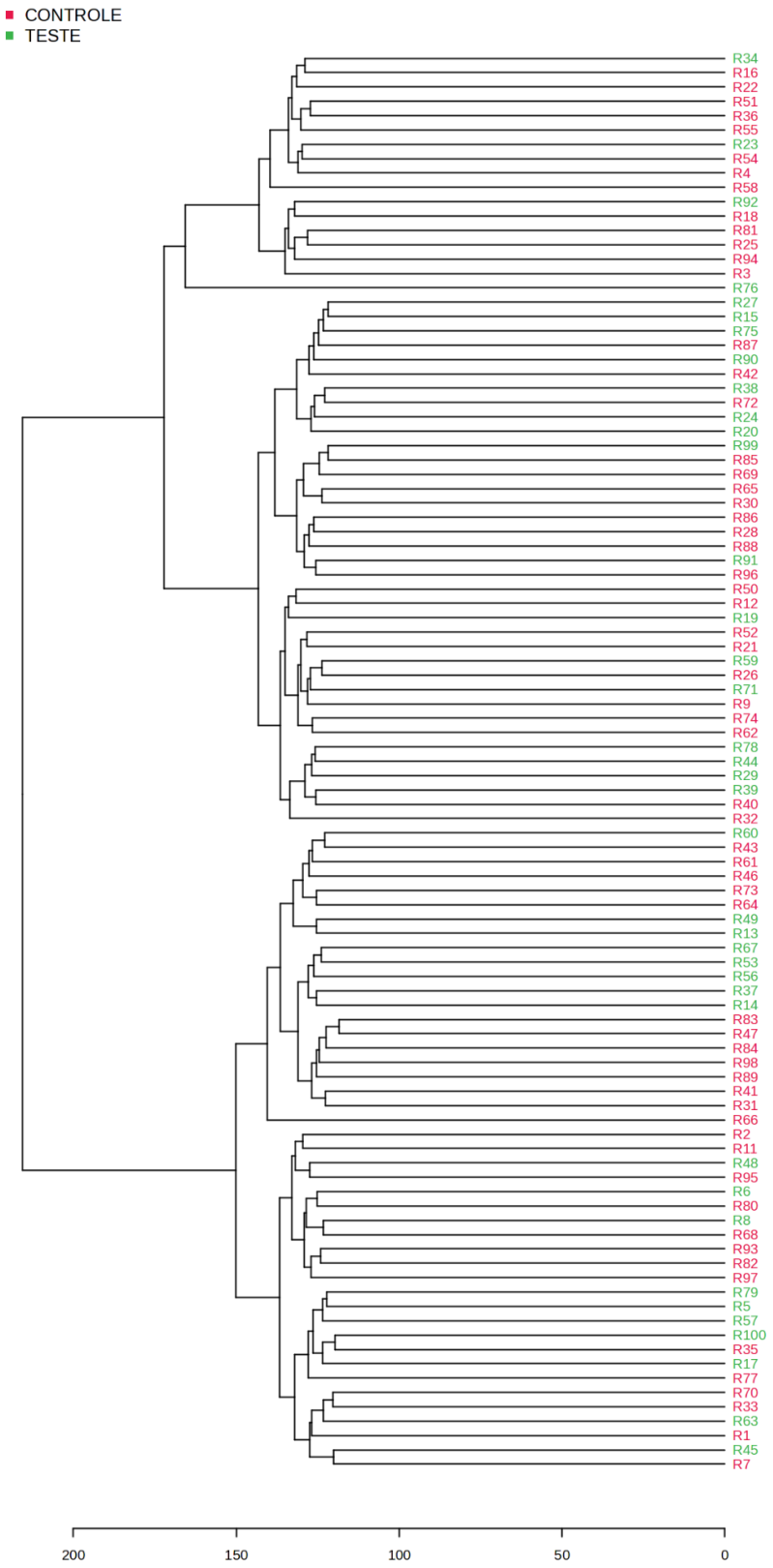
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 17 - Dendrograma gerado por análise hierárquica do conjunto P2.



Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 18 - Dendrograma gerado por análise hierárquica do conjunto P3.



Fonte: MetaboAnalyst (2026), dados processados pela autora.

5 CONSIDERAÇÕES FINAIS

O presente trabalho teve como objetivo investigar o potencial da metabolômica global, baseada em espectrometria de massas de alta resolução, para identificação de metabólitos diferenciais associados à COVID-19 em amostras de urina. Embora essa matriz apresente diversas vantagens práticas, os resultados obtidos evidenciam que o potencial teórico desta aplicação é acompanhado por desafios metodológicos significativos.

Um dos principais pontos discutidos ao longo do estudo se referiu à elevada variabilidade intrínseca da urina, fortemente influenciada pelo estado de hidratação, hábitos alimentares e características fisiológicas individuais. Apesar da avaliação de diferentes estratégias de pré-processamento e processamento dos dados, não foi possível identificar metabólitos que fossem significativamente distintos nos dois grupos em estudo.

Por se tratar de um estudo de metabolômica global, sua natureza foi essencialmente exploratória. A heterogeneidade do conjunto amostral analisado, ao mesmo passo que, poderia resultar em metabólitos representativos caso encontrados, já que perpassaria faixa etária ampla, distintas comorbidades, estados vacinais e sintomas relacionados à doença, impõe desafios substanciais à análise quimiométrica. A ausência de diferenciação estatisticamente entre os grupos Teste e Controle (p-valor > 0.05), mesmo diante da detecção de milhares de *features*, indica que, sob as condições de avaliação descritas, a urina não se mostrou capaz de expressar de maneira consistente alterações associadas à COVID-19.

É importante reconhecer que estudos dessa natureza estão intrinsecamente sujeitos a esse tipo de limitação. A ausência de discriminação metabólica sustentada pelos diversos métodos estatísticos descritos não invalida a abordagem, mas contribui para delimitar seus limites de aplicabilidade. Assim, o estudo cumpre o papel de fornecer uma avaliação crítica do uso desta matriz na investigação da COVID-19, oferecendo subsídios metodológicos e conceituais que podem auxiliar futuras pesquisas futuras.

Como perspectiva, destaca-se a necessidade de ampliação do número amostral e a inclusão de variáveis clínicas adicionais, que trariam a possibilidade de melhor estratificação dos grupos e controle de fatores de confusão. Essas estratégias podem contribuir para uma compreensão mais robusta das alterações metabólicas associadas à doença e para o aprimoramento das abordagens metabolômicas aplicadas a essa matriz biológica.

REFERÊNCIAS

ALBA, E. *et al*. Comparison between machine learning algorithms for the identification of seasonally dry tropical forest. **Anuário do Instituto de Geociências**, v. 45, 2022. Disponível em: https://doi.org/10.11137/1982-3908_2022_45_40758. Acesso em: 5 out. 2024.

ALZUBI, J.; NAYYAR, A.; KUMAR, A. Machine learning from theory to algorithms: an overview. **Journal of Physics: Conference Series**, v. 1142, n. 1, 2018. Disponível em: https://doi.org/10.1088/1742-6596/1142/1/012012. Acesso em: 20 abr. 2025.

ARAUJO, B. R.. **Metabolômica untargeted em urina de portadores de Cri Du Chat utilizando cromatografia a gás e espectrometria de massas (GC-MS)**. 2020. Tese (Doutorado em Química Analítica e Inorgânica) – Instituto de Química de São Carlos, Universidade de São Paulo, São Carlos, 2020. doi:10.11606/T.75.2020.tde-23102020-102820.

ATAÍDE, A. R. *et al.* Perfil químico: revisão bibliográfica dos parâmetros urinários. **Revista JRG de Estudos Acadêmicos**, v. 7, n. 15, 2024. Disponível em: https://doi.org/10.55892/jrg.v7i15.1568.

BORGES, R. M. *et al*. Guia para processamento de dados de cromatografia acoplada a espectrometria de massas. **Química Nova**, v. 45, n. 5, 608–620, 2022. Disponível em: https://doi.org/10.21577/0100-4042.20170838. Acesso em: 23 dez. 2025.

CANUTO, G. A. B. *et al*. Metabolômica: Definições, estado-da-arte e aplicações representativas. **Química Nova**, v. 41, n. 1, p. 75–91, jan. 2018.

