

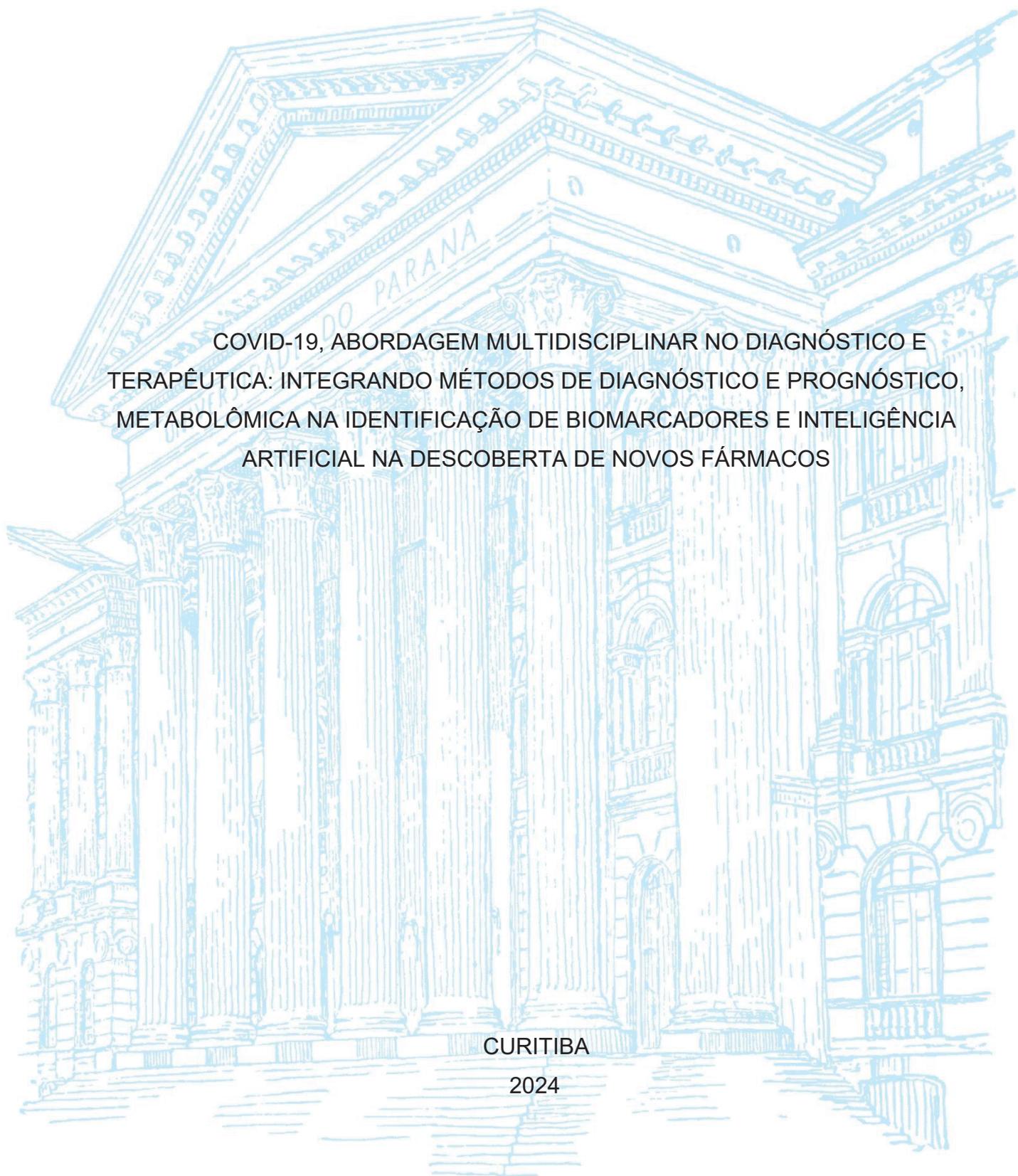
UNIVERSIDADE FEDERAL DO PARANÁ

ALEXANDRE DE FÁTIMA COBRE

COVID-19, ABORDAGEM MULTIDISCIPLINAR NO DIAGNÓSTICO E
TERAPÊUTICA: INTEGRANDO MÉTODOS DE DIAGNÓSTICO E PROGNÓSTICO,
METABOLÔMICA NA IDENTIFICAÇÃO DE BIOMARCADORES E INTELIGÊNCIA
ARTIFICIAL NA DESCOBERTA DE NOVOS FÁRMACOS

CURITIBA

2024



ALEXANDRE DE FÁTIMA COBRE

COVID-19, ABORDAGEM MULTIDISCIPLINAR NO DIAGNÓSTICO E
TERAPÊUTICA: INTEGRANDO MÉTODOS DE DIAGNÓSTICO E PROGNÓSTICO,
METABOLÔMICA NA IDENTIFICAÇÃO DE BIOMARCADORES E INTELIGÊNCIA
ARTIFICIAL NA DESCOBERTA DE NOVOS FÁRMACOS

Tese apresentada ao curso de Pós-Graduação em Ciências Farmacêuticas, Setor de Ciências da Saúde, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Doutor em Ciências Farmacêuticas.

Orientador: Prof. Dr. Roberto Pontarolo

CURITIBA

2024

Cobre, Alexandre de Fátima

COVID-19, abordagem multidisciplinar no diagnóstico e terapêutica [recurso eletrônico]: integrando métodos de diagnóstico e prognóstico, metabolômica na identificação de biomarcadores e inteligência artificial na descoberta de novos fármacos / Alexandre de Fátima Cobre – Curitiba, 2024.

1 recurso online : PDF

Tese (doutorado) – Programa de Pós-Graduação em Ciências Farmacêuticas. Setor de Ciências da Saúde, Universidade Federal do Paraná, 2024.

Orientador: Prof. Dr. Roberto Pontarolo

1. COVID-19. 2. Fatores de risco. 3. Diagnóstico. 4. Vacinação. 5. Metabolômica. 6. Biomarcadores. 7. Inteligência artificial. I. Pontarolo, Roberto. II. Universidade Federal do Paraná. III. Título.

CDD 616.24144

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação CIÊNCIAS FARMACÊUTICAS da Universidade Federal do Paraná foram convocados para realizar a arguição da tese de Doutorado de **ALEXANDRE DE FÁTIMA COBRE** intitulada: **COVID-19, ABORDAGEM MULTIDISCIPLINAR NO DIAGNÓSTICO E TERAPÊUTICA: INTEGRANDO MÉTODOS DE DIAGNÓSTICO E PROGNÓSTICO, METABOLÔMICA NA IDENTIFICAÇÃO DE BIOMARCADORES E INTELIGÊNCIA ARTIFICIAL NA DESCOBERTA DE NOVOS FÁRMACOS**, sob orientação do Prof. Dr. ROBERTO PONTAROLO, que após terem inquirido o aluno e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa. A outorga do título de doutor está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 28 de Março de 2024.

Assinatura Eletrônica
05/04/2024 14:15:31.0
ROBERTO PONTAROLO
Presidente da Banca Examinadora

Assinatura Eletrônica
05/04/2024 12:00:40.0
ANDERSON LUIZ ARA SOUZA
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ /
ESTATÍSTICA)

Assinatura Eletrônica
09/04/2024 21:45:29.0
FLAVIO DA SILVA EMERY
Avaliador Externo (UNIVERSIDADE DE SÃO PAULO - USP)

Assinatura Eletrônica
05/04/2024 12:24:58.0
YANNA DANTAS RATTMANN
Avaliador Externo (UNIVERSIDADE FEDERAL DO PARANÁ)

Às noites sem dormir, às dúvidas persistentes e às descobertas inesperadas.
Cada obstáculo foi um degrau crucial rumo ao topo.

AGRADECIMENTOS

Ao meu amado vilarejo de Muatala, na província de Nampula, Moçambique, cujas raízes são meu alicerce e inspiração. Cada passo nesta jornada é um tributo à simplicidade e humildade que moldaram quem sou. Sinto-me honrado em representar Muatala como o primeiro doutor, e meu coração se enche de orgulho ao saber que sou parte de uma comunidade que celebra coletivamente as conquistas. Que este seja um testemunho de que os sonhos podem transcender fronteiras e inspirar outros a acreditar no poder da educação. Com gratidão profunda e orgulho no coração, compartilho este momento com todos vocês.

A Dona Zilá e sua família, minha madrinha da igreja Nossa Senhora de Lourdes aqui no Brasil, merecem meu eterno amor e gratidão. Desde minha chegada em 2018, eles têm sido minha família, cuidando de mim como se eu fosse seu próprio filho. Sou imensamente grato por toda sua generosidade e carinho. Que Deus os abençoe abundantemente por todo amor e cuidado.

A minha gratidão aos meus amados familiares, especialmente à minha querida mãe, Dona Maria de Fátima Fernando, meu pai, Senhor Damião Cobre, meus queridos irmãos: Fernando de Fátima Cobre, Orlando de Fátima Cobre, Maria Alice Cobre e Albertina de Fátima Cobre, que mesmo estando a milhares de quilômetros de distância, vocês sempre estiveram ao meu lado ao longo desses seis anos. O apoio de vocês foi minha âncora, meu porto seguro, e cada conquista que alcanço é também de vocês. Obrigado por estarem sempre presentes, mesmo quando a distância parecia insuperável. Enfrentamos desafios juntos, e mesmo diante das adversidades, conseguimos vencer. Amo vocês.

Ao Prof. Dr. Dile Stremel, Prof. Dr. Alexessander Couto Alves, Msc. Moises Maia e Dra. Mónica Surek, minha gratidão pelas suas colaborações incríveis. Aos colegas da UFPR (CEB e GEATS), minha gratidão pela acolhida e colaboração.

À minha coorientadora, Profa. Dra. Fernanda Stumpf Tonin, e às minhas amigas, Profa. Dra. Mariana Millan Fachi e Dra. Beatriz Boger, não tenho palavras para expressar minha gratidão pelas suas inestimáveis contribuições ao longo desta minha jornada acadêmica. Vocês foram mais que amigas e orientadoras; foram âncoras nos momentos difíceis, me confortaram e me orientaram quando mais precisei. Suas presenças, mesmo à distância, foram minhas fontes de força e inspiração. Sou grato por suas generosidades, e pelo compromisso de me ajudarem a alcançar meus sonhos.

Por fim, agradeço ao meu orientador Prof. Dr. Roberto Pontarolo, meu anjo da guarda e um verdadeiro bom samaritano que, sem me conhecer, estendeu a mão solidária que mudou o curso da minha vida, que não apenas me orientou academicamente, mas também me forneceu moradia, sustento e apoio desde minha chegada ao Brasil. Além disso, também ele financiou minhas viagens para visitar minha família em Moçambique e para realizar o meu doutorado sanduiche na Inglaterra. Sua generosidade e bondade transcendem fronteiras, e sou profundamente grato por tudo que fez por mim. Obrigado por acreditar no potencial de um estrangeiro em busca de oportunidades e por ser o mentor exemplar que sempre esteve ao meu lado.

Minha jornada como cientista
começou na infância, quando a ciência se
tornou minha cor favorita, pintando minha
mente com fascínio e curiosidade
(Alexandre de Fátima Cobre, 2024)

NOTAS DO AUTOR

A presente tese foi redigida com algumas ressalvas de formatação:

- O documento foi estruturado em capítulos, referentes aos principais estudos publicados durante esta tese para facilitar a compreensão de todo o trabalho;
- As ilustrações e tabelas foram numeradas por capítulo com objetivo de permitir uma leitura mais dinâmica;
- Foi utilizado o sistema numérico de citação para facilitar a leitura do texto, uma vez que muitas referências são usualmente mencionadas em estudos da área de inteligência artificial aplicada a descoberta de fármacos e metabolômica;
- Os principais termos da área, sempre que possível, foram traduzidos para o português, porém, aqueles cuja tradução ainda não está bem estabelecida na literatura foram mantidos em inglês para evitar interpretações errôneas;
- As abreviaturas e siglas dos principais termos da área foram mantidas em inglês;
- As figuras foram mantidas em seu formato e idioma originais com notas e legendas explicativas adicionadas sempre que necessário.

RESUMO

A presente tese de doutorado é fruto de uma pesquisa iniciada no primeiro trimestre de 2020, impulsionada por uma bolsa de estudo emergencial, fornecida pela CAPES para o desenvolvimento de projeto sobre COVID-19. O objetivo primordial deste trabalho, foi gerar evidências científicas cruciais para o enfrentamento da pandemia em curso. Esta tese está organizada em sete capítulos desenvolvidos ao longo do período pandêmico, e objetivou produzir respostas às novas e emergentes questões desencadeadas pela COVID-19. Iniciando com os desafios de diagnóstico e identificação de fatores de risco, a pesquisa avança para explorar soluções terapêuticas e preventivas. Descobertas relevantes foram obtidas, destacando-se associações entre mortalidade e variáveis como tempo de diagnóstico, sexo e localização geográfica, sublinhando a necessidade premente de intervenções eficazes desde o início da crise sanitária. Além disso, a tese ressalta a importância crucial da nutrição na recuperação da COVID-19, identificando alimentos e nutrientes com impacto positivo no desfecho da doença, especialmente em contextos socioeconômicos subdesenvolvidos. A utilização da espectroscopia infravermelha é enfatizada como uma ferramenta válida e confiável para diagnóstico, especialmente em regiões com recursos limitados. Modelos de *machine learning* mostraram-se eficazes na previsão do diagnóstico e gravidade da COVID-19, melhorando os processos de triagem e diagnóstico em ambientes de saúde. Além disso, biomarcadores foram identificados, fornecendo *insights* valiosos para uma melhor compreensão da fisiopatologia da doença e o desenvolvimento de estratégias de tratamento mais eficazes. A pesquisa também investigou o potencial de compostos bioativos naturais como inibidores do vírus, utilizando simulações computacionais para identificar candidatos promissores para o tratamento da doença, oferecendo novas perspectivas para o desenvolvimento de terapias antivirais. No geral, esta tese representa uma contribuição significativa para a compreensão e o combate da COVID-19 em escala global, fornecendo uma análise abrangente e perspicaz sobre os desafios e avanços relacionados à pandemia.

Palavras-chave: COVID-19; fatores de risco; diagnóstico; vacinação, metabólica; biomarcadores; descoberta de novos fármacos; inteligência artificial; *machine learning*.

ABSTRACT

The present doctoral thesis is the result of research initiated in the first quarter of 2020, propelled by an emergency scholarship provided by CAPES for the development of a project on COVID-19. The primary objective of this work was to generate crucial scientific evidence for addressing the ongoing pandemic. This thesis is organized into seven chapters developed over the pandemic period, aiming to provide answers to the new and emerging questions triggered by COVID-19. Starting with the challenges of diagnosis and identification of risk factors, the research progresses to explore therapeutic and preventive solutions. Relevant findings were obtained, highlighting associations between mortality and variables such as time of diagnosis, gender, and geographical location, underscoring the urgent need for effective interventions from the onset of the health crisis. Additionally, the thesis emphasizes the crucial importance of nutrition in COVID-19 recovery, identifying foods and nutrients with a positive impact on disease outcome, especially in underdeveloped socioeconomic contexts. The use of infrared spectroscopy is emphasized as a valid and reliable tool for diagnosis, especially in regions with limited resources. Machine learning models proved effective in predicting COVID-19 diagnosis and severity, improving screening and diagnostic processes in healthcare settings. Furthermore, biomarkers were identified, providing valuable insights for a better understanding of the disease's pathophysiology and the development of more effective treatment strategies. The research also investigated the potential of natural bioactive compounds as virus inhibitors, using computational simulations to identify promising candidates for disease treatment, offering new perspectives for the development of antiviral therapies. Overall, this thesis represents a significant contribution to the understanding and combating of COVID-19 on a global scale, providing a comprehensive and insightful analysis of the challenges and advances related to the pandemic.

Keywords: COVID-19; risk factors; diagnosis; vaccination; metabolomics, biomarkers; drug discovery; artificial intelligence; machine learning.

PRODUÇÃO CIENTÍFICA RESULTANTE DA TESE

1. Cobre AF, Surek M, Vilhena RO, Böger B, Fachi MM, Momade DR, Tonin FS, Sarti FM, Pontarolo R. Influence of foods and nutrients on COVID-19 recovery: A multivariate analysis of data from 170 countries using a generalized linear model. Clin Nutr. 2022 Dec;41(12):3077-3084. doi: 10.1016/j.clnu.2021.03.018.
2. Cobre AF, Böger B, Fachi MM, Vilhena RO, Domingos EL, Tonin FS, Pontarolo R. Risk factors associated with delay in diagnosis and mortality in patients with COVID-19 in the city of Rio de Janeiro, Brazil. Cien Saude Colet. 2020 Oct;25(suppl 2):4131-4140. doi: 10.1590/1413-812320202510.2.26882020.
3. Cobre AF, Böger B, Vilhena RO, Fachi MM, Dos Santos JMMF, Tonin FS. A multivariate analysis of risk factors associated with death by COVID-19 in the USA, Italy, Spain, and Germany. Z Gesundh Wiss. 2022;30(5):1189-1195. doi: 10.1007/s10389-020-01397-7.
4. Cobre AF, Stremel DP, Böger B, Fachi MM, Borba HHL, Tonin FS, Sarti FM, Pontarolo R. The impact of COVID-19 vaccine rejection on hospital admission and variants spread worldwide: implications for healthcare policy. Research, Society and Development, v. 11, n. 11, p. e189111133435-e189111133435, 2022. doi: <https://doi.org/10.33448/rsd-v11i11.33434>.
5. Cobre AF, Alves AC, Gotine ARM, Domingues KZA, Lazo REL, Ferreira LM, Tonin FS, Pontarolo R. Novel COVID-19 biomarkers identified through multi-omics data analysis: N-acetyl-4-O-acetylneuraminic acid, N-acetyl-L-alanine, N-acetyltryptophan, palmitoylcarnitine, and glycerol 1-myristate. Intern Emerg Med. 2024 Feb 28. doi: 10.1007/s11739-024-03547-1.
6. Cobre AF, Stremel DP, Noletto GR, Fachi MM, Surek M, Wiens A, Tonin FS, Pontarolo R. Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? Comput Biol Med. 2021

Jul;134:104531. doi: 10.1016/j.compbio.2021.104531. Epub 2021 May 29. PMID: 34091385; PMCID: PMC8164361.

7. Cobre AF, Surek M, Stremel DP, Fachi MM, Lobo Borba HH, Tonin FS, Pontarolo R. Diagnosis and prognosis of COVID-19 employing analysis of patients' plasma and serum via LC-MS and machine learning. *Comput Biol Med.* 2022 Jul;146:105659. doi: 10.1016/j.compbio.2022.105659.

8. Cobre AF, Maia Neto M, de Melo EB, Fachi MM, Ferreira LM, Tonin FS, Pontarolo R. Naringenin-4' glucuronide as a new drug candidate against the COVID-19 Omicron variant: a study based on molecular docking, molecular dynamics, MM/PBSA and MM/GBSA. *J Biomol Struct Dyn.* 2023 Jul 2:1-14. doi: 10.1080/07391102.2023.2229446. Epub ahead of print. PMID: 37394802

9. Cobre AF, Böger B, Fachi MM, Ehrenfried AC, Stremel DP, Melo EB, Tonin FS, Pontarolo R. Machine learning-based virtual screening, molecular docking, drug-likeness, pharmacokinetics and toxicity analyses to identify new natural inhibitors of the glycoprotein spike (s1) of sars-cov-2. *Química Nova*, v. 46, p. 450-459, 2023. Doi: <http://dx.doi.org/10.21577/0100-4042.20230038>.

10. Domingues KZA, Cobre AF, Lazo REL, Amaral LS, Ferreira LM, Tonin FS, Pontarolo R. Systematic review and evidence gap mapping of biomarkers associated with neurological manifestations in patients with COVID-19. *J Neurol.* 2024 Jan;271(1):1-23. doi: 10.1007/s00415-023-12090-6.

11. Böger B, Fachi MM, Vilhena RO, Cobre AF, Tonin FS, Pontarolo R. Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. *Am J Infect Control.* 2021 Jan;49(1):21-29. doi: 10.1016/j.ajic.2020.07.011.

Artigos científicos publicados na tese do autor e foram selecionados pela Organização Mundial da Saúde como literatura de referência para aplicação na prática clínica visando o combate da pandemia da COVID-19

1. Cobre AF, Surek M, Vilhena RO, Böger B, Fachi MM, Momade DR, Tonin FS, Sarti FM, Pontarolo R. Influence of foods and nutrients on COVID-19 recovery: A multivariate analysis of data from 170 countries using a generalized linear model. Clin Nutr. 2022 Dec;41(12):3077-3084. doi: 10.1016/j.clnu.2021.03.018. Epub 2021 Mar 22. PMID: 33933299; PMCID: PMC7982641.

2. Cobre AF, Böger B, Fachi MM, Vilhena RO, Domingos EL, Tonin FS, Pontarolo R. Risk factors associated with delay in diagnosis and mortality in patients with COVID-19 in the city of Rio de Janeiro, Brazil. Cien Saude Colet. 2020 Oct;25(suppl 2):4131-4140. doi: 10.1590/1413-812320202510.2.26882020.

3. Cobre AF, Böger B, Vilhena RO, Fachi MM, Dos Santos JMMF, Tonin FS. A multivariate analysis of risk factors associated with death by COVID-19 in the USA, Italy, Spain, and Germany. Z Gesundh Wiss. 2022;30(5):1189-1195. doi: 10.1007/s10389-020-01397-7.

4. Cobre AF, Stremel DP, Noleto GR, Fachi MM, Surek M, Wiens A, Tonin FS, Pontarolo R. Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? Comput Biol Med. 2021 Jul;134:104531. doi: 10.1016/j.compbimed.2021.104531.

5. Cobre AF, Surek M, Stremel DP, Fachi MM, Lobo Borba HH, Tonin FS, Pontarolo R. Diagnosis and prognosis of COVID-19 employing analysis of patients' plasma and serum via LC-MS and machine learning. Comput Biol Med. 2022 Jul;146:105659. doi: 10.1016/j.compbimed.2022.105659.

6. Domingues KZA, Cobre AF, Lazo REL, Amaral LS, Ferreira LM, Tonin FS, Pontarolo R. Systematic review and evidence gap mapping of biomarkers associated with neurological manifestations in patients with COVID-19. J Neurol. 2024 Jan;271(1):1-23. doi: 10.1007/s00415-023-12090-6.

7. Böger B, Fachi MM, Vilhena RO, Cobre AF, Tonin FS, Pontarolo R. Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. *Am J Infect Control*. 2021 Jan;49(1):21-29. doi: 10.1016/j.ajic.2020.07.011.

Apresentação oral em eventos científicos internacionais no Reino Unido, no Peru e no Brasil

1. Cobre AF, Surek M, Stremel DP, Fachi MM, Alves AC, Tonin FS, Pontarolo R. Diagnosis, and prognosis of COVID-19 employing biochemical tests and machine learning. In: GLOBAL VIROLOGY CONGRESS - Virology & Advances in Clinical & Celular Immunology. London, United Kingdom, 11-12 September 2023.

2. Cobre AF, Stremel DP, Fachi MM, Tonin FS, Pontarolo R. Un enfoque integrador revela candidatos a fármacos multifacéticos para el tratamiento simultáneo de COVID-19, hepatitis B y C, dengue y HIV utilizando inteligencia artificial & machine learning multi-target y polifarmacología. In: Seminario, en conmemoración al día del químico farmacêutico del Ciudad de Cusco, Peru, 9-11 Mayo de 2023.

3. Cobre AF, Stremel DP, Fachi MM, Tonin FS, Pontarolo R. Diagnosis and prognosis of COVID-19 employing patients' metabolomics (LC-MS) and biochemical data and machine learning. In: First Workshop UFPR CAPES-PrInt: New insights on Health Sciences. Curitiba city, Paraná, Brazil, January 20, 2023.

Apresentação de pôsteres em eventos científicos internacionais

1. Cobre AF, Junkert A, Fachi MM, Böger B, Surek M, Wiens A, Tonin F, Pontarolo R. POSB422 Machine Learning-Based Virtual Screening, Molecular Docking and Drug-Likeness to Discover New Inhibitors of the Glycoprotein Spike (S1) of SARS-CoV-2. Value Health. 2022 Jan;25(1):S274. doi: 10.1016/j.jval.2021.11.1333. Epub 2022 Jan 19. PMID: PMC8769606.

2. Domingues K, Cobre A, Tonin FS, Pontarolo R. PD34 Neuron-specific Biomarkers Associated With Neurological Manifestations In COVID-19: An Evidence Mapping Systematic Review. *International Journal of Technology Assessment in Health Care*. 2022;38(S1):S102-S102. doi:10.1017/S0266462322002938

LISTA DE ILUSTRAÇÕES

REVISÃO DE LITERATURA	– Fluxo do desenvolvimento de um novo fármaco.....	50
REVISÃO DE LITERATURA	– Papel da inteligência artificial (IA) na descoberta de medicamentos.....	52
CAPÍTULO 1	– FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 E O IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO AUMENTO DE INTERNAÇÕES HOSPITALARES.....	58
FIGURA 1.1	– Curva de Kaplan-Meier pelo Índice de Desenvolvimento Social do tempo desde o início dos sintomas até o diagnóstico no Rio de Janeiro, Brasil (fevereiro-abril de 2020).....	69
CAPÍTULO 3	– ACURÁCIA DA TÉCNICA DE ESPETROFOTOMETRIA DE INFRAVERMELHO NO DIAGNOSTICO DE COVID-19: UM ESTUDO DE REVISÃO SISTEMÁTICA COM META-ANÁLISE.....	114
FIGURA 3.1	– Fluxograma da revisão sistemática.....	119
FIGURA 3.2	– Qualidade metodológica dos estudos incluídos seguindo o QUADAS-2.....	129
CAPÍTULO 4	– DESENVOLVIMENTO DE MODELOS PREDITIVOS E IDENTIFICAÇÃO DE BIOMARCADORES PROGNÓSTICOS EM COVID-19, HIV E TUBERCULOSE POR MEIO DE INTELIGÊNCIA ARTIFICIAL E MACHINE LEARNING.....	135
FIGURA 4.1	– Fluxograma do estudo I: análise de dados dos exames bioquímicos, hematológicos e de urinálise de pacientes COVID-19 atendidos no Hospital Israelita Albert Einstein (Brasil) visando a predição do diagnóstico e investigação de potenciais Biomarcadores prognósticos.....	140
FIGURA 4.2	– Fluxograma do estudo envolvendo dados dos exames bioquímicos e hematológicos de pacientes COVID-19, HIV, TB e co-infectados HIV/TB atendidos no Hospital Geral de Marrere (Moçambique) visando a predição do diagnóstico dessas doenças.....	144
FIGURA 4.3	– Fluxograma do estudo envolvendo análise de dados clínicos e demográficos dos pacientes atendidos nas diferentes farmácias privadas no Brasil visando o desenvolvimento de modelos de machine learning para predição do diagnóstico da COVID-19.....	153
FIGURA 4.4	– Análise exploratória. Modelo de análise de componentes principais (PCA) de discriminação de amostras negativas e positivas (A) e amostras de pacientes com doença grave e não grave (B).....	157

FIGURA 4.5	– Gráfico de leverage versus resíduos de student para detecção de amostras discrepantes. Para dados de diagnóstico: análise de outliers de amostras negativas (A) e amostras positivas (B). Para dados de gravidade: análise de outliers para amostras de pacientes sem gravidade (C) e com gravidade (D).....	158
FIGURA 4.6	– Curvas ROC de acurácia dos modelos de machine learning. Artificial Neural Network (ANN): diagnóstico (A) e gravidade (B). Decision tree (DT): diagnóstico (C) e gravidade (D). Partial Least Squares Discriminant Analysis (PLS-DA): diagnóstico (E) e gravidade (F). K-Nearest Neighbors (KNN): diagnóstico (G) e gravidade (H).....	159
FIGURA 4.7	– Análise exploratória do conjunto de dados COVID-19. Em (A) é mostrado o modelo PCA das amostras de sangue de 816 pacientes com COVID-19 diagnosticados por RT-PCR são representadas pelos triângulos vermelhos e as amostras de sangue de 920 controles com RT-PCR negativo são representadas por círculos verdes. Em (B) é mostrado o gráfico de Hotelling T^2 versus Q-resíduos do modelo PCA para detectar valores discrepantes em dados de amostras de pacientes com COVID-19. Neste gráfico, uma amostra é considerada outlier se e somente se apresentar simultaneamente altos valores de Hotelling T^2 e altos valores de Q-residual. De acordo com o gráfico, apesar de algumas amostras apresentarem valores elevados de Q-Resíduos, elas não podem ser consideradas outliers porque estão dentro do intervalo de confiança de 95% do Hotelling T^2 . Em (C) é mostrado o gráfico de cargas que representa as variáveis mais importantes na discriminação de amostras do grupo COVID-19 e controles. Apenas o gráfico de cargas do primeiro componente principal é mostrado por ser o que capturou a maior variância explicada dos dados originais, que foi de 74,39%.....	164
FIGURA 4.8	– Modelo PCA dos pacientes do grupo controle (n=4.520).....	166
FIGURA 4.9	– Análise exploratória do conjunto de dados de imunodeficiência humana (HIV)/AIDS, tuberculose (TB) e coinfeção HIV/TB.....	167
FIGURA 4.10	– Gráfico de leverage versus resíduos de student para detecção de amostras outliers.....	169
FIGURA 4.11	– Raiz quadrática do erro médio de validação cruzada (RMSECV) versus número da variável latente. Um total de 2 variáveis latentes (LV) foram selecionadas para a construção do modelo PLS-DA para predição do diagnóstico de COVID-19, por apresentarem menores valores de RMSECV.....	170