CHAMBERS, M. C *et al*. A cross-platform toolkit for mass spectrometry and proteomics. **Nat Biotechnol**, v. 30, n. 10, p. 918–20, 2012. Disponível em: https://doi.org/10.1038/nbt.2377.

CLISH, C. B. Metabolomics: an emerging but powerful tool for precision medicine. **Molecular Case Studies**, v. 1, n. 1, p. a000588, 2015.

DA SILVA, T. P. B.; AMARAL, R. R.; MARTINS, A. M. A. Metabolômica em plasma de pacientes com carcinoma hepatocelular (CHC): avaliação de biomarcadores pela espectrometria de massas. Programa de Iniciação Científica – **PIC/UniCEUB: Relatórios de Pesquisa**, v. 3, n. 1, 2017.

DALMAN, M. R. *et al.* Fold change and p-value cutoffs significantly alter microarray interpretations. **BMC Bioinformatics**, v. 13, 2012. Disponível em: doi:10.1186/1471-2105-13-S2-S11.

DOS SANTOS, E. K. P.; CANUTO, G. A. B. Optimizing XCMS parameters for GC-MS metabolomics data processing: a case study. **Metabolomics: Official Journal of the Metabolomic Society**, v. 19, n. 4, p. 26, 2023. Disponível em: https://doi.org/10.1007/s11306-023-01992-1. Acesso em: 3 maio 2025.

DUTTA, D. *et al*. COVID-19 Diagnosis: A Comprehensive Review of the RT-qPCR Method for Detection of SARS-CoV-2. **Diagnostics**, v. 12, n. 6, p. 1–18, 2022.

EBRAHIM, S. H. *et al*. Covid-19 and community mitigation strategies in a pandemic. **Bmj**, v. 368, 2020.

FEITOSA, T. M. O.; CHAVES, A. M.; MUNIZ, G. T. S.; CRUZ, M. C. C.; JUNIOR, I. F. C. Comorbidades e COVID-19: Uma revisão integrativa. **Revista Interfaces: Saúde, Humanas e Tecnologia**, v. 8, n. 3, 2020.

FERREIRA, M. M. C. Quimiometria III-Revisitando a análise exploratória dos dados multivariados. **Química Nova**, v. 45, n. 10, p. 1251-1264, 2022. Disponível em: https://doi.org/10.21577/0100-4042.20170910.

FONSECA, F. C. S.; BELTRAME, W. A. R. Aplicações práticas dos algoritmos de clusterização K-means e Bisecting K-means. **Vitória: UFES**, 2010. Disponível em: https://www.researchgate.net/publication/327121358. Acesso em: 5 maio 2025.

KARKI, K. *et al.* Review on current race for Covid-19 diagnosis. **Biosensors and Bioelectronics: X**, v. 16, 2024. Elsevier Ltd. Disponível em: https://doi.org/10.1016/j.biosx.2023.100432. Acesso em: 5 maio 2025.

LEITE, V. dos S. A. **Metabolômica global de amostras de alho brasileiro (Allium sativum L.) por GC-MS e PS-MS**. 2024. Tese (Doutorado em Química) – Universidade Federal de Viçosa, Rio Paranaíba, 2024.

LINDEN, R. Técnicas de agrupamento. **Revista de Sistemas de Informação da FSMA**, n. 4, p. 18–36, 2009. Disponível em: https://www.researchgate.net/publication/267710538. Acesso em: 4 mar. 2025.

LÖTSCH, J.; KRINGEL, D.; ULTSCH, A. Revisiting Fold-Change Calculation: Preference for Median or Geometric Mean over Arithmetic Mean-Based Methods. **Biomedicines**, v. 12, n. 8, 2024. Disponível em: https://doi.org/10.3390/biomedicines12081639.

LOVATTI, B. P. O. **Métodos de aprendizagem de máquina em Química Analítica: floresta randômica aplicada na avaliação de petróleo.** 2019. Tese (Doutorado em Química) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2019.

MAKOANA, K. M. *et al*. Integration of metabolomics and chemometrics with in-silico and in-vitro approaches to unravel SARS-CoV-2 inhibitors from South African plants. **PLOS ONE**, v. 20, n. 3, e0320415, 2025. Disponível em: https://doi.org/10.1371/journal.pone.0320415. Acesso em: 02 jan. 2026.

MENEZES, M. E.; LIMA, L. M.; MARTINELLO, F. Diagnóstico laboratorial do SARS-CoV-2 por transcrição reversa seguida de reação em cadeia da polimerase em tempo real (RT-PCR). **Rev RBAC**, v. 52, n. 2, p. 122-30, 2020.

METZ, J.; MONARD, M. C. **Estudo e análise das diversas representações e estruturas de dados utilizadas nos algoritmos de clustering hierárquico**. São Carlos: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2006. (Relatórios Técnicos do ICMC, n. 269). Disponível em: https://repositorio.usp.br/directbitstream/8fd381d5-4dce-4385-a3f0-0bb92638af5c/Relat%C3%B3rio%20T%C3%A9cnico_269_2006.pdf. Acesso em: 07 maio 2026. , 2006

MEUMANN, E. M.; ROBSON, J. M. B. Testing for COVID-19: a 2023 update. **Australian Prescriber**, v. 46, n. 1, p. 13–17, jun. 2023. Disponível em: https://australianprescriber.tg.org.au/assets/AP/pdf/p13-Meumann-Robson.pdf. Acesso em: 28 mar. 2025.

MIOT, H. A. Avaliação da normalidade dos dados em estudos clínicos e experimentais. **Jornal Vascular Brasileiro**, v. 16, n. 2, p. 88-91. 2017.

MORALES-ANGULO, C. *et al*. Nasopharyngeal swab for the diagnosis of COVID-19. **Rev ORL**, v. 5, 2020.

MOREIRA, O. B. de O. **Desenvolvimento e Otimização de Estratégias Analíticas Alternativas Aplicáveis ao Diagnóstico de Dengue, Zika, Chikungunya e COVID-19**. 2024. 140 p. Tese (Doutorado em Química) - Departamento de Química, Universidade Federal de Juiz de Fora, Juiz de fora, 2024.

MOREIRA, O. B. de O. *et al*. Factorial design applied to LC-ESI-QTOF mass spectrometer parameters for untargeted metabolomics. **Analytical Methods**, v. 15, n. 20, p. 2512-2521, 2023.

NUNES, D. H. F. Um breve estudo sobre o algoritmo k-means. Dissertação (Mestrado em Matemática) - Faculdade de Ciências e Tecnologia, Universidade de Coimbra, Coimbra, 2016.OLIVEIRA, M. A. L. *et al*. Testes diagnósticos para o SARS-COV-2: uma reflexão crítica. **Química Nova**, v. 45, n. 6, p. 760–766, 2022.

OLIVEIRA, E. de S.; MATOS, M. F.; MORAIS, A. C. L. N. de. Perspectiva de resultados falso-negativos no teste de RT-PCR quando realizado tardiamente para o diagnóstico de covid-19. **InterAmerican Journal of Medicine and Health**, v. 3, p. 1–7, 2020. Disponível em: https://doi.org/10.31005/iajmh.v3i0.90. Acesso em: 6 mar. 2025.

ORGANIZAÇÃO PAN-AMERICANA DA SAÚDE. Histórico da pandemia de COVID-19 - OPAS/OMS | **Organização Pan-Americana da Saúde**. Disponível em: https://www.paho.org/pt/covid19/historico-da-pandemia-covid-19. Acesso em: 07 jun. 2024.

PADOVANI, P. C. **Metabólitos do triptofano como potenciais biomarcadores de injúria renal aguda em amostras de plasma de pacientes com COVID-19 através da metabolômica alvo com HPLC-MS/MS**. 2025. 1 recurso online (132 p.) Dissertação (mestrado) - Universidade Estadual de Campinas (UNICAMP), Instituto de Química, Campinas, SP. Disponível em: 20.500.12733/36552. Acesso em: 01 fev. 2026.