FIGURA 4.12	– Raiz quadrática do erro médio de validação cruzada (RMSECV) versus número de variáveis latentes. Foram selecionadas um total de 4 variáveis latentes para a construção do modelo PLS-DA para predição do diagnóstico de HIV/AIDS, TB e coinfeção HIV/TB, por apresentarem menores valores de RMSECV.....	171
FIGURA 4.13	– Área sob a curva ROC do desempenho do modelo PLS-DA para a predição de diagnóstico de pacientes COVID-19, HIV, TB, e co-infectados HIV/TB.....	172
FIGURA 4.14	– Gráfico de Importância da Variável na Projeção (VIP) dos biomarcadores mais importantes para diagnóstico de COVID-19.....	176
FIGURA 4.15	– Gráfico de importância variável na projeção (VIP) dos biomarcadores mais importantes para o diagnóstico de HIV, TB e coinfeção HIV/TB.....	177
FIGURA 4.16	– Análise exploratória dos dados. Modelo PCA de pacientes com COVID-19 testados pelo método de antígeno. O PCA foi capaz de discriminar entre pacientes positivos (triângulos vermelhos) e negativos (quadrados verdes) com COVID-19....	178
FIGURA 4.17	– Análise exploratória dos dados. Modelo PCA de pacientes com COVID-19 testados pelo método anti-IgG. O PCA foi capaz de discriminar entre pacientes positivos (triângulos vermelhos) e negativos (quadrados verdes) com COVID-19....	178
FIGURA 4.18	– Análise exploratória dos dados. Modelo PCA de pacientes com COVID-19 testados pelo método anti-IgM. O PCA foi capaz de discriminar entre pacientes positivos (triângulos vermelhos) e negativos (quadrados verdes) com COVID-19....	179
CAPÍTULO 5	– EXPLORANDO A METABOLÔMICA INTEGRATIVA POR LC-MS, GC-MS e RMN COM INTELIGÊNCIA ARTIFICIAL NA DESCOBERTA DE NOVOS BIOMARCADORES ASSOCIADOS AO DIAGNOSTICO E PROGNÓSTICO DA COVID-19.....	197
FIGURA 5.1	– Fluxograma do estudo I: diagnóstico e prognóstico de COVID-19 empregando análise de plasma e soro via LC-MS e Machine learning.....	201
FIGURA 5.2	– Fluxograma do estudo II: Análise de dados metabolômicos de pacientes com COVID-19 de vários países revela novos biomarcadores para diagnóstico precoce e prognóstico da infecção por SARS-CoV-2.....	207
FIGURA 5.3	– Análise exploratória de dados de pacientes do estudo I.....	216
FIGURA 5.4	– Área sob a curva ROC do desempenho do modelo PLS-DA para a predição do diagnóstico e severidade de COVID-19 usando dados de pacientes do estudo I.....	218

FIGURA 5.5	– Gráfico de importância variável na projeção dos biomarcadores mais importantes para diagnóstico de COVID-19 (10 principais) do estudo I.....	219
FIGURA 5.6	– Gráfico de importância variável na projeção dos biomarcadores mais importantes para gravidade e letalidade da COVID-19 do estudo I.....	220
FIGURA 5.7	– Perfil dos 10 principais biomarcadores sanguíneos associados ao diagnóstico de COVID-19. Os resultados estão agrupados de acordo com as classes: saudável (n=25), não-COVID-19 (n=25) e COVID-19 (n=46).....	221
FIGURA 5.8	– Perfil dos 10 principais biomarcadores sanguíneos associados à gravidade e letalidade da COVID-19.....	222
FIGURA 5.9	– Gráfico de Importância Variável na Projeção (VIP) dos biomarcadores mais importantes para diagnóstico de COVID-19 (servidor web Metaboanalyst 5.0).....	223
FIGURA 5.10	– Gráfico de Importância Variável na Projeção (VIP) dos biomarcadores mais importantes para gravidade/letalidade da COVID-19 (servidor web Metaboanalyst 5.0).....	224
FIGURA 5.11	– Modelo PCA para discriminação de amostras de pacientes com COVID-19 (círculos vermelhos) e voluntários saudáveis (círculos verdes) do estudo II.....	226
FIGURA 5.12	– Modelo PCA para discriminação de amostras de dados de diagnóstico e gravidade da COVID-19, do estudo II.....	227
FIGURA 5.13	– Gráfico de leverage versus resíduos de estudantes para detecção de valores discrepantes no conjunto de sete conjuntos de dados de amostras de pacientes da Espanha que foram analisadas por MNR (a), de amostras da China que foram analisadas por GC-MS (b), de amostras dos Estados Unidos que foram analisados por LC-MS (c), de amostras da Itália que foram analisadas por GC-MS (d), de amostras da França que foram analisadas por LC-MS (e), de amostras da Itália que foram analisadas por LC-MS (f) e as amostras da Itália que também foram analisadas por GC-MS (g).....	229
FIGURA 5.14	– Gráficos da Raiz Quadrática Média do Erro da Validação Cruzada (RMSECV) versus Número de Variáveis Latentes (LV).....	230
FIGURA 5.15	– Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 utilizando dados da Espanha (conjunto de dados 1, dados MNR).....	234
FIGURA 5.16	– Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 utilizando dados da China (conjunto de dados 2, GC-MS Data).....	234

FIGURA 5.17	– Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 usando conjunto de dados da Espanha (conjunto de dados 3, dados LC-MS).....	234
FIGURA 5.18	– Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 utilizando dados da Itália (conjunto de dados 4, GC-MS Data).....	235
FIGURA 5.19	– Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 utilizando dados da França (conjunto de dados 5, GC-MS Data).....	235
FIGURA 5.20	– Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 utilizando dados da Itália (conjunto de dados 7, dados LC-MS).....	236
FIGURA 5.21	– Importância variável no gráfico de projeção (VIP) dos biomarcadores mais importantes para a gravidade da COVID-19 usando dados da Itália (dados LC-MS, conjunto de dados 6).....	236
FIGURA 5.22	– Perfil dos principais biomarcadores sanguíneos associados à gravidade da COVID-19 utilizando dados da Itália (conjunto de dados LC-MS, conjunto de dados 6). Os resultados são agrupados de acordo com classes: pacientes com COVID-19 grave (n = 16), COVID-19 moderado (n =16), COVID-19 leve (n = 20) e voluntários saudáveis (n=9). As caixas indicam intervalos interquartis (mediana); linhas horizontais indicam valores mínimos e máximos.....	238
FIGURA 5.23	– Perfil dos principais biomarcadores sanguíneos associados ao diagnóstico de COVID-19 utilizando dados de Espanha (conjunto de dados MNR). Os resultados estão agrupados de acordo com as classes: saudável (n = 280) e COVID-19 (n = 261). As caixas indicam intervalos interquartis (mediana); linhas horizontais indicam valores mínimos e máximos.....	239
FIGURA 5.24	– Perfil dos principais biomarcadores sanguíneos associados ao diagnóstico de COVID-19 utilizando dados da China (conjunto de dados GC-MS). Os resultados estão agrupados de acordo com as classes: saudável (n = 57), não-COVID-19 (n = 30) e COVID-19 (n = 60). As caixas indicam intervalos interquartis (mediana); linhas horizontais indicam valores mínimos e máximos.....	239
FIGURA 5.25	– Perfil dos principais biomarcadores sanguíneos associados ao diagnóstico de COVID-19 utilizando dados dos da Itália (conjunto de dados LC-MS). Os resultados estão agrupados	

	de acordo com as classes: saudável (n = 133) e COVID-19 (n = 254). As caixas indicam intervalos interquartis (mediana); linhas horizontais indicam valores mínimos e máximos.....	240
CAPÍTULO 6	– ABORDAGEM INTEGRATIVA NA BUSCA POR INIBIDORES DA PROTEÍNA SPIKE (RBD) DO SARS-COV-2: TRIAGEM VIRTUAL, ANÁLISE DE DRUG-LIKENESS, PREDIÇÕES ADMET, DINÂMICA MOLECULAR E INTELIGÊNCIA ARTIFICIAL PARA O TRATAMENTO DA COVID-19.....	255
FIGURA 6.1	– Fluxograma do estudo.....	259
FIGURA 6.2	– Identificação da proteína Spike do tipo selvagem (PDB ID 6M17) e da mutação ômicron (PDB ID 7T9L) do vírus SARS-CoV-2.....	261
FIGURA 6.3	– Modelagem de docking molecular dos quatro principais ligantes. O acoplamento dos ligantes naringenina-4'-O-Glucuronídeo (ZINC000045789238), ergolóide (ZINC000003995616), ohioensina A (ZINC000004098448) e prunetrina (ZINC000008662732) são mostrados nas Figuras A, C, E e G, respectivamente.....	274
FIGURA 6.4	– Modelagem de docking molecular do remdesivir (medicamento de referência). As interações do complexo proteína-ligante Spike são mostradas em linhas tracejadas. Estruturas amarelas e azuis representam o remdesivir e a proteína spike (S1), respectivamente.....	275
FIGURA 6.5	– Simulações de dinâmica molecular dos quatro principais ligantes. As Figuras A, B, C e D mostram o perfil de RMSD, RMSF, ligação de hidrogênio total e acessibilidade ao solvente para as simulações de dinâmica molecular no tempo de 100 ns.....	280
FIGURA 6.6	– Simulações de dinâmica molecular do remdesivir (medicamento de referência). As Figuras A, B, C e D mostram o perfil de RMSD, RMSF, ligação de hidrogênio total e acessibilidade ao solvente para as simulações de dinâmica molecular no tempo de 100 ns.....	281
FIGURA 6.7	– Resultados da análise de docking de poses de compostos com maior afinidade com a espícula (S1) do SARS-CoV-2. São mostrados apenas os docking dos ligantes que apresentaram semelhança com o fármaco, e que também apresentaram melhores resultados na análise ADMET (NBC5, NBC14, NBC15 e NBC27).....	284
FIGURA 6.8	– Estruturas 2D de interações proteína-ligante. São mostradas apenas as estruturas dos ligantes mais promissoras contra a proteína Spike do SARS-CoV-2 (NBC5, NBC14, NBC15 e NBC27). As interações dos ligantes NBC5, NBC14, NBC15 e	

	NBC27 são mostradas nas Figuras A, B, C e D, respectivamente.....	298
FIGURA 6.9	– Análise do espaço químico usando os descritores de Lipinski. Em (A) são mostrados a classe dos compostos bioativos. Houve uma diferença significativa das classes de bioatividade segundo o LogP (B), pIC50 (C), peso molecular (D) grupos doadores (E) e aceptores de ligações de hidrogênio (F).....	301
FIGURA 6.10	– Modelo PCA. Discriminação entre os compostos ativos (IC50<100 nM), inativos (IC50 entre 100-10000 nM) e com atividade intermediária (IC50 entre 100-10000 nM) contra o vírus SARS-CoV-2.....	302
FIGURA 6.11	– Em (A), (C) e (E) os valores da bioatividade experimental e predita dos modelos de Random forest (RF), Xgboost e Hisogradient Boosting (HGB), respectivamente. Em (B), (D) e (E) são mostradas as variáveis (grupos funcionais) mais importantes dos compostos na predição bioatividade contra o -SARS-CoV-2 nos modelos RF, Xgboost e HGB, respectivamente.....	304
CAPÍTULO 7	– ABORDAGEM INTEGRATIVA REVELA CANDIDATOS MULTIFACETADOS A FÁRMACOS PARA O TRATAMENTO SIMULTÂNEO DE COVID-19, HEPATITE, DENGUE E HIV USANDO INTELIGÊNCIA ARTIFICIAL MULTI-TARGET, E POLIFARMACOLOGIA.....	313
FIGURA 7.1	– Fluxograma usado para coleta e limpeza dos dados, treinamento e validação dos modelos de machine learning baseados em QSAR para predição de candidatos a fármacos com bioatividade multi-target (SARS-CoV-2, HBV, HCV, Dengue e HIV).....	318
FIGURA 7.2	– Análise exploratória dos compostos bioativos usados para o desenvolvimento de modelos de machine learning para predição de bioatividade de compostos anti SARS-CoV-2, HIV, HBV, HCV e dengue.....	326
FIGURA 7.3	– Gráfico dos valores experimentais e preditos da bioatividade antiviral multi-target (HIV, SARS-CoV-2, dengue, e hepatite B e C) baseados na análise de relação estrutura atividade quantitativa (QSAR) dos 19 mil compostos químicos usando os top cinco melhores modelos de machine learning. Em (A), (B), (C), (D), e (E) é referente aos modelos Extra Trees regression, random Forest (RF), Extreme Gradient Boosting (XGBoost), Histogram-based Gradient Boosting, respectivamente.....	328
FIGURA 7.4	– Estrutura química de Nummularine B (A), Telaprevir(B) e Entecavir (C), com seus respectivos valores médios de	

	bioatividade multi-target anti SARS-CoV-2, HBV, HCV, Dengue e HIV.....	330
FIGURA 7.5	– Gráficos de energia livre de ligação entre nummularine B, telaprevir e Entecavir com as proteina salvos dos virus SARS-CoV-2, HBV, HCV, dengue e HIV.....	331
FIGURA 7.6	– Gráfico de energia livre de ligação MM/GBSA dos quatro ligantes (nummularine B, telaprevir e Entecavir) com a proteina MPro do SARS-CoV-2 (Figura A), complex isomerase do HBV (Figura B), RNA polimerase de HCV (Figura C), a serina protease de NS3 da dengue (Figura D), e a transcriptase reversa de HIV (Figura E).....	333
FIGURA 7.7	– Variação das energias livres de ligação MM/GBSA de cada aminoacido da proteina MPro do SARS-CoV-2 com os ligantes nummularine B (A), telaprevir (B), Entecavir (C) e nirmaltegrevir (D).....	336
FIGURA 7.8	– Variação das energias livres de ligação MM/GBSA de cada aminoacido da enzima complex isomerase do virus da hepatite B com os ligantes nummularine B (A), telaprevir (B), Entecavir (C) e nirmaltegrevir (D).....	337
FIGURA 7.9	– Variação das energias livres de ligação MM/GBSA de cada aminoacido da enzima RNA polimerase do virus da hepatite C com os ligantes nummularine B (figura 8A), telaprevir (B), Entecavir (C) e nirmaltegrevir (D).....	338
FIGURA 7.10	– Variação das energias livres de ligação MM/GBSA de cada aminoacido da enzima serina protease de NS3 do virus da dengue com os ligantes nummularine B (figura 8A), telaprevir (figura 9B), Entecavir (C) e nirmaltegrevir (D).....	339
FIGURA 7.11	– Variação das energias livres de ligação MM/GBSA de cada aminoacido da enzima transcriptase reversa do virus HIV-1 com os ligantes nummularine B (figura 8A), telaprevir (B), entecavir (C) e nirmaltegrevir (D).....	340

LISTA DE TABELAS

CAPÍTULO 1	– FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 E O IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO AUMENTO DE INTERNAÇÕES HOSPITALARES.....	59
TABELA 1.1	– Características dos pacientes com COVID-19 no Rio de Janeiro, Brasil (fevereiro-abril de 2020).....	67
TABELA 1.2	– Resultados da análise multivariada e univariada da regressão de Cox com covariáveis dependentes do tempo dos fatores de risco associados ao atraso no diagnóstico da COVID-19 no Rio de Janeiro, Brasil (fevereiro-abril de 2020).....	70
TABELA 1.3	– Evolução da COVID-19 de acordo com as características dos pacientes no Rio de Janeiro, Brasil (fevereiro-abril de 2020)....	71
TABELA 1.4	– Resultados das análises multivariada e univariada da regressão logística dos fatores de probabilidade associados à mortalidade por COVID-19 no Rio de Janeiro, Brasil (fevereiro-abril de 2020).....	73
TABELA 1.5	– Resultados da análise de sensibilidade do modelo de regressão logística multivariada e univariada dos fatores de probabilidade associados à mortalidade por COVID-19 no Rio de Janeiro, Brasil (fevereiro a abril de 2020).....	74
TABELA 1.6	– Previsão do número médio de leitos, dispositivos de ventilação e internações hospitalares por dia para tratamento de COVID-19 nos EUA, Itália, Espanha e Alemanha (01/03/2020 - 08/04/2020).....	76
TABELA 1.7	– Teste qui-quadrado para análise de coeficiente de significância, testes Nalgelkerk R^2 e Hosmer-Lemeshow para análise de ajuste do modelo de regressão logística multivariada de todas as variáveis estudadas.....	76
TABELA 1.8	– Análise multivariada dos fatores de risco de mortalidade por COVID-19 nos EUA, Itália, Espanha e Alemanha.....	78
TABELA 1.9	– Percentagem média da população que se recusa a ser vacinada contra a COVID-19 em alguns países, março-dezembro de 2021.....	81
TABELA 1.10	– Modelo de regressão de Poisson do efeito da taxa de rejeição do país no aumento do número de internações em UTI por COVID-19 e no aumento do número de casos de variantes da COVID-19 do SARS-CoV-2.....	82
CAPÍTULO 2	– AVALIAÇÃO DO EFEITO DOS ALIMENTOS E NUTRIENTES COMO ABORDAGENS COMPLEMENTARES NA RECUPERAÇÃO DE PACIENTES COM COVID-19 EM 170 PAÍSES.....	91