PANG, Z. *et al.* MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation, **Nucleic Acids Research**, v.52, n. W1, W398–W406, 2024. Disponível em: https://doi.org/10.1093/nar/gkae253. Acesso em: 03 jan. 2026.

PEREIRA, A.; DA CRUZ, K. A. T.; LIMA, P. S. Principais aspectos do novo coronavírus sars-cov-2: uma ampla revisão. **Arquivos do MUDI**, v. 25, n. 1, p. 73-90, 2021.

PILON, A. C. *et al*. Metabolômica de plantas: Método e desafios. **Química Nova**, v. 43, n. 3, p. 329–354, mar. 2020.

PINO, F. A. A questão da não normalidade: uma revisão. **Revista de Economia Agrícola**, v. 61, n. 2, p. 17-33, 2014.

RANJBARZADEH, R. *et al.* Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools. **Computers in Biology and Medicine**, v. 152, n. 106405. 2023. Disponível em: https://doi.org/10.1016/j.compbiomed.2022.106405.

ROSA, J. R. da. **Avaliação do perfil metabolômico urinário no binômio diabetes mellitus e SARS-CoV-2**. 2023. Dissertação (Mestrado em Ciências da Saúde) – Universidade São Francisco, Bragança Paulista, 2023.

SANTANA, F. B. de *et al.* Experimento didático de quimiometria para classificação de óleos vegetais comestíveis por espectroscopia no infravermelho médio combinado com análise discriminante por mínimos quadrados parciais: Um tutorial, parte V. **Química Nova**, v. 43, n. 3, p 371-381, 2020. Disponível em: http://dx.doi.org/10.21577/0100-4042.20170480.

SANTOS, L. A. O. *et al*. Análise da taxa de eficácia dos testes sorológicos rápidos para COVID-19 registrados na ANVISA, uma revisão sistemática na literatura. **Research, Society and Development**, v. 10, n. 11, p. e264101119615-e264101119615, 2021.

SCHIAFFINO, M. C. **Desenvolvimento de um método para classificação de comportamentos de ratos Wistar utilizando o algoritmo de aprendizado supervisionado Florestas Aleatórias (Random Forests).** 2020. Dissertação (Mestrado em Neurociência) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2020.

SMITH, C. A. *et al.* Package 'xcms'. **xcms Reference Manual**, 2015.

SOUZA, A. M. de.; POPPI, R. J. Experimento didático de quimiometria para análise exploratória de óleos vegetais comestíveis por espectroscopia no infravermelho médio e análise de componentes principais: um tutorial, parte I. **Química Nova**, v. *35*, n. 1, p. 223–229. 2012. Disponível em: https://doi.org/10.1590/S0100-40422012000100039. Acesso em: 07 out. 2025.

SOUSA, N. F. *et al*. Uma análise comparativa entre algoritmos de agrupamentos de dados. **ENCONTRO DE COMPUTAÇÃO DO OESTE POTIGUAR – ECOP-POCKET**, 2., s.d., v. 5, 2021. Disponível em: https://periodicos.ufersa.edu.br/index.php/ecop. Acesso em: 9 fev. 2025

SOUZA, A. S. R. *et al*. General aspects of the COVID-19 pandemic. **Revista Brasileira de Saúde Materno Infantil**, v. 21, p. 29–45, fev. 2021.

SOUZA, M. de A. **Metabolômica de macrofungos cultivados em torta do caroço de algodão por espectrometria de massas.** 2020. Dissertação (Mestrado em Química) – Universidade Federal de Goiás, Instituto de Química, Goiânia, 2020.

TAUTENHAHN, R. *et al*. XCMS Online: a web-based platform to process untargeted metabolomic data. **Analytical Chemistry**, v. 84, n. 11, p. 5035-5039, 2012.

TRISTÁN, A. I. et al. Metabolomic profiling of COVID-19 using serum and urine samples in intensive care and medical ward cohorts. **Scientific Reports**, v. 14, n. 1, p. 23713, 2024. Disponível em: https://doi.org/10.1038/s41598-024-74641-9. Acesso em: 5 maio 2025.

VAES, E.; KHAN, M.; MOMBAERTS, P. Statistical analysis of differential gene expression relative to a fold change threshold on NanoString data of mouse odorant receptor genes. **BMC Bioinformatics**, v. 15, 2014. Disponível em: doi:10.1186/1471-2105-15-39.

VALDÉS, A.; CIFUENTES, A. Técnicas bioquímicas para la detección de la COVID-19. **Metabolómica en tiempos de COVID**. 2020.

VERAS, G. *et al*. Perfil cientométrico da Quimiometria no Brasil. **Química Nova**, v. *45*, n. 10, p. 1315–1321, 2022. Disponível em: https://doi.org/10.21577/0100-4042.20170930. Acesso em: 21 dez. 2025.

VEROTTI, M. P. *et al*. Testes diagnósticos para COVID-19 registrados na Agência Nacional de Vigilância Sanitária: sensibilidade e especificidade reportadas pelos fabricantes. **Comunicação em Ciências da Saúde**, Brasília, v. 31, p. 217-229, 2020. Supl. 1.

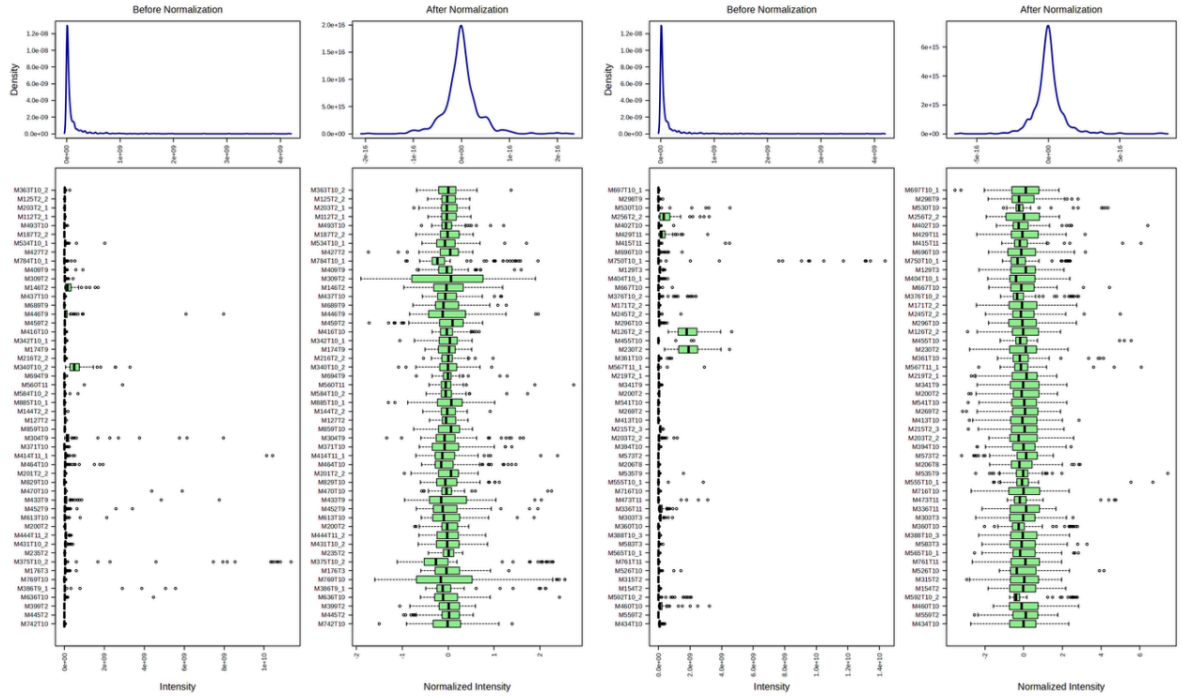
WOLD, S.; ESBENSEN, K.; GELADI, P. Principal component analysis. **Chemometrics and Intelligent Laboratory Systems**, v. 2, n. 1–3, p. 37–52, 1987. Disponível em: https://doi.org/10.1016/0169-7439(87)80084-9. Acesso em: 1 maio 2025.

WORLD HEALTH ORGANIZATION (WHO). Emergencies. Diseases. Coronavirus disease (COVID19). Technical guidance. **Naming the coronavirus disease (COVID19) and the virus that causes it.** 2020. Disponível em: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. Acesso em: 09 jun. 2024.

YÜCE, M.; FILIZTEKIN, E.; ÖZKAYA, K. G. COVID-19 diagnosis —A review of current methods. **Biosensors and Bioelectronics**, v. 172, n. 112752, 2020. Disponível em: https://doi.org/10.1016/j.bios.2020.112752. Acesso em: 20 jun. 2025.

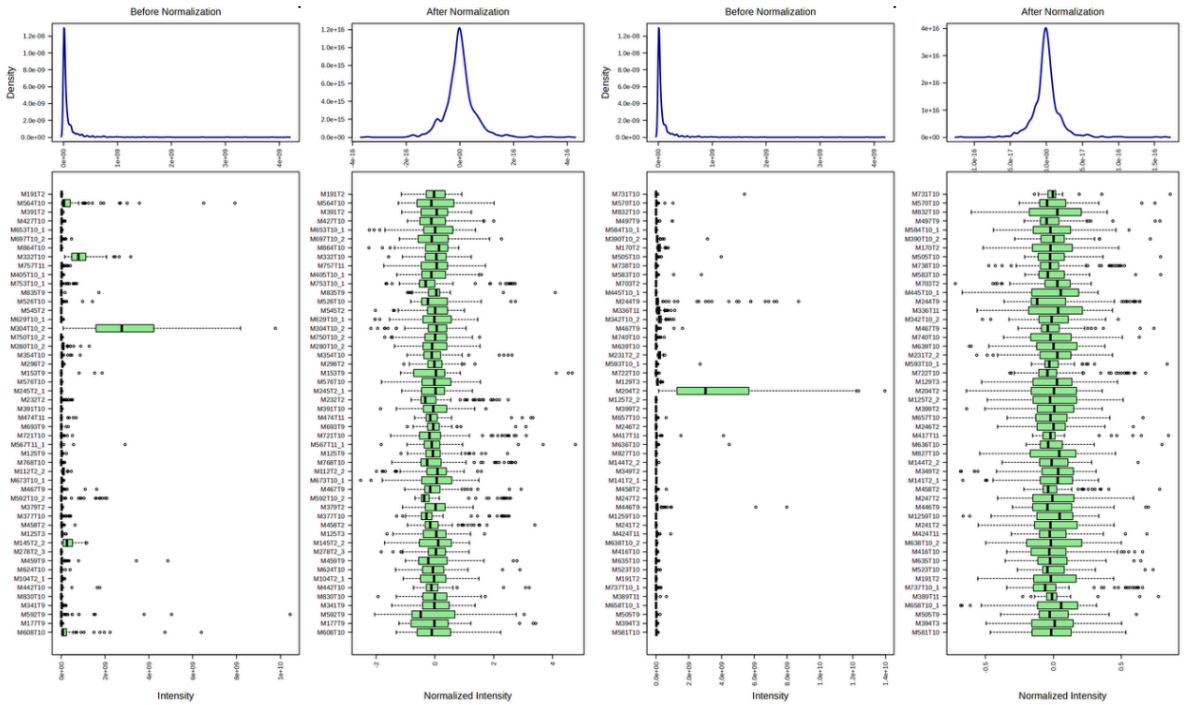
APÊNDICE A - Comparação de representações gráficas dos procedimentos de processamento testado

Figura 19 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 1 e 2, respectivamente.



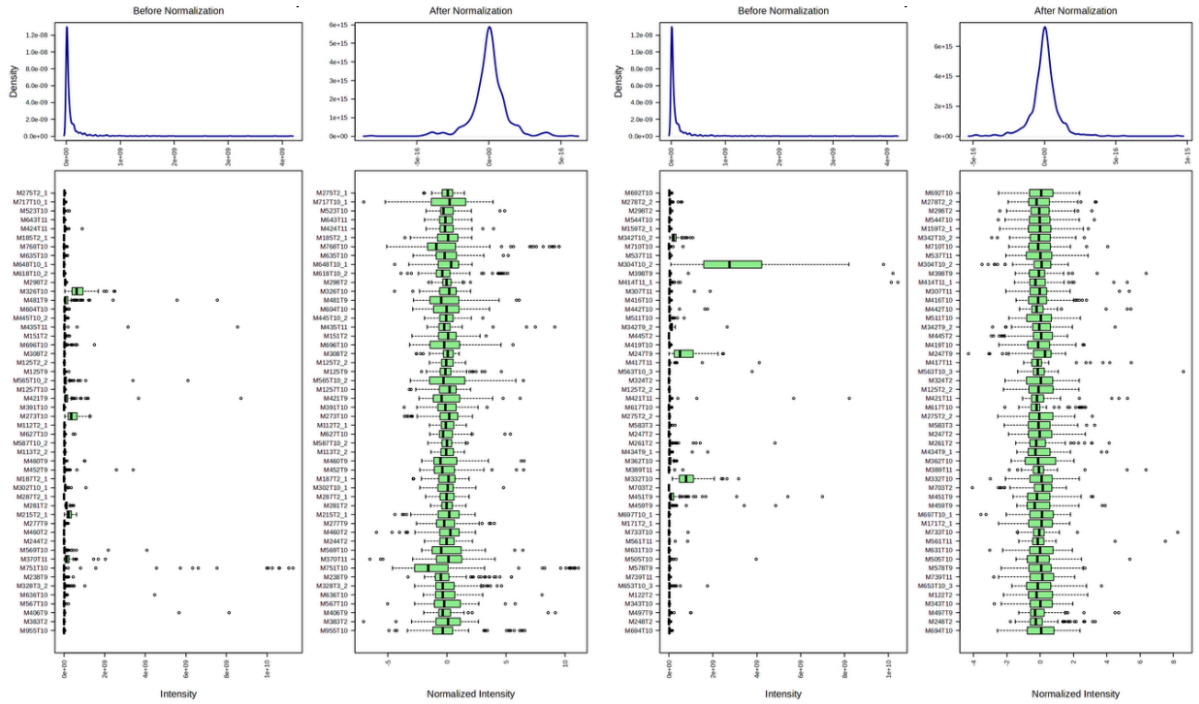
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 20 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 3 e 4, respectivamente.



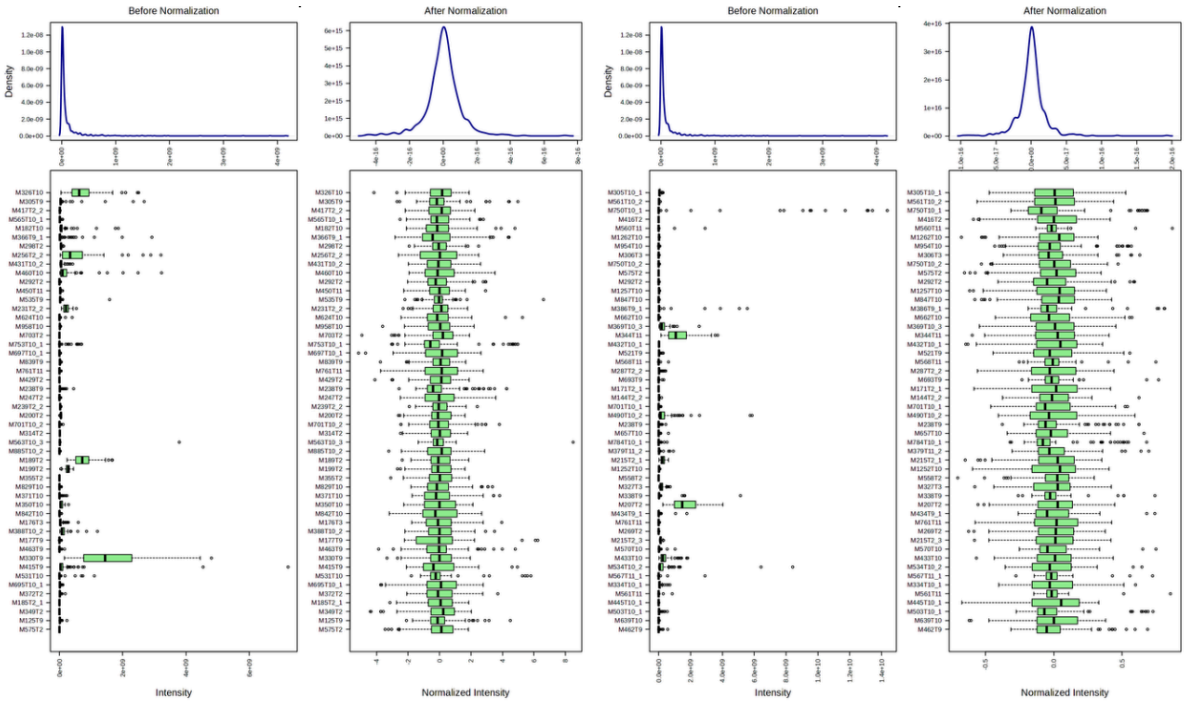
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 21 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 5 e 6, respectivamente.