TABELA 2.1	– Análise multivariada de modelo linear generalizado para estimar o efeito da quantidade de proteína consumida (modelo 1), da quantidade de lipídios consumidos (modelo 2), da quantidade de carboidratos consumidos (modelo 3) e da quantidade de alimento consumidos em geral (modelo 4) na recuperação da COVID-19 em 170 países.....	101
TABELA 2.2	– Coeficientes do modelo linear generalizado para a magnitude dos efeitos dos alimentos consumidos na recuperação da COVID-19, segundo grupos do índice global de fome.....	104
CAPÍTULO 3	– ACURÁCIA DA TÉCNICA DE ESPETROFOTOMETRIA DE INFRAVERMELHO NO DIAGNOSTICO DE COVID-19: UM ESTUDO DE REVISÃO SISTEMÁTICA COM META-ANÁLISE.....	114
TABELA 3.1	– Característica basais dos estudos e pacientes incluídos.....	121
TABELA 3.2	– Parâmetros analíticos reportados dos métodos por MIR-FTIR dos estudos incluídos na revisão sistemática.....	124
TABELA 3.3	– Meta-análise dos parâmetros de acurácia para as diferentes técnicas diagnósticas.....	128
CAPÍTULO 4	– DESENVOLVIMENTO DE MODELOS PREDITIVOS E IDENTIFICAÇÃO DE BIOMARCADORES PROGNÓSTICOS EM COVID-19, HIV E TUBERCULOSE POR MEIO DE INTELIGÊNCIA ARTIFICIAL E MACHINE LEARNING.....	135
TABELA 4.1	– Exames Bioquímicos, hematológicos, de urina, virológicos e bacteriológicos suados para o desenvolvimento de modelos de <i>Machine learning</i> para a predição de diagnóstico e prognóstico de COVID-19.....	140
TABELA 4.2	– Subconjunto de dados de treinamento e de teste utilizados para o desenvolvimento de modelos de <i>machine learning</i> para prever o diagnóstico e severidade de COVID-19.....	142
TABELA 4.3	– Exames bioquímicos e hematológicos, de urina, virológicos e bacteriológicos suados para o desenvolvimento de modelos de <i>Machine learning</i> para a predição de diagnóstico de COVID-19, HIV/AIDS, TB e coinfeção HIV/TB.....	146
TABELA 4.4	– Subconjunto de dados de calibração e validação usados para o desenvolvimento de modelos de <i>machine learning</i> para prever o diagnóstico de pacientes com COVID-19, HIV/AIDS, TB e HIV/TB.....	149
TABELA 4.5	– Níveis de variação dos biomarcadores bioquímicos, hematológicos e urinários de pacientes positivos e com a doença grave em uma escala normalizada de pacientes.....	155
TABELA 4.6	– Comparação de desempenho dos modelos de <i>machine learning</i> para COVID-19.....	159

TABELA 4.7	– Variáveis importantes dos modelos de <i>machine learning</i> na classificação de positividade e severidade de COVID-19.....	160
TABELA 4.8	– Biomarcadores bioquímicos e hematológicos utilizados no estudo.....	162
TABELA 4.9	– Comparação de desempenho dos modelos de <i>machine learning</i> para COVID-19, HIV/AIDS, Tuberculose e coinfeção HIV/TB.....	173
TABELA 4.10	– Avaliação do desempenho preditivo do modelo PLS-DA na predição de amostras externas de pacientes sem nenhuma das doenças estudadas (COVID-19, HIV e TB), mas com outras comorbidades.....	177
TABELA 4.11	– Desempenho de modelos de Machine Learning para predição do diagnóstico de COVID-19.....	180
CAPÍTULO 5	– EXPLORANDO A METABOLÔMICA INTEGRATIVA POR LC-MS, GC-MS e RMN COM INTELIGÊNCIA ARTIFICIAL NA DESCOBERTA DE NOVOS BIOMARCADORES ASSOCIADOS AO DIAGNOSTICO E PROGNÓSTICO DA COVID-19.....	197
TABELA 5.1	– Conjuntos de dados multi-ômicos de pacientes COVID-19 e seus respectivos países.....	208
TABELA 5.2	– Desempenho preditivo dos modelos de <i>machine learning</i>	216
TABELA 5.3	– Biomarcadores para predição do diagnóstico e gravidade/severidade da COVID-19 de acordo com os modelos PLS-DA do software SOLO vs. e Metaboanalyst 5.0..	225
TABELA 5.4	– Resultados de desempenho dos modelos PLS-DA na previsão do diagnóstico e gravidade da COVID-19 em cada um dos sete conjuntos de dados avaliados.....	232
TABELA 5.5	– Biomarcadores importantes na previsão do diagnóstico e gravidade da COVID-19 foram comuns em duas ou mais bases de dados diferentes investigadas no estudo.....	241
CAPÍTULO 6	– ABORDAGEM INTEGRATIVA NA BUSCA POR INIBIDORES DA PROTEÍNA SPIKE (RBD) DO SARS-COV-2: TRIAGEM VIRTUAL, ANÁLISE DE DRUG-LIKENESS, PREDIÇÕES ADMET, DINÂMICA MOLECULAR E INTELIGÊNCIA ARTIFICIAL PARA O TRATAMENTO DA COVID-19.....	255
TABELA 6.1	– Triagem virtual e docking molecular dos quatro principais compostos naturais que tiveram maior afinidade pela proteína <i>Spike (RBD)</i> do SARS-CoV-2 da mutação ômicron.....	273
TABELA 6.2	– Comparação de interações do complexo ligante da proteína <i>Spike (RBD)</i> da variante ômicron do SARS-CoV-2 a partir de triagem virtual e análises de docking molecular.....	276

TABELA 6.3	– Predição dos parâmetros farmacocinéticos dos quatro principais ligantes que tiveram a maior afinidade pela proteína <i>Spike</i> (S1) RBD da mutação ômicron.....	277
TABELA 6.4	– Predição de toxicidade aguda e crônica dos quatro principais ligantes que tiveram a maior afinidade pela proteína <i>Spike</i> (S1) RBD da mutação ômicron.....	278
TABELA 6.5	– Energia livre de ligação de <i>MM/PBSA</i> e <i>MM/GBSA</i> do complexo entre a proteína <i>Spike</i> (RBD) da variante ômicron e os quatro ligantes e remdesivir (fármaco controle).....	283
TABELA 6.6	– Comparação dos resultados de docking molecular dos ligantes com maior afinidade pela proteína <i>Spike</i> (RBD) do SARS-CoV-2 tipo selvagem com os fármacos controle (molnupinavir e remdesivir).....	286
TABELA 6.7	– Análise de similaridade medicamentosa (druglikeness) dos principais compostos naturais com maior afinidade com a glicoproteína <i>Spike</i> (RBD) do SARS-CoV-2 tipo selvagem	289
TABELA 6.8	– Predição de parâmetros farmacocinéticos dos compostos naturais que simultaneamente apresentaram melhores afinidades de ligação com S1 e possuem características de semelhança com medicamentos.....	293
TABELA 6.9	– Predição de toxicidade aguda e toxicidade crônica dos seis compostos que tiveram os melhores resultados farmacocinéticos.....	296
TABELA 6.10	– Análise descritiva das variáveis da regra dos 5 de Lipinski de todos 10,057 compostos incluídos no estudo.....	300
TABELA 6.11	– Desempenho preditivo dos modelos de machine learning após avaliação da performance preditiva dos modelos top três de machine learning, após a otimização dos seus hiperparâmetros.....	303
TABELA 6.12	– Valores de bioatividade (pIC50) dos cinco compostos mais promissores identificados por dinâmica molecular e dos fármacos controles (remdesivir e molnupinavir) preditos pelos algoritmos Random Forest, XGBoost e Histogram gradient Boosting usando a abordagem QSAR-3D.....	305
CAPÍTULO 7	– ABORDAGEM INTEGRATIVA REVELA CANDIDATOS MULTIFACETADOS A FÁRMACOS PARA O TRATAMENTO SIMULTÂNEO DE COVID-19, HEPATITE, DENGUE E HIV USANDO INTELIGÊNCIA ARTIFICIAL MULTI-TARGET, E POLIFARMACOLOGIA.....	313
TABELA 7.1	– Descrição das características e funções biológicas das proteínas alvos dos vírus SARS-CoV-2, dengue, hepatite B, hepatite C e HIV.....	322

TABELA 7.2	– Top 5 melhores modelos de machine learning multi-target para predição dos compostos com bioatividade simultânea contra o SARS-CoV-2, HIV, dengue e hepatite B e C.....	327
TABELA 7.3	– Predição da bioatividade <i>multi-target</i> (HIV, SARS-CoV-2, dengue, hepatite B e C) para os 113.682 compostos do <i>Human Metabolome Database</i> , utilizando modelos de machine learning como <i>Extra Trees regression</i> , <i>Random Forest (RF)</i> , <i>Extreme Gradient Boosting (XGBoost)</i> e <i>Histogram-based Gradient Boosting</i> . Apresentamos apenas os 10 principais compostos que exibiram valores elevados de bioatividade média (média de pIC50). Para uma compreensão mais clara dos resultados, os valores médios de pIC50 foram convertidos para IC50 (nM).....	330

LISTA DE ABREVIATURAS OU SIGLAS

ACE-2	– Angiotensin-converting enzyme 2
ADA	– Adaptive boosting
AhR	– Aryl hydrocarbon Receptor
AR-LBD	– Androgen receptor ligand binding domain
AIDS	– Acquired immunodeficiency syndrome
ADMET	– Absorção, distribuição, metabolismo, excreção e Toxicidade
AIC	– Akaike information criterion
ATAD5	– Atpase family AAA domain-containing protein 5
ANN-DA	– Artificial neural networks discriminant analysis
AST	– Enzima aspartato aminotransferase
ATR	– Attenuated total reflection
AUC	– Área sob a Curva
CADD	– <i>Computer-aided drug design</i>
CCM	– Correlação de Matthew
CoM	– Center of Mass
CtO ₂	– Central venous oxygen saturation
CPK	– Creatine phosphokinase
CYP	– Cytochrome P450 enzymes
DL	– <i>Deep learning</i>
DM	– Dinâmica molecular
DNNs	– Deep neural networks
ELISA	– Enzyme-linked immunosorbent assay
ER	– Estrogen receptor alpha
ER-LBD	– Estrogen receptor ligand binding domain
ESPEN	– European Society for Clinical Nutrition and Metabolism
ET	– Extra Trees regression/classification
EPO	– External parameter orthogonalization
FP	– Falso positivo
FN	– Falso negativo

FTIR	– Fourier transform infrared spectroscopy
FiO ₂	– Fraction of Inspired Oxygen
GSK	– Glaxosmithkline
GLM	– Generalized Linear Model
GLSW	– Generalized least squares weighting
GLIM	– Iniciativa de Liderança Global sobre Desnutrição
GPxs	– Glutathione peroxidases
HBV	– Hepatitis b virus
HCV	– Hepatitis c virus
HIV	– Virus de imunodeficiência Humana
HR	– Hazard ratio
HSE	– Heat shock factor response element
IA	– Inteligência artificial
IC	– Intervalo de Confiança
IC50	– Half-maximal inhibitory concentration
IgG	– Immunoglobulin G
IgM	– Immunoglobulin M
INR	– International normalized ratio
KNN	– K-nearest neighbours
LC-MS	– Liquid chromatography coupled with mass spectrometry
LREG	– Logistic regression discriminant analysis
LV	– Latent variable
GC-MS	– Cromatografia gasosa acoplada à espectrometria de massa
SIMCA	– Soft independent modelling of class analogy
LDH	– Lactato desidrogenase
LDH	– Linear discriminant analysis
MAE	– Mean absolute error
MAPE	– Mean absolute percentage error
MCH	– Mean corpuscular hemoglobin
MCHC	– Mean corpuscular hemoglobin concentration
MMP	– Mitochondrial membrane potential
ML	– Machine learning

MM/GBSA	– Mechanics/generalized born surface area
MM/PBSA	– Molecular mechanics/poisson-boltzmann surface area
MPro	– Protease principal do vírus SARS-cov-2
MCV	– Mean corpuscular volume
MSC	– Multiplicative scatter correction
NB	– Naive bayes
NLR	– Razão de verossimilhança negativa
OMS	– Organização Mundial da Saúde
OSC	– Orthogonal signal correction
pIC50	– Negative logarithm (base 10) of the IC50 value
PLS-DA	– Discriminant analysis by partial least squares
PCA	– Principal component analysis
PDB	– Protein data bank
PDBQT	– Protein data bank, partial charge
PLS	– Partial least squares
PCR	– Proteína C reativa
pCO2	– Pressão arterial ou venosa de dióxido de carbono
PPAR-Gamma	– Peroxisome proliferator activated receptor gamma
QDA	– Quadratic discriminant analysis
QSAR	– Relação quantitativa estrutura-atividade
QSPR	– Relação quantitativa estrutura-propriedade
QUADAS-2	– Quality Assessment of Diagnostic Accuracy Studies-2
R ²	– Coefficient of Determination
RBD	– Receptor-binding domain
RDW	– Red blood cell distribution width
RF	– Random forest
RT-PCR	– Real-Time reverse Transcription Polymerase Chain Reaction
ROC	– Receiver operating characteristic
RMN	– Ressonância magnética nuclear
RMSEC	– Root-Mean-Square Error of Cross-Validation
RMSEC	– Root-Mean-Square Error of Calibration

RMSD	– Root mean square deviation
RMSF	– Root mean square fluctuation
RNA	– Ácido Ribo nucléico
SVM	– Support vector machine
SARS-CoV-2	– Severe acute respiratory syndrome coronavirus 2
SBDD	– <i>Structure based drug design</i>
SNV	– Standard normal variate
TB	– Tuberculose pulmonar
TrxRs	– Tiorredoxina redutase
PLR	– Razão de verossimilhança positiva (PLR)
PRISMA	– <i>Preferred Reporting Items for Systematic Reviews and Meta-Analysis</i>
PROSPERO	– <i>International Prospective Register of Systematic Reviews</i>
UTI	– <i>Unidade de Tratamento Intensivo</i>
VS	– <i>Virtual screening</i>
VP	– Verdadeiro positivo
VN	– Verdadeiro negativo
VIP	– Variable Importance in projection
XGBoostDA	– Gradient boosted tree discriminant analysis