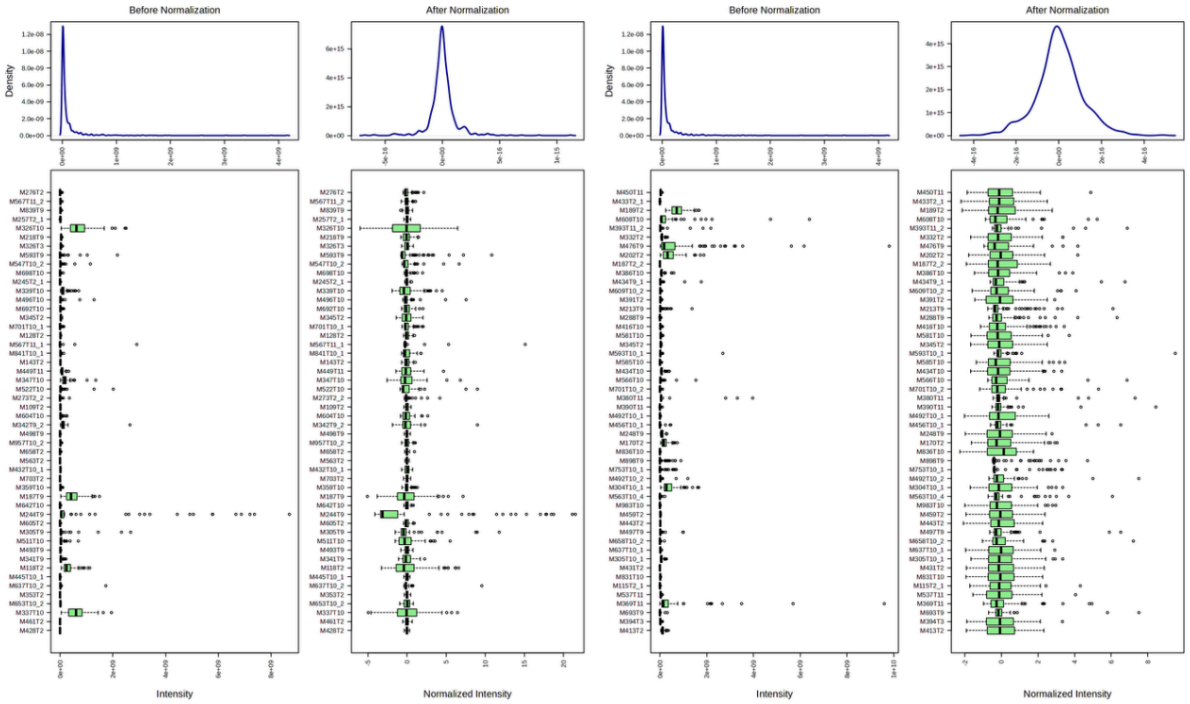
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 22 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 7 e 8, respectivamente.



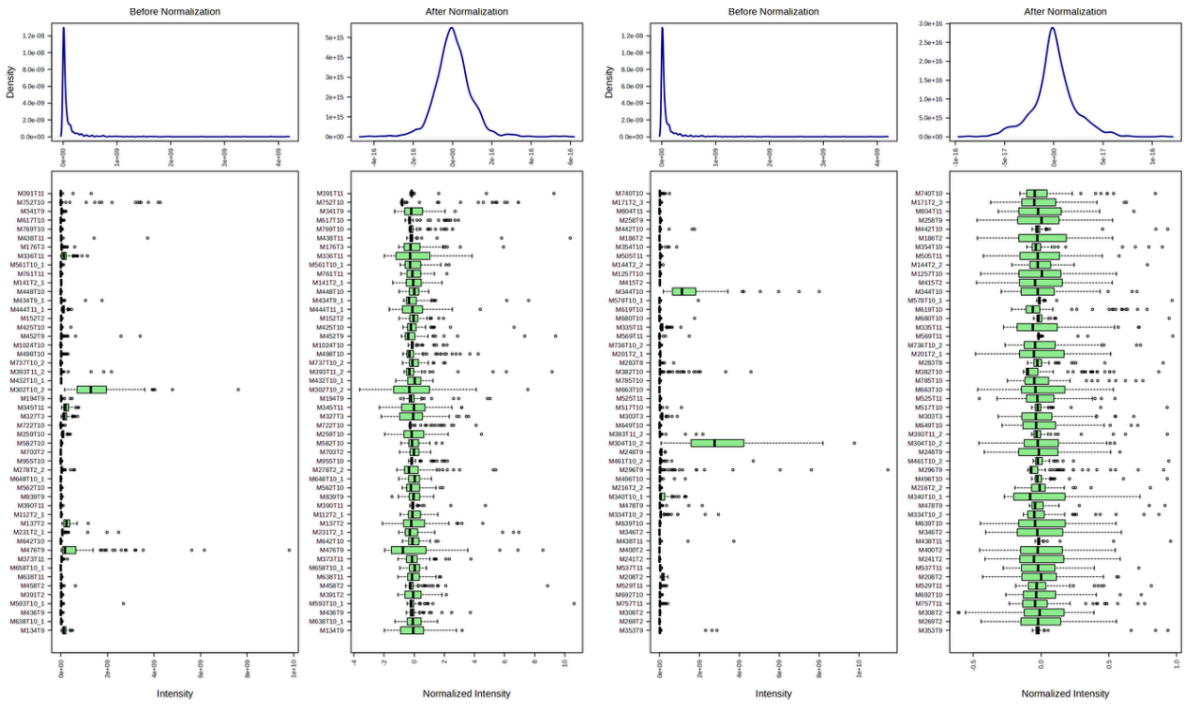
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 23 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 9 e 10, respectivamente.



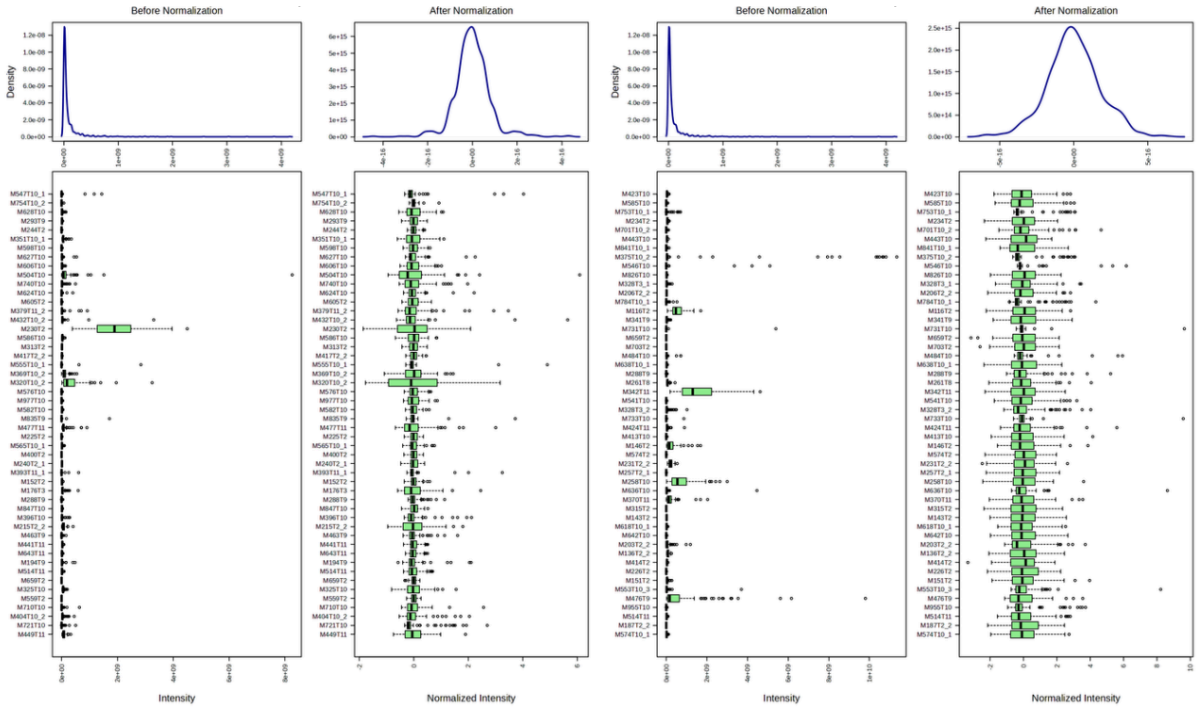
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 24 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 11 e 12, respectivamente.