SUMÁRIO

1 INTRODUÇÃO	32
1.1 OBJETIVOS	33
1.1.1 Objetivo geral	33
1.1.2 Objetivos específicos	33
1.2 JUSTIFICATIVA	34
2 REVISÃO DE LITERATURA	35
2.1 BREVE HISTÓRICO DO SURGIMENTO DA COVID-19	35
2.2 DADOS EPIDEMIOLÓGICOS DA COVID-19 NO BRASIL E NO MUNDO	36
2.3 FATORES DE RISCO PARA COVID-19	37
2.4 COVID LONGA: O NOVO DESAFIO DA PANDEMIA	38
2.5 DIAGNÓSTICO E PROGNÓSTICO DA COVID-19	39
2.5.1 Biomarcadores laboratoriais e machine learning no diagnóstico e prognóstico da COVID-19	39
2.5.2 Metabolômica integrativa na pesquisa de biomarcadores contra a COVID-19	42
2.5.3 Espectroscopia de infravermelho na detecção da COVID-19: uma abordagem inovadora	44
2.6 ABORDAGENS NÃO FARMACOLÓGICAS NO TRATAMENTO DE PACIENTES COM COVID-19	46
2.6.1. Alimentação e nutrição como abordagem complementar na recuperação dos pacientes COVID-19	46
2.7 ABORDAGENS FARMACOLÓGICAS NO TRATAMENTO DE PACIENTES COM COVID-19	47
2.7.1 Planejamento racional de fármacos como uma abordagem promissora na descoberta de novos medicamentos	47
2.7.2 Inteligência artificial e machine learning na descoberta de novos fármacos	51
2.7.3 Inteligência artificial na triagem de novos fármacos	54
2.7.4 Inteligência artificial na previsão das propriedades físico-químicas	54
2.7.5 Inteligência artificial na previsão da biodisponibilidade	55
2.7.6 Inteligência artificial na previsão da toxicidade	57
1 CAPÍTULO I - FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 E O IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO AUMENTO DE INTERNAÇÕES HOSPITALARES	59

1.1 RESUMO.....	60
1.2 INTRODUÇÃO	61
1.3 OBJETIVOS	62
1.3.1 Objetivo geral	62
1.3.2 Objetivos específicos	62
1.4 MATERIAL E MÉTODO	63
1.4.1 ESTUDO I: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS AO ATRASO NO DIAGNÓSTICO E MORTALIDADE POR COVID-19 NO ÂMBITO NACIONAL.....	63
1.4.2 ESTUDO II: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 NO ÂMBITO INTERNACIONAL.....	64
1.4.3 ESTUDO III: IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO BRASIL E NO MUNDO SOBRE NO AUMENTO DO NÚMERO DE INTERNAÇÕES POR COVID-19 65	
1.5 RESULTADOS	67
1.5.1 ESTUDO I: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS AO ATRASO NO DIAGNÓSTICO E MORTALIDADE POR COVID-19 NO ÂMBITO NACIONAL.....	67
1.5.2 ESTUDO II: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 NO ÂMBITO INTERNACIONAL.....	75
1.5.3 ESTUDO III: IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO BRASIL E NO MUNDO SOBRE NO AUMENTO DO NÚMERO DE INTERNAÇÕES POR COVID-19 80	
1.6 DISCUSSÃO	82
1.6.1 ESTUDO I: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS AO ATRASO NO DIAGNÓSTICO E MORTALIDADE POR COVID-19 NO ÂMBITO NACIONAL.....	82
1.6.2 ESTUDO II: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 NO ÂMBITO INTERNACIONAL.....	84
1.6.3 ESTUDO III: IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO BRASIL E NO MUNDO SOBRE NO AUMENTO DO NÚMERO DE INTERNAÇÕES POR COVID-19 88	
1.7 CONCLUSÃO.....	90

2 CAPÍTULO II - AVALIAÇÃO DO EFEITO DOS ALIMENTOS E NUTRIENTES COMO ABORDAGENS COMPLEMENTARES NA RECUPERAÇÃO DE PACIENTES COM COVID-19 EM 170 PAÍSES.....	91
2.1 RESUMO.....	92
2.2 INTRODUÇÃO	93
2.3 OBJETIVOS	94
2.3.1 Objetivo geral:	94
2.3.2 Objetivos específicos:.....	94
2.4 MATERIAL E MÉTODOS.....	94
2.4.1 Desenho do estudo	94
2.4.2 Conjunto de dados	95
2.4.3 Variáveis do estudo.....	96
2.4.4 Análise estatística	97
2.5 RESULTADOS	98
2.5.1 Considerações gerais dos resultados	98
2.5.2 Abastecimento de alimentos e taxas de recuperação da COVID-19.....	98
2.5.3 Análise multivariada do modelo linear generalizado	99
2.5.3.2 Efeito do abastecimento alimentar nas taxas de recuperação da COVID-19 segundo grupos de países	103
2.6 DISCUSSÃO	107
2.6.1 Efeitos do fornecimento de alimentos nas taxas de recuperação da COVID-19	
107	
2.6.2 Efeitos do fornecimento de macro e micronutrientes nas taxas de recuperação da COVID-19.....	109
2.7 CONCLUSÃO.....	113
3 CAPÍTULO III - ACURÁCIA DA TÉCNICA DE ESPETROFOTOMETRIA DE INFRAVERMELHO NO DIAGNOSTICO DE COVID-19: UM ESTUDO DE REVISÃO SISTEMÁTICA COM META-ANÁLISE.....	114
3.1 RESUMO.....	115
3.2 INTRODUÇÃO	116
3.3 OBJETIVOS	116
3.3.1 Objetivo geral:	116
3.3.2 Objetivos específicos:.....	117
3.4 MATERIAL E MÉTODOS.....	117

3.4.1 Meta-análise.....	118
3.5 RESULTADOS	119
3.5.1 Parâmetros analíticos.....	123
3.5.2 Teste de acurácia diagnóstica.....	127
3.6 DISCUSSÃO	129
3.7 CONCLUSÃO.....	134
4 CAPÍTULO IV - DESENVOLVIMENTO DE MODELOS PREDITIVOS E IDENTIFICAÇÃO DE BIOMARCADORES PROGNÓSTICOS EM COVID-19, HIV E TUBERCULOSE POR MEIO DE INTELIGÊNCIA ARTIFICIAL E MACHINE LEARNING.....	135
4.1 RESUMO.....	136
4.2 INTRODUÇÃO	137
4.3 OBJETIVOS	137
4.3.1 Objetivo geral	137
4.3.2 Objetivos específicos	138
4.4 MATERIAL E MÉTODOS.....	138
4.4.1 Estudo I: Análise de dados dos exames bioquímicos, hematológicos e de urinálise de pacientes COVID-19 atendidos no Hospital Israelita Albert Einstein (Brasil) visando a predição do diagnóstico e investigação de potenciais Biomarcadores prognósticos.....	139
4.4.1.2 Pré-processamento de dados para Machine learning	141
4.4.2 Estudo II: Análise de dados dos exames bioquímicos e hematológicos de pacientes COVID-19, HIV, Tuberculose e co-infectados HIV/TB atendidos em um Hospital Regional de referência do Norte de Moçambique (Hospital Geral de Marrere, Província de Nampula) visando predição do diagnóstico e investigação de biomarcadores associados a essas doenças.....	144
4.4.2.8 Análise univariada	148
4.4.3 Estudo III: Predição de diagnóstico de COVID-19 usando dados clínicos de pacientes COVID-19 atendidos na rede de Farmácia Drugstore distribuída em todo território Nacional.	150
4.5 RESULTADOS	154
4.5.1 Estudo I: Análise de dados dos exames bioquímicos, hematológicos e de urinálise de pacientes COVID-19 atendidos no Hospital Israelita Alberto Einstein	

(Brasil) visando a predição do diagnóstico e investigação de potenciais Biomarcadores prognósticos.....	154
4.4.2 Estudo II: Análise de dados dos exames bioquímicos e hematológicos de pacientes COVID-19, HIV, Tuberculose e co-infectados HIV/TB atendidos em um Hospital Regional de referência do Norte de Moçambique (Hospital Geral de Marrere, Província de Nampula) visando predição do diagnóstico e investigação de biomarcadores associados a essas doenças.....	161
4.5.3 Estudo III: Predição de diagnóstico de COVID-19 usando dados clínicos de pacientes COVID-19 atendidos na rede de Farmácia Drugstore distribuída em todo território Nacional.....	178
4.6 DISCUSSÃO.....	182
4.6.1 Estudo I: Análise de dados dos exames bioquímicos, hematológicos e de urinálise de pacientes COVID-19 atendidos no Hospital Israelita Albert Einstein (Brasil) visando a predição do diagnóstico e investigação de potenciais Biomarcadores prognósticos.....	182
4.6.2 Estudo II: Análise de dados dos exames bioquímicos e hematológicos de pacientes COVID-19, HIV, Tuberculose e co-infectados HIV/TB atendidos em um Hospital Regional de referência do Norte de Moçambique (Hospital Geral de Marrere, Província de Nampula) visando predição do diagnóstico e investigação de biomarcadores associados a essas doenças.....	186
4.6.3 Estudo III: Predição de diagnóstico de COVID-19 usando dados clínicos de pacientes COVID-19 atendidos na rede de Farmácia Drugstore distribuída em todo território Nacional.....	190
4.7 CONCLUSÃO.....	194
5 CAPÍTULO V - EXPLORANDO A METABOLÔMICA INTEGRATIVA POR LC-MS, GC-MS E RMN COM INTELIGÊNCIA ARTIFICIAL NA DESCOBERTA DE NOVOS BIOMARCADORES ASSOCIADOS AO DIAGNOSTICO E PROGNÓSTICO DA COVID-19.....	197
5.1 RESUMO.....	198
5.2 INTRODUÇÃO.....	199
5.3 OBJETIVO.....	200
5.3.1 Objetivo geral:.....	200
5.3.2 Objetivos específicos:.....	200
5.4 MATERIAL E MÉTODOS.....	201

5.4.1. Estudo I: Diagnóstico e prognóstico de COVID-19 empregando análise de plasma e soro via LC-MS e Machine learning.....	201
5.4.2 Estudo II: Novos biomarcadores COVID-19 identificados por meio de análise de dados multiômicos: ácido N-acetil-4-O-acetilneuramínico, N-acetil-L-alanina, N-acetiltryptofano, palmitoilcarnitina e 1-miristato de glicerol	206
5.5 RESULTADOS	215
5.5.1. Estudo I: Diagnóstico e prognóstico de COVID-19 empregando análise de plasma e soro via LC-MS e Machine learning	215
5.5.2 Estudo II: Novos biomarcadores COVID-19 identificados por meio de análise de dados multi-ômicos: ácido N-acetil-4-O-acetilneuramínico, N-acetil-L-alanina, N-acetiltryptofano, palmitoilcarnitina e 1-miristato de glicerol	225
5.6 DISCUSSÃO	242
5.6.1. ESTUDO I: DIAGNÓSTICO E PROGNÓSTICO DE COVID-19 EMPREGANDO ANÁLISE DE PLASMA E SORO VIA LC-MS E MACHINE LEARNING	242
5.6.2 Estudo II: Novos biomarcadores COVID-19 identificados por meio de análise de dados multi-ômicos: ácido N-acetil-4-O-acetilneuramínico, N-acetil-L-alanina, N-acetiltryptofano, palmitoilcarnitina e 1-miristato de glicerol	250
5.7 CONCLUSÃO.....	254
6 CAPÍTULO VI - ABORDAGEM INTEGRATIVA NA BUSCA POR INIBIDORES DA PROTEÍNA SPIKE (RBD) DO SARS-COV-2: TRIAGEM VIRTUAL, ANÁLISE DE DRUG-LIKENESS, PREDIÇÕES ADMET, DINÂMICA MOLECULAR E INTELIGÊNCIA ARTIFICIAL PARA O TRATAMENTO DA COVID-19.....	255
6.1 RESUMO.....	256
6.2 INTRODUÇÃO	257
6.3 OBJETIVOS	258
6.3.1 Objetivo geral:	258
6.3.2 Objetivos específicos:.....	258
6.4 MATERIAL E MÉTODOS	259
6.4.1 Fluxograma do estudo.....	259
6.4.2 Seleção de proteína Spike (RBD): variante selvagem e Ômicron	259
6.4.3 Preparo da proteína alvo	260
6.4.4 Coleta e preparo dos ligantes.....	261
6.4.5 Triagem virtual e docking molecular	262

6.4.6 Avaliação da influência do estado estereoisomérico, tautomérico e estado de protonação em pH fisiológico dos ligantes na sua afinidade pela RBD.....	263
6.4.7 Análise de drug-likeness e farmacocinética (ADME).....	263
6.4.8 Predição da toxicidade aguda e crónica.....	264
6.4.9 Simulações de dinâmica molecular	264
6.4.10 Cálculo de energia livre de ligação MM/GBSA.....	266
6.4.11 Cálculo de energia livre de ligação MM/PBSA	266
6.4.11 Simulações de pós – dinâmica molecular	267
6.4.12 Desenvolvimento de modelos de Inteligência artificial e machine learning baseados em QSAR-3D.....	267
6.5 RESULTADOS	271
6.5.1 Estudo I: Naringenina-4'-glicuronídeo como novo candidato a fármaco contra a variante Omicron da COVID-19: um estudo baseado em docking molecular, dinâmica molecular, MM/PBSA e MM/GBSA	271
6.5.2 Estudo II: Triagem virtual, docking molecular, análise de drug-likeness e predições ADMET: proteína Spike (RBD) do SARS-CoV-2 tipo selvagem	285
6.5.3 Validação dos resultados das análises in sílico do estudo I e II via Machine learning	299
6.6 DISCUSSÃO	305
6.6.1 Estudo I: Naringenina-4'-glicuronídeo como novo candidato a fármaco contra a variante Omicron da COVID-19: um estudo baseado em docking molecular, dinâmica molecular, MM/PBSA e MM/GBSA	306
6.6.2 Estudo II: Triagem virtual, docking molecular, análise de drug-likeness e predições ADMET: proteína Spike (RBD) do SARS-CoV-2 tipo selvagem	309
6.7 CONCLUSÃO.....	312
7 CAPÍTULO VII - ABORDAGEM INTEGRATIVA REVELA CANDIDATOS MULTIFACETADOS A FÁRMACOS PARA O TRATAMENTO SIMULTÂNEO DE COVID-19, HEPATITE, DENGUE E HIV USANDO INTELIGÊNCIA ARTIFICIAL MULTI-TARGET, E POLIFARMACOLOGIA	313
7.1 RESUMO.....	314
7.2 INTRODUÇÃO	315
7.3 OBJETIVOS	316
7.3.1 Objetivo geral:	316
7.3.2 Objetivos específicos:.....	316

7.4 MATERIAL E MÉTODOS	317
7.4.1 Fluxograma do estudo.....	317
7.4.2 Descrição do banco de dados utilizado para Machine learning: ChEMBL Database.....	318
7.4.3 Pré-processamento dos dados para machine learning	319
7.4.4 Seleção de recursos.....	320
7.4.5 Divisão de dados de treinamento e de teste	320
7.4.6 Inteligência artificial e machine learning	320
7.4.7 Investigação dos descritores moleculares mais importantes na atividade biológica multi-target usando SHAP values	321
7.4.8 Validação dos resultados de machine learning multi-target via modelagem por docking molecular e simulações de dinâmica molecular	321
7.4.9 Docking molecular.....	323
7.4.10 Simulações de dinâmica molecular	323
7.4.11 Cálculo de energia livre de ligação MM/GBSA.....	324
7.5 RESULTADOS	324
7.5.1 Análise do espaço químico.....	324
7.5.2 Machine learning.....	326
7.5.3 Utilização dos cinco modelos de machine learning de melhor desempenho em um conjunto de dados externo: Human Metabolome Database.....	329
7.5.4 Análises de docking molecular	331
7.5.5 Simulações de dinâmica molecular	332
7.6 DISCUSSÃO	341
7.7 CONCLUSÃO.....	344
8 COMENTÁRIOS FINAIS.....	346
REFERÊNCIAS.....	348
APÊNDICE I – BIOGRAFIA DO AUTOR.....	401
APÊNDICE II – GALERIA DE FOTOS DURANTE A PARTICIPAÇÃO COMO PALESTRANTE EM EVENTOS CIENTÍFICOS INTERNACIONAIS NO REINO UNIDO (INGLATERRA) E NO BRASIL.....	402
APÊNDICE III – ARTIGOS CIENTÍFICOS PUBLICADOS (PRIMEIRA PÁGINA)..	403

1 INTRODUÇÃO

A pandemia de COVID-19, desencadeada pelo SARS-CoV-2, emergiu como um dos desafios mais prementes para a saúde global, demandando uma resposta científica e clínica robusta e multifacetada ¹. Diante dessa complexidade, uma abordagem multidisciplinar no diagnóstico e tratamento da doença torna-se não apenas pertinente, mas essencial ^{2,3}.