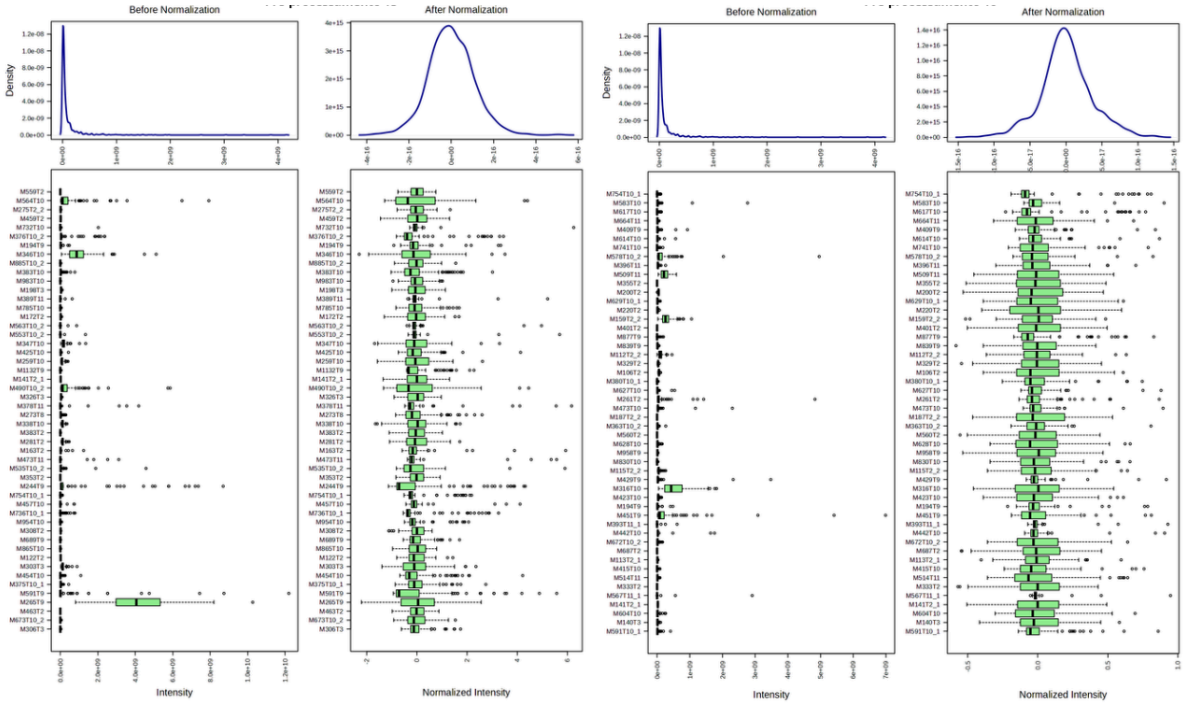
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 25 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 13 e 14, respectivamente.



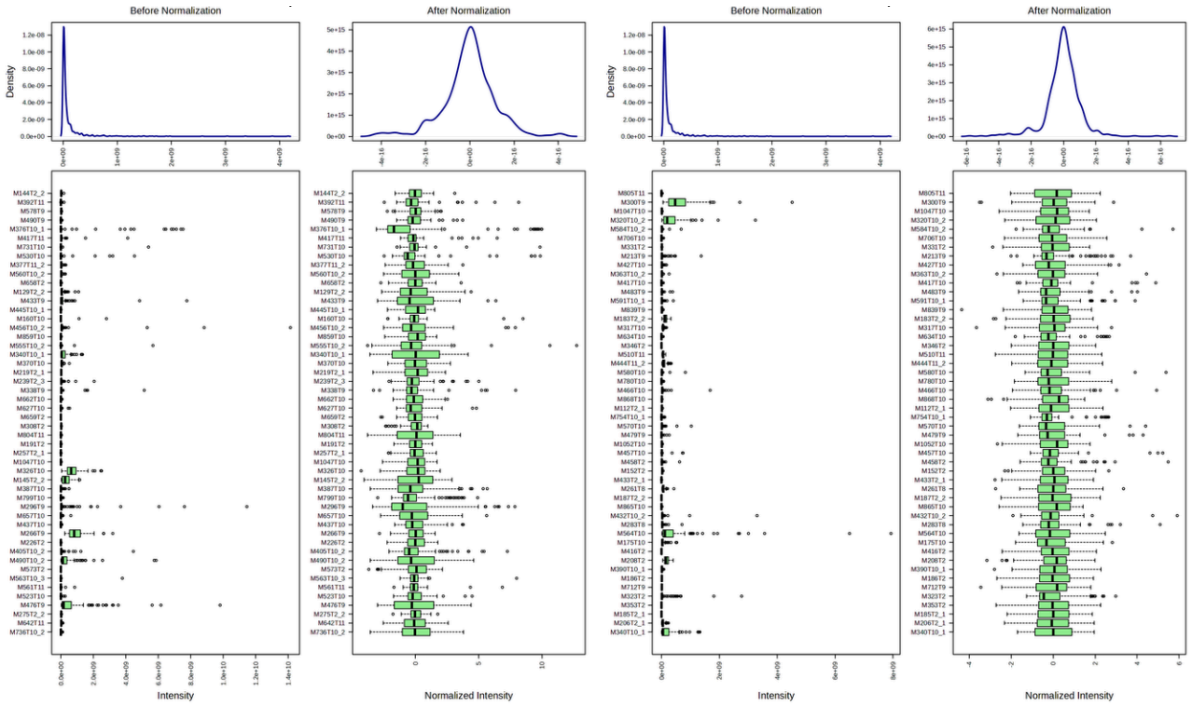
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 26 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 15 e 16, respectivamente.