Este estudo, intitulado "COVID-19, abordagem multidisciplinar no diagnóstico e terapêutica: integrando métodos de diagnóstico e prognóstico, metabolômica na identificação de biomarcadores e inteligência artificial na descoberta de novos fármacos", propõe uma investigação que não se limita apenas à análise dos aspectos clínicos e biológicos da doença, mas também se estende à compreensão de sua epidemiologia em diferentes contextos, incluindo o âmbito nacional e de outros países ao redor do mundo ^{2,4,5}.

A primeira dimensão deste trabalho visa aprimorar os métodos de diagnóstico e prognóstico da COVID-19, visando à detecção precoce e precisa da infecção, bem como à identificação de fatores de risco e prognóstico associados à gravidade da doença ⁶⁻⁸. Para isso, serão exploradas técnicas avançadas de inteligência artificial e *machine learning* para análise de dados clínicos e laboratoriais dos pacientes COVID-19⁶, considerando também as variações na epidemiologia da doença em diferentes regiões geográficas ^{7,9}.

Em seguida, esta pesquisa abordará o papel da metabolômica na identificação de biomarcadores para a COVID-19. A análise do perfil metabólico dos pacientes pode fornecer informações valiosas sobre as alterações bioquímicas associadas à infecção viral, permitindo a identificação de marcadores moleculares para diagnóstico, prognóstico e monitoramento da resposta ao tratamento, considerando também as particularidades epidemiológicas de diferentes populações ¹⁰.

Além disso, este estudo explorará o potencial da inteligência artificial na descoberta de novos fármacos para o tratamento da COVID-19, levando em consideração as variações genéticas do vírus e as características epidemiológicas de diferentes regiões ^{11,12}. Por meio de técnicas avançadas de *machine learning* e modelagem computacional, serão desenvolvidos modelos preditivos para identificar compostos com atividade antiviral e potencial terapêutico contra o SARS-CoV-2 ^{11,13}.

Ao integrar essas diferentes abordagens e considerar a epidemiologia da COVID-19 no Brasil e em outros países ao redor do mundo, este estudo visa fornecer *insights* importantes para o combate a essa pandemia e para o enfrentamento de futuras emergências de saúde pública. Espera-se que os resultados desta pesquisa possam contribuir para uma resposta mais eficaz e adaptada às diversas realidades epidemiológicas da doença.

1.1 OBJETIVOS

1.1.1 Objetivo geral

O objetivo geral deste estudo foi desenvolver uma abordagem multidisciplinar para compreender a epidemiologia e a fisiopatologia da COVID-19, desenvolver novos métodos de diagnóstico, investigar novos biomarcadores prognósticos e novos fármacos para o tratamento da doença, integrando métodos de diagnóstico e prognóstico aprimorados, metabolômica e inteligência artificial.

1.1.2 Objetivos específicos

Capítulo I – Investigar os fatores de risco associados a mortalidade de COVID-19 e o impacto da rejeição da vacinação no aumento de internações hospitalares em diversos países do Mundo, incluindo Brasil;

Capítulo II – Avaliar o efeito dos alimentos e nutrientes como abordagem complementar na recuperação de pacientes com COVID-19 em 170 países;

Capítulo III – Conduzir uma revisão sistemática com meta-análise visando avaliar a acurácia da técnica de espectroscopia de infravermelho no diagnóstico de COVID-19;

Capítulo IV – Identificar potenciais biomarcadores para o diagnóstico e prognóstico da COVID-19, HIV, tuberculose e coinfeção utilizando exames bioquímicos, inteligência artificial e machine learning;

Desenvolver modelos preditivos de inteligência artificial e *machine learning* visando a identificação de potenciais biomarcadores diagnósticos e prognósticos em COVID-19, HIV e tuberculose;

Capítulo V – Explorar a metabolômica integrativa por cromatografia líquida acoplada a espectrometria de massas (LC-MS), cromatografia a gás acoplada a espectrometria de massas (GC-MS) e Ressonância Magnética Nuclear (RMN) com inteligência artificial e *machine learning* visando a descoberta de novos biomarcadores associados ao diagnóstico e prognóstico da COVID-19;

Capítulo VI – Utilizar uma abordagem integrativa visando a busca por candidatos a fármacos inibidores da proteína Spike (RBD) do SARS-CoV-2 – triagem virtual, análise de *drug-likeness*, previsões das propriedades farmacocinéticas e de toxicidade (ADMET), simulações de dinâmica molecular e inteligência artificial para o tratamento da COVID-19;

Capítulo VII – Utilizar uma abordagem integrativa visando a identificação de novos candidatos multifacetados a fármacos para o tratamento simultâneo de COVID-19, hepatite, dengue e HIV usando inteligência artificial *multi-target* e polifarmacologia.

1.2 JUSTIFICATIVA

A pandemia de COVID-19 representou um desafio sem precedentes para a saúde global, exigindo uma resposta científica e clínica urgente e abrangente. Diante da complexidade e gravidade da situação, torna-se imprescindível uma abordagem multidisciplinar para compreender, prevenir, diagnosticar e tratar eficazmente essa doença.

Este estudo de doutorado, intitulado "COVID-19, abordagem multidisciplinar no diagnóstico e terapêutica: integrando métodos de diagnóstico e prognóstico, metabolômica na identificação de biomarcadores e inteligência artificial na descoberta de novos fármacos", justifica-se pela necessidade premente de desenvolver estratégias inovadoras e integradas para enfrentar a pandemia.

A primeira dimensão deste estudo visa aprimorar os métodos de diagnóstico e prognóstico da COVID-19, utilizando técnicas avançadas de inteligência artificial e *machine learning* para analisar dados clínicos e laboratoriais dos pacientes. A detecção precoce e precisa da infecção, juntamente com a identificação de fatores de risco e prognóstico, é crucial para guiar intervenções terapêuticas e mitigar os impactos da doença.

Além disso, a investigação do perfil metabólico dos pacientes, por meio da metabolômica, permitirá a identificação de biomarcadores que possam auxiliar no

diagnóstico, prognóstico e monitoramento da resposta ao tratamento da COVID-19. Essa abordagem é fundamental para personalizar a terapia e melhorar os resultados clínicos, considerando as particularidades epidemiológicas de diferentes populações.

Por fim, a utilização de inteligência artificial na descoberta de novos fármacos para o tratamento da COVID-19 visa fornecer soluções terapêuticas inovadoras e eficazes. A consideração das variações genéticas do vírus e as características epidemiológicas de diferentes regiões são essenciais para desenvolver terapias adaptadas e promissoras.

Ao integrar essas abordagens multidisciplinares e considerar a epidemiologia da COVID-19 no Brasil e em outros países ao redor do mundo, este estudo pretende oferecer contribuições significativas para o combate à pandemia. Espera-se que os resultados desta pesquisa possam informar políticas de saúde pública, orientar práticas clínicas e inspirar futuras investigações na área, visando uma resposta mais eficaz e adaptada às diversas realidades epidemiológicas da doença

2 REVISÃO DE LITERATURA

2.1 BREVE HISTÓRICO DO SURGIMENTO DA COVID-19

Desde 8 de dezembro de 2019, vários casos de pneumonia de etiologia desconhecida foram relatados em Wuhan, província de Hubei, na China ^{14–16}. Em 7 de janeiro, um novo coronavírus foi identificado pelo Centro Chinês de Controle e Prevenção de Doenças (CDC) a partir de uma amostra de esfregaço da garganta de um paciente, e foi posteriormente denominado 2019-nCoV pela Organização Mundial de Saúde (OMS) ¹⁷.

Quase um mês após a confirmação do novo coronavírus pelas autoridades chinesas ainda havia muitas perguntas sem respostas sobre a doença, tais como: qual era o efeito do vírus no corpo humano? qual era o espectro dos sintomas possíveis? e quais eram os grupos mais vulneráveis?

No dia 30 de janeiro de 2020, um conjunto de 20 médicos que estavam na linha de combate ao surto no Hospital de Jinyintan, em Wuhan (China), publicaram um estudo científico no periódico *The Lancet* respondendo algumas dessas perguntas, no

qual descreveram uma análise detalhada dos primeiros 99 pacientes infectados que eles trataram. Todos os 99 pacientes chegaram ao hospital com pneumonia causada pelo vírus 2019-nCoV¹⁸. A maioria dos pacientes trabalhava ou vivia próximo ao mercado atacadista local de frutos do mar de Huanan, onde também eram vendidos animais vivos. A idade média foi de 55,5 anos, sendo a maioria homens. Todos os pacientes testaram positivo para o vírus por RT-PCR (do inglês, reação de transcriptase reversa seguida de reação em cadeia da polimerase) em tempo real. Mais da metade tinham doenças crônicas. Os sintomas mais comuns foram febres, tosse e falta de ar. Alguns também apresentaram outras manifestações como dores musculares, confusão, dor de cabeça e dor de garganta. A maioria dos pacientes tinha pneumonia nos dois pulmões, alguns tinham manchas múltiplas nos pulmões e um paciente teve pneumotórax. Cerca de 17% desenvolveram síndrome do desconforto respiratório agudo, com 11% piorando rapidamente e morrendo por falência de múltiplos órgãos¹⁸.

Em 26 de fevereiro de 2020 foi diagnosticado o primeiro caso da doença, denominada COVID-19, na América do Sul, e foi de um homem de nacionalidade brasileira de 61 anos, diagnosticado no Hospital Albert Einstein, em São Paulo (Brasil), após retornar de viagem na região de Lombardia, Itália¹⁹.

Em 11 de Março de 2020, a COVID-19 foi caracterizada pela OMS como uma pandemia²⁰. O termo “pandemia” se refere à distribuição geográfica de uma doença e não à sua gravidade. Portanto, a partir de 11 de março de 2020 considerou-se que havia surtos de COVID-19 em vários países e regiões do mundo²¹.

Em 30 de Maio de 2020, a OMS decretou a COVID-19 como uma Emergência de Saúde Pública de Importância Internacional (ESPII), o mais alto nível de alerta da Organização, conforme previsto no Regulamento Sanitário Internacional. Foram levados em conta vários aspectos epidemiológicos, incluindo o potencial de transmissão, a população suscetível, a severidade da doença, a capacidade de impactar viagens internacionais, entre outros fatores específicos²¹.

2.2 DADOS EPIDEMIOLÓGICOS DA COVID-19 NO BRASIL E NO MUNDO

A COVID-19 sobrecarregou os sistemas globais de saúde de uma forma sem precedentes, com impacto importante nas atividades dos serviços de cuidados

primários, hospitais, urgências e unidades de cuidados intensivos (UCI). Além disso, a rápida propagação da doença e o número significativo de mortes associadas levaram os governos a implementar medidas de contenção da doença que causaram efeitos devastadores nas vidas das populações e nas economias em todo o mundo²²⁻²⁴.

Segundo as estimativas da OMS, até fevereiro de 2024 foram registrados um total acumulado de 774,4 milhões de casos confirmados 7,0 milhões de mortes em todo o mundo²⁵. Em termos do número total de casos confirmados, o Brasil ocupa a sexta posição com um total acumulado de 37,5 milhões de casos, ficando apenas atrás dos Estados Unidos (103,4 milhões), China (99,3 milhões), Índia (45 milhões), França (39 milhões) e Alemanha (38,4 milhões). Em relação ao número de óbitos, o Brasil ocupa a segunda posição com 702,1 mil óbitos, ficando apenas atrás dos Estados Unidos (1,2 milhão de mortes)²⁶.

2.3 FATORES DE RISCO PARA COVID-19

De acordo com revisões sistemáticas e meta-análises (RSMA) da literatura, a idade tem sido apontada como fator de risco mais importante na mortalidade e complicações por COVID-19²⁷⁻²⁹. Adicionalmente, alguns fatores de risco metabólicos também estão associados com a mortalidade pela doença, como é apresentado na RSMA de Bahram (2020), que analisou 1.124 estudos e demonstrou que fatores como obesidade (29%, intervalo de confiança 95% - IC 95%: 14 – 47%), diabetes (22%, IC 95%: 12% - 33%) e hipertensão (32%, IC 95%: 12% - 56%) são fatores associados com a mortalidade e complicações da COVID-19³⁰.

Para além dos fatores mencionados anteriormente, algumas revisões sistemáticas (RS) apontaram que a disparidade social, racial/étnica e econômica também foram fatores de risco associados a mortalidade por COVID-19, e existem vários fatores interligados que contribuem para essa disparidade³¹⁻³³. Dentre eles destacam-se:

- (i) Acesso a cuidados de saúde: comunidades marginalizadas muitas vezes têm acesso limitado a serviços de saúde de qualidade, incluindo testes, tratamento e vacinação, que pode resultar em diagnóstico tardio e cuidados inadequados, aumentando o risco de complicações e morte pela doença;

- (ii) Condições socioeconômicas: pessoas de grupos étnicos minoritários ou de baixa renda pode estar mais expostas a condições de trabalho que não permitem o distanciamento social adequado, como empregos essenciais ou informais, conseqüentemente aumentando a probabilidade de contrair o vírus;
- (iii) Habitação superlotada: em muitas comunidades marginalizadas, a habitação superlotada é comum, o que torna difícil o isolamento e aumenta a transmissão do vírus entre os membros da família;
- (iv) Acesso à informação: barreiras linguísticas, falta de acesso à tecnologia ou à informação precisa sobre a pandemia também podem contribuir para taxas mais altas de infecção e mortalidade em certas comunidades ³¹⁻³³.

2.4 COVID LONGA: O NOVO DESAFIO DA PANDEMIA

A maioria das pessoas que apresenta COVID-19 recupera totalmente, mas as evidências atuais sugerem que aproximadamente 10–20% das pessoas experimentam uma variedade de efeitos a médio e longo prazo após recuperarem da doença inicial, a esse tipo de síndrome pós-COVID é chamada COVID longa ³⁴.

Segundo a recente RSMA de Chen (2022), os sintomas mais comuns associados à condição pós-COVID-19 incluem fadiga, falta de ar e disfunção cognitiva (por exemplo, confusão, esquecimento ou falta de foco ou clareza mental) ³⁵. A condição pós-COVID-19 pode afetar a capacidade de uma pessoa realizar atividades diárias, como trabalho ou tarefas domésticas ^{34,35}.

No Brasil, um recente estudo foi conduzido pela Fundação Oswaldo Cruz (Fiocruz) ³⁶, envolvendo 646 pacientes com COVID-19 acompanhados durante 14 meses. Destes, 50,2% apresentaram síndrome de COVID-longa. Vinte e três sintomas diferentes foram relatados. Os mais frequentes foram fadiga (35,6%), tosse persistente (34,0%), dispneia (26,5%), perda de olfato/paladar (20,1%) e dores de cabeça frequentes (17,3%). Também foram relatados transtornos mentais (20,7%), alteração da pressão arterial (7,4%) e trombose (6,2%). A maioria dos pacientes apresentou 2 a 3 sintomas ao mesmo tempo. A COVID longa começou após infecção leve, moderada e grave em 60%, 13% e 27% dos casos, respectivamente, e não se restringiu a faixas etárias específicas ³⁶. O estudo concluiu que pacientes mais velhos

tendem a apresentar sintomas mais graves, levando a um período pós-COVID-19 mais longo. A presença de sete comorbidades foi correlacionada com a gravidade da infecção, sendo a própria gravidade o principal fator que determinou a duração dos sintomas nos casos longos de COVID ³⁶.

A causa de desenvolvimento da COVID-19 é pouco conhecida, e de fato existe uma recente RS que evidenciou que não há associação entre a COVID-19 longa com os fatores de risco clássicos da COVID-19, ou seja, a idade, comorbidades e sexo ³⁷. Possíveis hipóteses das causas do aparecimento da COVID longa são levantadas por recentes estudos de Theoharides (2022) e Grandjean (2020), que correlacionam a persistência de sintomas e danos ocasionados pela invasão viral com a capacidade da proteína Spike (S) em evadir a resposta imune inata e adaptativa ^{38,39}.

Há um grande desafio para o diagnóstico de COVID-19 longa e quatro principais pontos ajudam a entender:

- (i) Pacientes que tiveram histórico de sintomas típicos de COVID-19 agudo com resultado de RT-PCR (do inglês, reação de transcriptase reversa seguida de reação em cadeia da polimerase) positivo, apresentando sintomas de longa duração, o diagnóstico de COVID longo é direto;
- (ii) Pacientes com sintomas agudos de COVID-19 com resultados RT-PCR negativo, apresentando sintomas longos, representam um desafio real na prática clínica;
- (iii) Uma proporção significativa de pacientes com COVID-19 inicialmente assintomático que desenvolvem sintomas longos (aumentando a confusão diagnóstica);
- (iv) Existe uma variação do tempo de duração dos sintomas agudos entre os pacientes, adicionando ainda mais as dificuldades em diferenciar pacientes COVID-19 longa e COVID-19 agudo ⁴⁰.

2.5 DIAGNÓSTICO E PROGNÓSTICO DA COVID-19

2.5.1 Biomarcadores laboratoriais e machine learning no diagnóstico e prognóstico da COVID-19

A utilização de biomarcadores laboratoriais por meio de inteligência artificial e *machine learning* tem emergido como uma abordagem promissora no diagnóstico e

prognóstico da COVID-19 ⁴¹. A pandemia global desencadeada pelo coronavírus SARS-CoV-2 desafiou os sistemas de saúde a nível mundial, destacando a necessidade de ferramentas precisas e eficazes para identificar e monitorar a doença ⁴². Nesse contexto, biomarcadores laboratoriais, que são substâncias mensuráveis no corpo humano que refletem a presença ou a gravidade de uma condição específica, têm sido amplamente investigados como indicadores úteis na detecção e avaliação da COVID-19 ⁴¹. Além disso, o uso de técnicas avançadas de *machine learning*, um ramo da inteligência artificial, tem permitido a análise de grandes conjuntos de dados clínicos e laboratoriais para desenvolver modelos preditivos precisos que podem auxiliar no diagnóstico precoce e no prognóstico da doença ⁴³.

Inteligência artificial (IA) é um campo da ciência da computação que se concentra no desenvolvimento de sistemas capazes de realizar tarefas que normalmente exigiriam inteligência humana ⁴⁴. A IA visa criar máquinas capazes de aprender, raciocinar, perceber, tomar decisões e resolver problemas de maneira semelhante aos seres humanos. Já, *machine learning* é uma subcategoria da inteligência artificial que se concentra no desenvolvimento de algoritmos e técnicas estatísticas que permitem aos computadores aprender e melhorar automaticamente a partir de experiências passadas, sem serem explicitamente programados para isso. Em vez de seguir instruções específicas, os sistemas de *machine learning* usam dados para aprender padrões e tomar decisões ⁴⁴.

Machine learning é uma ferramenta eficaz e inovadora capaz de auxiliar profissionais de saúde, legisladores e outras partes interessadas durante os processos de tomada de decisão. No campo do diagnóstico clínico da COVID-19, estas análises preditivas baseadas em biomarcadores podem ajudar a otimizar o rastreio de pacientes com doença grave, minimizando a mortalidade e a hospitalização, e reduzindo os atrasos no atendimento ⁴⁵. *Machine learning* tem também ajudado na identificação e interpretação dos biomarcadores diagnósticos e prognósticos da COVID-19 ⁴⁶. Por exemplo no estudo de Brunati (2020) foram desenvolvidos dois modelos de *machine learning* (*Random forest* e *decision tree*) para o diagnóstico de COVID-19 considerando exames clínicos de rotina (contagem de leucócitos, níveis plasmáticos de plaquetas, proteína C reativa, aspartato aminotransferase, alanina aminotransferase, gama glutamil transferase, fosfatase alcalina e lactato desidrogenase) de 279 pacientes internados nos serviços de emergência do Hospital San Raffaele (Milão, Itália) com COVID-19. A acurácia

diagnóstica dos modelos de *random forest* e de *decision tree* foram de 82% e 86%, ao passo que a sensibilidade foi de 92% e 95%, respectivamente. Ainda nesse estudo, a dosagem da enzima aspartato aminotransferase (AST), dosagem dos níveis dos linfócitos e de lactato desidrogenase (LDH) foram os três biomarcadores mais importantes no diagnóstico da COVID-19 ⁴⁶.

Assim, é essencial identificar potenciais biomarcadores prognósticos para cuidados mais precoces e direcionados, especialmente considerando que alguns pacientes com COVID-19 desenvolvem doença grave, que está associada a um maior risco de hospitalização. Os biomarcadores fornecem uma abordagem dinâmica e poderosa para a compreensão do espectro da doença, com aplicações em epidemiologia observacional e analítica, ensaios clínicos randomizados, triagem e diagnóstico e prognóstico ⁴⁷. Recentemente, estudos que investigam biomarcadores para diagnosticar COVID-19 em estágios iniciais têm sido incentivados em todo o mundo, com o objetivo de proporcionar um encaminhamento mais rápido para o tratamento e assim, reduzir problemas de saúde associados à doença ^{48,49}.

Segundo a RSMA de Kermali (2020), envolvendo 34 estudos, identificou os seguintes biomarcadores como sendo importantes no prognóstico da COVID-19: proteína C reativa (PCR), amiloide A sérica, interleucina-6, lactato desidrogenase, relação neutrófilos-linfócitos, dímero D, troponina cardíaca, biomarcadores renais, linfócitos e contagem de plaquetas. A maioria apresentou níveis significativamente mais elevados em pacientes com complicações graves de infecção por COVID-19 em comparação com os seus homólogos não graves. Somente a contagem de linfócitos e plaquetas mostrou níveis significativamente mais baixos em pacientes graves em comparação com pacientes não graves ⁵⁰. Adicionalmente, uma recente RS de Dominguez (2023), envolvendo estudos (n=2.237 pacientes), identificou o neurofilamento de cadeia leve (NfL) e proteína glial fibrilar ácida (GFAP) como biomarcadores associados a complicações neuronais causadas por COVID-19 ⁵¹.

A tempestade de citocinas é uma reação imunológica desregulada do corpo, na qual o sistema imunológico libera uma quantidade excessiva de citocinas (proteínas de sinalização celular) como resposta a uma infecção ⁵². Na COVID-19, algumas pessoas desenvolvem essa tempestade de citocinas, especialmente em casos graves. Durante a infecção pelo coronavírus, o sistema imunológico pode reagir de maneira exagerada, liberando grandes quantidades de citocinas. Isso pode causar uma inflamação generalizada no corpo, levando a danos nos tecidos e órgãos,

podendo até mesmo ser fatal ⁵². Os sintomas da tempestade de citocinas podem incluir febre alta, inflamação generalizada, dificuldade respiratória, queda abrupta na pressão arterial e danos aos órgãos. Vários biomarcadores têm sido associados à tempestade de citocinas na COVID-19. Estes são marcadores biológicos que podem indicar a presença ou gravidade da tempestade de citocinas e da resposta inflamatória exacerbada. Segundo a RS de Melo (2021) ⁵³, os biomarcadores mais importantes na tempestade de citocinas incluem:

- Interleucina-6 (IL-6): É uma citocina pró-inflamatória que costuma estar elevada em casos de tempestade de citocinas. Elevações significativas de IL-6 têm sido observadas em pacientes com COVID-19 grave.
- Interleucina-1B (IL-1B): Na COVID-19 grave, a liberação excessiva de IL-1B e outras citocinas pró-inflamatórias pode levar à inflamação generalizada e danos aos tecidos, incluindo os pulmões. Isso pode resultar em complicações graves, como síndrome do desconforto respiratório agudo (SDRA), falha de múltiplos órgãos e até mesmo morte.
- Fator de Necrose Tumoral alfa (TNF-alfa): Outra citocina pró-inflamatória que pode estar aumentada durante a tempestade de citocinas.
- PCR: Um marcador de inflamação sistêmica, a PCR pode estar elevada em casos de tempestade de citocinas.
- Ferritina: A ferritina sérica frequentemente aumenta em resposta à inflamação e pode estar elevada em casos graves de COVID-19 com tempestade de citocinas.
- Dímero-D: Este marcador está associado à coagulação sanguínea e pode estar elevado em casos de complicações tromboembólicas associadas à tempestade de citocinas.
- Linfócitos: A contagem de linfócitos, especialmente linfócitos T, pode estar reduzida durante a tempestade de citocinas devido à ativação excessiva do sistema imunológico ⁵³.

2.5.2 Metabolômica integrativa na pesquisa de biomarcadores contra a COVID-19

As técnicas de metabolômica, como LC-MS (do inglês, cromatografia líquida acoplada à espectrometria de massa), GC-MS (do inglês, cromatografia gasosa

acoplada à espectrometria de massa) e RMN (ressonância magnética nuclear), têm sido fundamentais na identificação de biomarcadores específicos para o diagnóstico e prognóstico de diversas doenças ⁵⁴.

Essas técnicas permitem a análise abrangente de metabólitos presentes em fluidos biológicos, tecidos ou células, fornecendo um perfil metabólico detalhado. LC-MS é altamente sensível e permite a identificação e quantificação de uma ampla gama de metabólitos em uma única amostra. Isso ajuda na descoberta de biomarcadores que podem ser úteis na identificação precoce de doenças, monitoramento de progressão e resposta ao tratamento ⁵⁵. GC-MS é especialmente útil na análise de metabólitos voláteis e termoestáveis ⁵⁶. Pode ser aplicado para identificar metabólitos específicos associados a certas condições patológicas, auxiliando no diagnóstico e na compreensão dos mecanismos subjacentes à doença ⁵⁶. Quanto ao RMN, embora seja menos sensível em comparação com LC-MS e GC-MS, a RMN fornece informações estruturais detalhadas sobre os metabólitos, o que pode ser crucial na identificação de biomarcadores específicos e na compreensão de vias metabólicas envolvidas em doenças ⁵⁷. A identificação de biomarcadores específicos através dessas técnicas não só pode facilitar o diagnóstico precoce, mas também ajudar na estratificação de pacientes, permitindo tratamentos mais personalizados e monitoramento da progressão da doença ^{55,56}.

Em relação a aplicação clínica na COVID-19, uma recente RSMA conduzida por Spick (2022) avaliou a acurácia de espectrometria de massas no diagnóstico de COVID-19, avaliando um total de 34 estudos com um total de 2.858 pacientes com COVID-19 e 2.544 controles. Os valores de sensibilidade e especificidade foram de 0,87(IC95%: 0,81 – 0,96) e 0,88 (IC95%: 0,82 – 0,98), respectivamente. No mesmo estudo, nas análises de subgrupo observou-se que os melhores resultados foram alcançados por análises proteômicas virais de *swabs* nasofaríngeos e análises metabolômicas de plasma e soro. O desempenho de outras matrizes de amostragem (respiração, sebo, saliva) foi pior, indicando que estes protocolos estão atualmente insuficientemente maduros para aplicação clínica ⁵⁸.

Em relação ao uso de LC-MS na COVID-19, Pang (2021) realizou uma análise abrangente *pathways* metanálise, utilizando sete bancos de dados distintos de experimentos de metabolômica em amostras de soro e plasma de pacientes com COVID-19 de três países distintos (Brasil, China e Estados Unidos) ⁵⁹. Essas amostras foram analisadas pela técnica de LC-MS e em seguida os dados foram analisados por

modelos de *machine learning* PCA (do inglês, análise de componentes principais) e PLS-DA (do inglês, partial least square discriminant analysis). Diversos biomarcadores foram identificados como sendo importantes na severidade e fatalidade da COVID-19, com mais destaque a beta-alanina, fenilalanina, vitamina B6, frutose, manose, tirosina, glutatona, cisteína, fenilalanina e metionina. *Pathway* metanálise revelou que as rotas metabólicas mais importantes na severidade e fatalidade da doença foram a rota de biossíntese da fenilalanina, tirosina e triptofano, e a biossíntese de aminoacil t-RNA, e glicólise e gliconeogênese, respectivamente ⁵⁹.

No estudo conduzido por Ruszkiewicz (2020), foram utilizadas amostras de hálito de pacientes diagnosticados com COVID-19 por meio de uma combinação de GC-MS e inteligência artificial. Os participantes do estudo eram provenientes de Dortmund (Alemanha) e Edinburgo (Escócia). A análise estatística multivariada por meio de *machine learning*, utilizando o modelo PCA, identificou aldeídos (como etanal e octanal), cetonas (incluindo acetona e butanona) e metanol como biomarcadores importantes na distinção da COVID-19 de outras condições de saúde ⁶⁰. Os resultados revelaram que a diferenciação entre pacientes com diagnóstico definitivo de COVID-19 e aqueles sem a doença foi alcançada com uma acurácia de 80% em Edimburgo e 81,5% em Dortmund, respectivamente. As taxas de sensibilidade e especificidade foram de 82,4% e 75%, respectivamente, enquanto a AUROC (Área sob a Curva ROC) atingiu 0,87 (IC 95%: 0,67 - 1,00), Já em Edimburgo, as taxas de sensibilidade e especificidade foram de 90% e 80%, respectivamente, com AUROC de 0,91 (IC 95%: 0,87 – 1,00) ⁶⁰.

2.5.3 Espectroscopia de infravermelho na detecção da COVID-19: uma abordagem inovadora

Os principais desafios dos testes de diagnóstico para COVID-19 são o custo e o tempo necessários para cada resultado de teste. O diagnóstico padrão-ouro por RT-PCR é caro, devido à escassez de instalações de teste, mesmo em países desenvolvidos, e pode levar mais de 2 dias para obter o resultado, porque as amostras devem ser transportadas para processamento para laboratórios frequentemente distantes ⁶¹, e isso não é adequado para testes em massa ⁶². Apesar do aumento global dos esforços de testes de RT-PCR, a pandemia não foi interrompida. Em contraste, há uma recorrência e uma segunda onda da doença porque muitos

pacientes infecciosos espalham a doença enquanto aguardam os resultados dos testes padrão ouro. Existem algumas empresas que desenvolvem testes mais rápidos e de baixo custo baseados em novos sensores ⁶³. Abordagens alternativas de detecção de antígenos ou anticorpos ainda não foram comprovadas, talvez criando novamente um viés estatístico que poderia afetar diretamente as políticas de saúde pública ⁶⁴. A sensibilidade destes testes é bem heterogênea, variando de 30% ⁶⁵ a 81% em diferentes ambientes ⁶⁶. Assim, os valores de sensibilidade e especificidade para testes de antígeno ainda precisam ser comprovados de forma robusta ⁶⁷. Assim, ainda há a necessidade não atendida de desenvolver abordagens de teste da COVID-19 que possam fornecer resultados em tempo real e no local ^{65,67,68}.