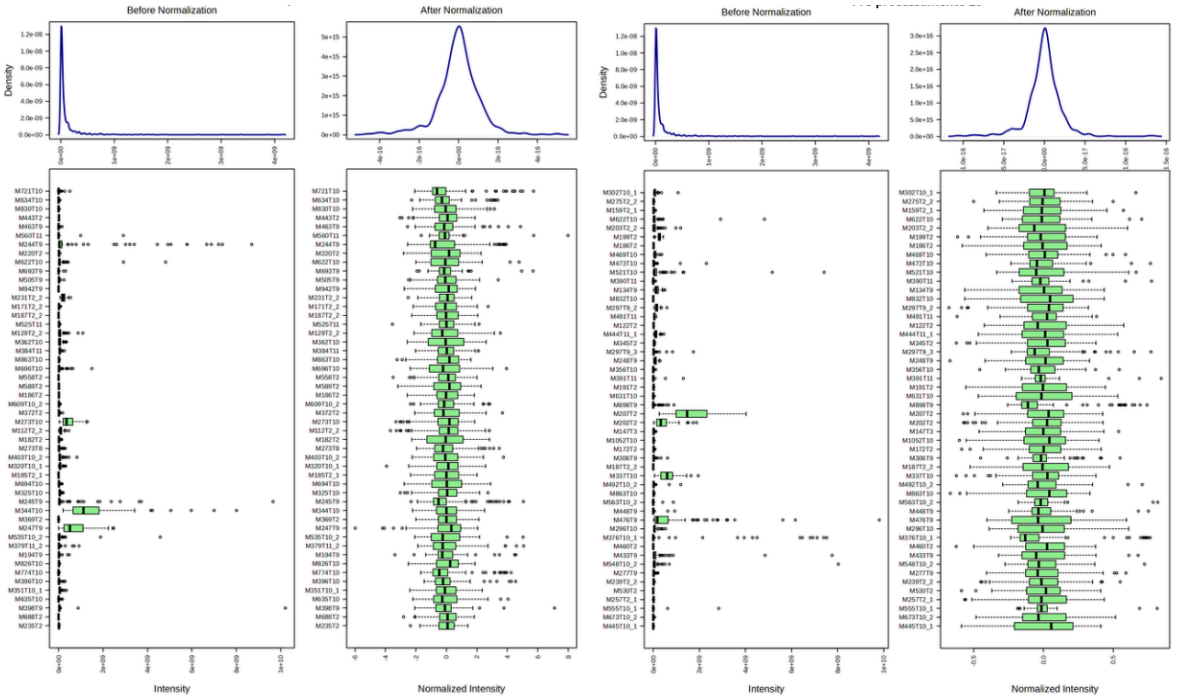
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 27 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 17 e 18, respectivamente.



Fonte: MetaboAnalyst (2026), dados processados pela autora.

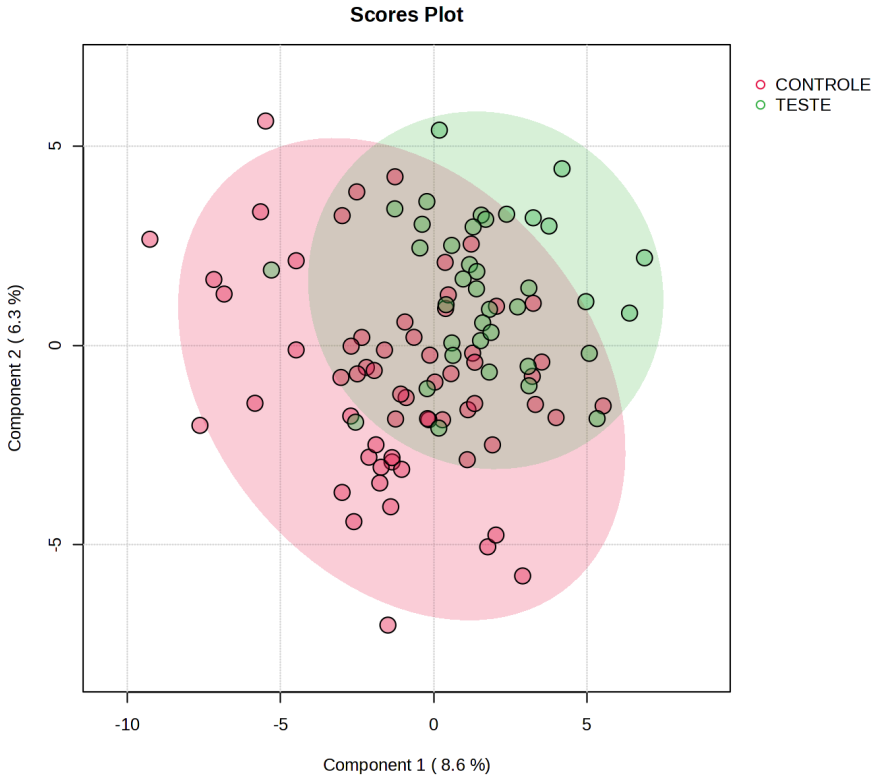
Figura 28 - Comparação da distribuição dos dados antes e após as etapas de normalização, transformação e escalonamento aplicadas nos processamentos 19 e 20, respectivamente.



Fonte: MetaboAnalyst (2026), dados processados pela autora.

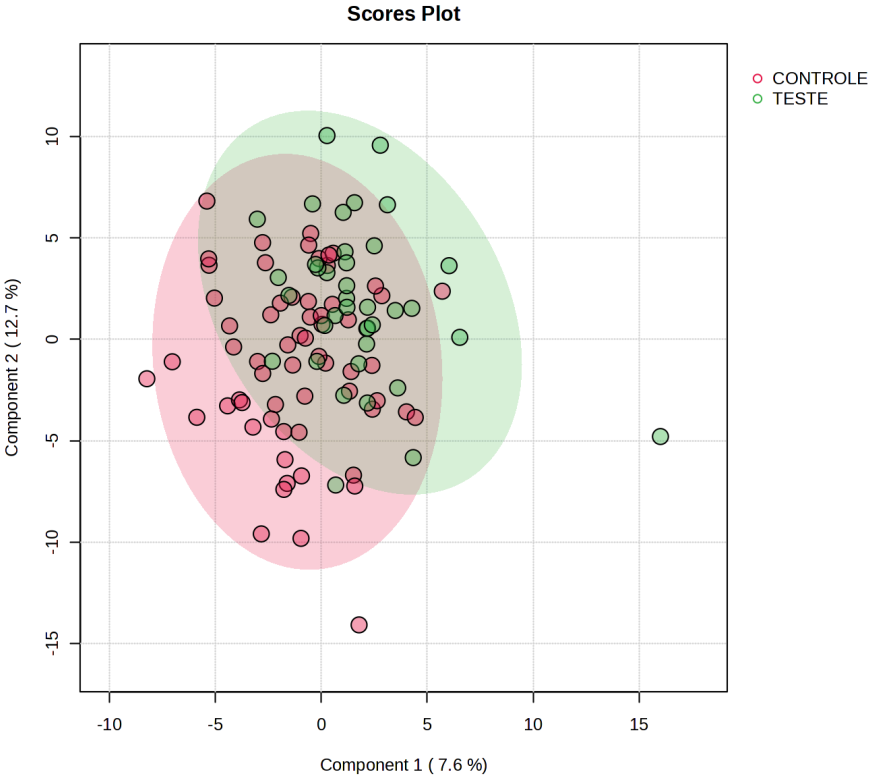
APÊNDICE B - Modelos PLS-DA

Figura 29 - PLS-DA aplicado ao conjunto N1.



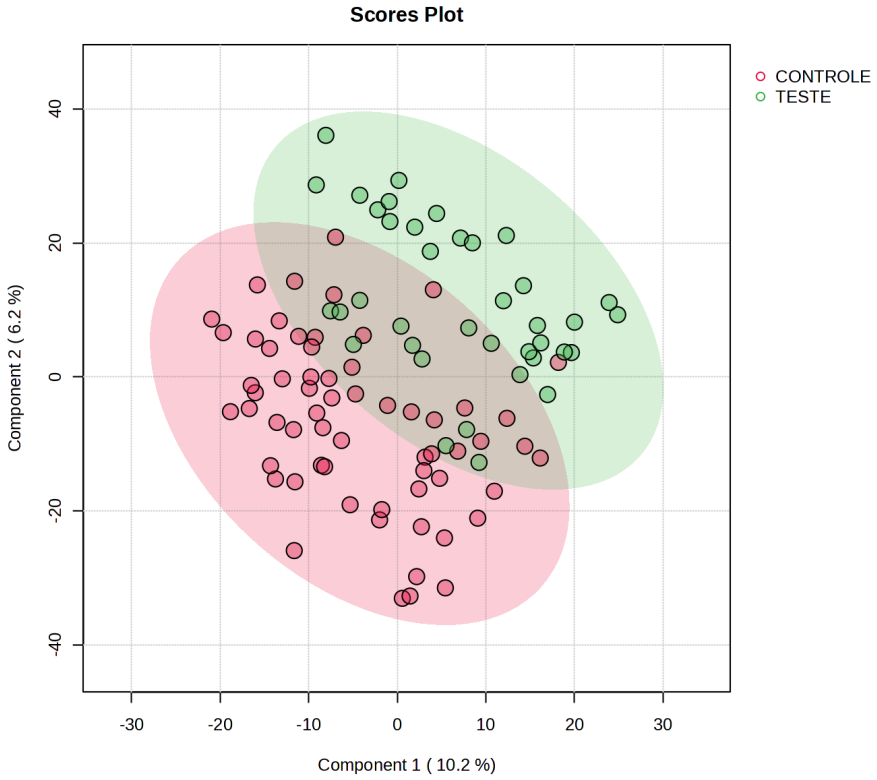
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 30 - PLS-DA aplicado ao conjunto N2.



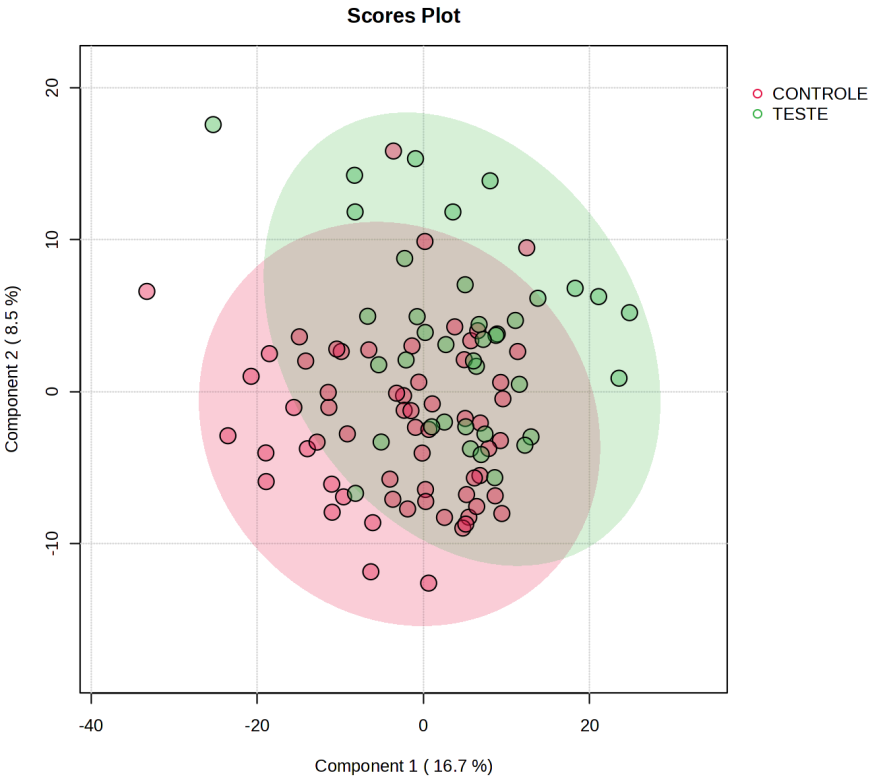
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 31 - PLS-DA aplicado ao conjunto N3.



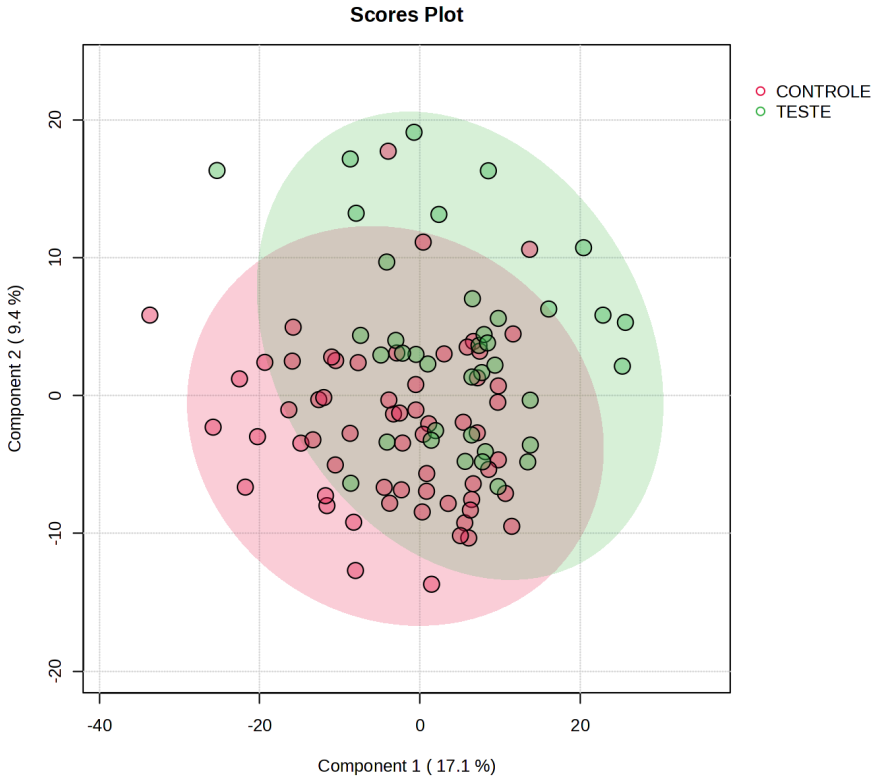
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 32 - PLS-DA aplicado ao conjunto P1.



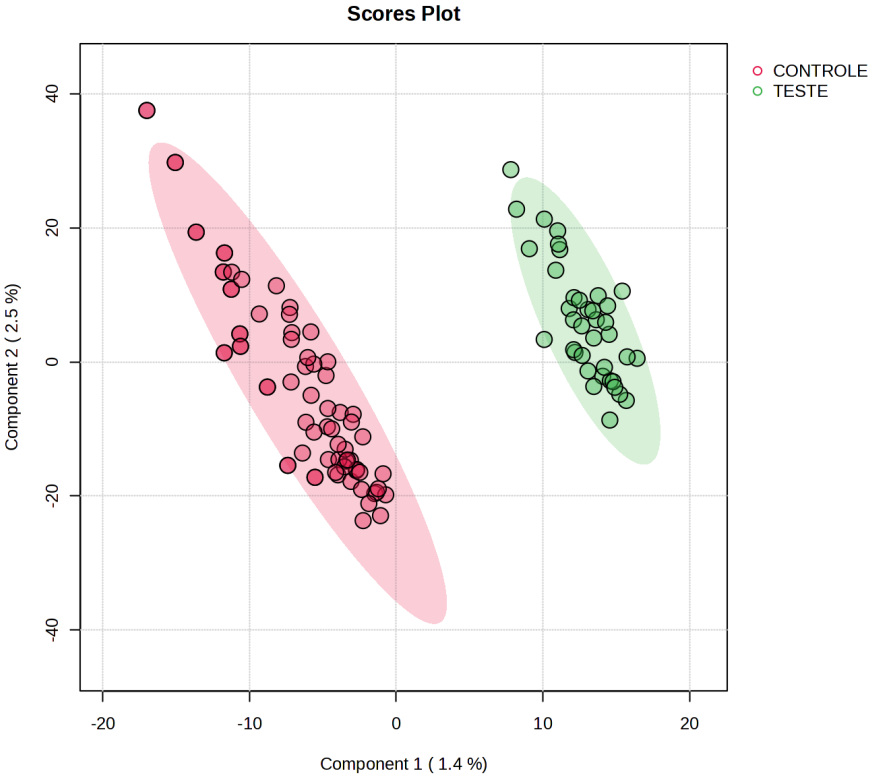
Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 33 - PLS-DA aplicado ao conjunto P2.



Fonte: MetaboAnalyst (2026), dados processados pela autora.

Figura 34 - PLS-DA aplicado ao conjunto P3.



Fonte: MetaboAnalyst (2026), dados processados pela autora.

ANEXOS

ANEXO I – Créditos de disciplinas necessárias para integralização do currículo.

Disciplina	Créditos
Ética e Segurança em Laboratórios de Pesquisa em Química (UFJF)	2
Tópicos Especiais em Química II (UFJF)	2
Química Analítica Avançada (UFJF)	4
Planejamento de Misturas (UFJF)	4
Seminários I	1
Estágio Docência I	1

ANEXO II – Carga horária dedicada a Atividade Prática Docente (Tutoria).

Disciplina de Graduação	Carga Horária
Laboratório de Química Analítica IV	60
Laboratório de Transformações Químicas	30
Laboratório de Química Inorgânica	30