A espectroscopia vibracional, incluindo a espectroscopia de infravermelho com transformada de Fourier de reflexão total atenuada (ATR-FTIR), tem sido amplamente utilizada para discriminar e classificar populações normais e patológicas usando diferentes tipos de material biológico (de células, tecidos ou biofluidos) ⁶⁹⁻⁷¹. Biofluidos facilmente acessíveis, como sangue plasma/soro, saliva ou urina são considerados ideais para implementação clínica devido aos métodos rotineiros de coleta, bem como ao preparo mínimo de amostras ⁶⁹. O interrogatório de amostras com técnicas espectroscópicas de infravermelho (IR) permite a geração de uma “impressão digital espectral”, o que subseqüentemente facilita a discriminação das diferentes populações e a identificação de potenciais biomarcadores ⁷². Nos últimos anos, a espectroscopia ATR-FTIR baseada em biofluidos tem sido usada para diagnosticar, rastrear ou monitorar a progressão/regressão em uma variedade de doenças ⁷³. As técnicas espectroscópicas são rápidas, econômicas e não destrutivas, o que as torna candidatas perfeitas para aplicação para a clínica, mesmo como complemento de métodos mais estabelecidos ⁷⁴⁻⁷⁶.

Conforme relatado em publicações recentes, a espectroscopia IV seguida de análise quimiométrica tem sido utilizada com sucesso para identificar a infecção por SARS-CoV-2 em vários fluidos biológicos. Por exemplo, no estudo de Guleken (2022) ⁷⁵ foram analisadas amostras de soro de 49 pacientes com COVID-19 visando a diferenciar pacientes com COVID-19 com diferentes níveis de anticorpos (Imunoglobulina G e M - IgG e IgM), usando FTIR e espectroscopia Raman. Os dados espectroscópicos foram analisados por análise multivariada, *machine learning* e métodos de redes neurais. Foi demonstrado que a análise do soro utilizando FTIR e espectroscopia Raman permitiu diferenciar os níveis de anticorpos entre 1 e 6 meses

através de biomarcadores espectrais de amidas II e I. Além disso, a análise multivariada mostrou que utilizar a espectroscopia Raman na faixa entre 1.317 cm^{-1} e 1.432 cm^{-1} , 2.840 cm^{-1} e 2.956 cm^{-1} permite distinguir pacientes após 1, 3 e 6 meses de COVID-19 com sensibilidade próxima a 100% ⁷⁵.

2.6 ABORDAGENS NÃO FARMACOLÓGICAS NO TRATAMENTO DE PACIENTES COM COVID-19

2.6.1. Alimentação e nutrição como abordagem complementar na recuperação dos pacientes COVID-19

A alimentação e a nutrição desempenham um papel crucial na recuperação de todo e qualquer tipo de doença, incluindo na COVID-19, uma vez que fornecem os nutrientes necessários para fortalecer o sistema imunológico, promover a cicatrização e melhorar a resposta do corpo à infecção viral ^{77,78}. Uma dieta equilibrada, rica em vitaminas, minerais e antioxidantes, pode ajudar a reduzir a gravidade dos sintomas, acelerar a recuperação e diminuir o risco de complicações ^{77,78}. Além disso, uma alimentação adequada pode ajudar a combater a inflamação, controlar o peso corporal e manter a saúde geral, aspectos essenciais durante o processo de recuperação da COVID-19 ⁷⁸.

Por outro lado, a pandemia de COVID-19 teve um enorme impacto na saúde, social e económico em todo o mundo durante os últimos anos ⁷⁹. O desfecho clínico é pior em vários grupos de pacientes em risco de desnutrição ^{79,80}. Segundo a ESPEN (*The European Society for Clinical Nutrition and Metabolism*), apesar da relevância da investigação nutricional na COVID-19, até dezembro de 2020 as evidências disponíveis eram extremamente limitadas ^{79,80}. Das mais de 75.000 publicações sobre COVID-19 disponíveis no PubMed em 2020, apenas 1.200 foram recuperadas ao adicionar nutrição na pesquisa, e apenas 169 quando desnutrição é usada como palavra-chave ⁸⁰.

Portanto, em 2021, o reforço das evidências sobre o potencial papel fundamental do estado nutricional e dos cuidados nutricionais na COVID-19 era, portanto, uma necessidade não satisfeita e uma prioridade urgente ⁸⁰.

A ESPEN e o Escritório Regional da Organização Mundial de Saúde para a Europa (OMS/Europa) discutiram este tema e concordaram em abrir uma chamada para submissão de artigos científicos sobre o estado nutricional e os cuidados nutricionais em doentes com COVID-19, a serem publicados nas revistas *ESPEN Clinical Nutrition* e *Clinical Nutrition ESPEN* ⁸⁰. A teleconferência foi editada por convidados do ESPEN e de representantes da OMS/Europa.

Atualmente, as evidências mostram que existem certos grupos populacionais com maior vulnerabilidade à doença, especialmente pacientes com condições subjacentes, como doenças crônicas não transmissíveis (por exemplo, hipertensão, diabetes tipo 2, doença cardíaca isquêmica, doença pulmonar obstrutiva crônica e cancro) ⁸¹. A gravidade da infecção por COVID-19, piores desfechos e mortalidade têm sido associados à idade, comorbidades e alto índice de massa corporal em diferentes países ⁸²⁻⁸⁶. É importante destacar que grande parte das doenças crônicas não transmissíveis citadas estão causalmente relacionadas ao consumo alimentar e às características do estilo de vida. Dietas ricas em carboidratos, gorduras saturadas e açúcares refinados contribuem para a prevalência da obesidade e diabetes tipo 2 e podem aumentar o risco de COVID-19 grave e mortalidade ⁸⁷. O alto consumo de carboidratos e gorduras saturadas está diretamente associado a quadros inflamatórios, que podem comprometer a saúde na prevenção e superação de infecções. Por outro lado, um pior estado nutricional está associado ao estresse oxidativo, que afeta negativamente o sistema imunológico. Portanto, uma dieta saudável e balanceada é essencial para a produção adequada de anticorpos e minimização do estresse oxidativo e do estado inflamatório, principalmente no que diz respeito à quantidade de proteína suficiente para promover uma resposta imune adequada, ajudando no combate a COVID-19 ^{88,89}.

2.7 ABORDAGENS FARMACOLÓGICAS NO TRATAMENTO DE PACIENTES COM COVID-19

2.7.1 Planejamento racional de fármacos como uma abordagem promissora na descoberta de novos medicamentos

Descobrir um novo medicamento é um processo multifacetado com três estratégias principais: a descoberta aleatória ⁹⁰, a triagem de alto rendimento ⁹¹ e o planejamento racional ⁹⁰⁻⁹². A descoberta aleatória, também conhecida como serendipidade, ocorre quando medicamentos são descobertos acidentalmente, muitas vezes durante a pesquisa de outras substâncias ⁹⁰. Um exemplo icônico é a descoberta da penicilina por Alexander Fleming em 1928 ⁹⁰. A triagem de alto rendimento envolve a avaliação de grandes bibliotecas de compostos químicos para identificar aqueles com atividade contra um alvo específico relacionado à doença ⁹¹. Embora eficiente, essa abordagem tem desafios, incluindo custos elevados, falsos positivos/negativos e limitações nos ensaios biológicos ⁹¹. Por fim, o planejamento racional usa conhecimento prévio sobre a biologia da doença para projetar moléculas que possam interagir com alvos específicos de forma mais direcionada ⁹².

Embora a descoberta aleatória e a triagem de alto rendimento sejam amplamente utilizadas, o planejamento racional de fármacos emerge como a abordagem mais promissora ⁹². Enquanto a descoberta aleatória depende da sorte e a triagem de alto rendimento enfrenta desafios como custo e falsos resultados, o planejamento racional usa conhecimento prévio para projetar moléculas com interações direcionadas, oferecendo maior eficácia e segurança potencial ⁹⁰⁻⁹³.

CADD (*Computer-Aided Drug Design*), ou *design* de medicamentos assistido por computador é uma estratégia que utiliza o conhecimento científico existente sobre a biologia de uma doença para identificar alvos terapêuticos específicos. Os cientistas identificam uma proteína, um gene ou um processo biológico associado à doença e procuram desenvolver um medicamento que possa interferir ou modular esse alvo de forma benéfica. Isso frequentemente envolve estudos detalhados de biologia molecular, genômica, e outros campos relacionados ⁹².

Considerando que o processo de descoberta e desenvolvimento de um novo fármaco é muito caro e demorado, que em geral dura em torno de 15 anos (**Figura 1**), a utilização de CADD, ou *design* de fármacos assistido por computador, oferece diversas vantagens na descoberta de novos medicamentos ⁹², como evidenciado a seguir:

- Economia de tempo e recursos: O CADD permite a triagem virtual de milhões de compostos em um curto período, poupando tempo e recursos que seriam gastos na síntese e teste de compostos físicos.

- Precisão na seleção de compostos: Os algoritmos utilizados no CADD são capazes de prever propriedades químicas, interações moleculares e estruturas tridimensionais, ajudando a identificar compostos mais promissores para estudos posteriores.
- Redução de falhas: Ao eliminar compostos menos propensos a serem eficazes ou seguros, o CADD reduz a probabilidade de falha em fases posteriores de desenvolvimento de medicamentos.
- Personalização e otimização: O CADD permite a modificação estrutural de compostos existentes para otimizar a atividade farmacológica ou minimizar efeitos colaterais, tornando o processo de desenvolvimento mais eficiente.
- Menor dependência de testes em animais: Ao reduzir a quantidade de compostos físicos que precisam ser testados em laboratório, o CADD pode contribuir para diminuir a necessidade de testes em animais, alinhando-se com práticas mais éticas.
- Ampla aplicação na pesquisa: O CADD pode ser aplicado em diversas áreas, desde a descoberta de novos medicamentos até o *design* de moléculas para terapias específicas, abrindo portas para uma variedade de aplicações na área da saúde.

Entre as diferentes técnicas de CADD incluem, destacam-se:

- (i) Docking Molecular: Este método simula a interação entre uma molécula (geralmente um potencial fármaco) e sua proteína-alvo. Ele prediz como e onde a molécula se encaixaria na proteína para formar um complexo estável. O objetivo é identificar moléculas que se encaixem de forma ideal na estrutura da proteína alvo, inibindo ou ativando sua função ⁹⁴.
- (ii) Relação quantitativa estrutura-atividade (QSAR): Este método analisa a relação entre a estrutura química de compostos e suas atividades biológicas. Usando dados experimentais, modelos estatísticos são criados para prever a atividade de compostos similares com base em sua estrutura química ⁹⁵.
- (iii) Simulações de Dinâmica Molecular: Esse método simula o movimento e a interação dos átomos em uma molécula ao longo do tempo. Ele fornece informações detalhadas sobre a dinâmica molecular, permitindo a

compreensão das propriedades estruturais e funcionais das macromoléculas biológicas ⁹⁶.

- (iv) Predição de propriedades farmacocinéticas e toxicidade: modelos computacionais são empregados para prever propriedades farmacocinéticas, como absorção, distribuição, metabolismo e excreção (ADME), bem como para prever a toxicidade potencial de compostos candidatos ⁹⁷.

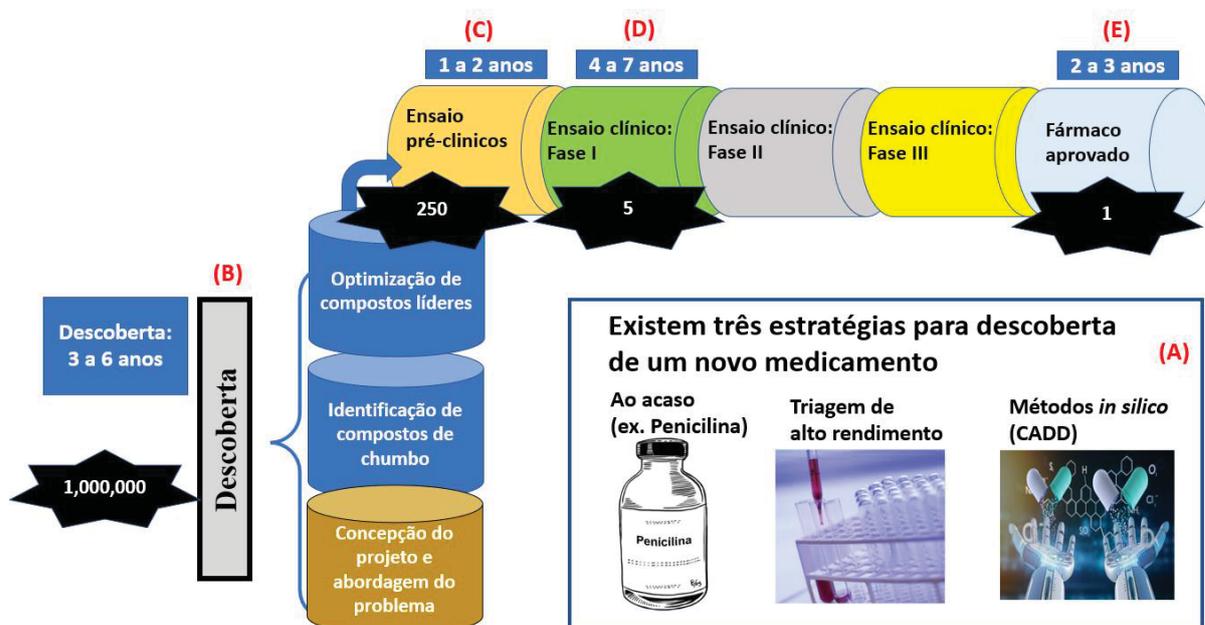


Figura 1. Fluxo do desenvolvimento de um novo fármaco.

Nota: Em (A) é mostrado as estratégias de descoberta de um novo fármaco. Em (B) é a etapa de descoberta de fármaco, que dura cerca entre 3-6 anos. Nesta etapa inclui desde a concepção do projeto e o problema a ser investigado, identificação e otimização dos compostos líderes. Em média em total de um milhão de moléculas são analisadas. Em (C), é a etapa dos testes *in vitro* e *in vivo* por exemplo, em roedores. Em (D) é a etapa dos ensaios clínicos em seres humanos, visando avaliar a eficácia, potência e segurança. Em (E) é a etapa de aprovação do medicamento pelas agências reguladoras. Dos cerca de 1 milhão de moléculas testadas na etapa de descoberta, apenas 1 molécula é aprovada para incorporação na terapêutica. Em média desde a etapa de descoberta até a aprovação dura 15 anos. Fonte: O Autor (2024)

Em relação ao uso de CADD na COVID-19, diversos estudos usando abordagem CADD estão disponíveis na literatura, e existem diversas RS comprovando a eficiência desses métodos na descoberta de fármacos contra COVID-

19⁹⁸⁻¹⁰⁰. Em um recente estudo de Hasan (2022), diversos modelos de CADD (docking e dinâmica molecular e predição ADMET) foram empregadas para testar um conjunto de 305 compostos provenientes de fontes naturais contra as cinco principais proteínas alvos envolvidas em vários estágios do ciclo de vida do SARS-CoV-2, ou seja, ligações virais (ACE2 e TMPRSS2), replicação viral e transcrição (Mpro, PLpro e RdRp). Nesta estudo, o fitoquímico vitanolide mostrou resultados promissores como candidato a fármaco para o tratamento da COVID-19. No estudo de Prajapat (2020), usando os mesmos métodos de CADD, foi capaz de identificar alguns antidiabéticos (acarbose), vitaminas (riboflavina e ácido levomefólico), agentes antiplaquetários (cangrelor), antibióticos aminoglicosídeos (canamicina, amicacina), broncodilatador (fenoterol), imunomodulador (lamivudina), e agentes antineoplásicos (mitoxantrona e vidarabina), que se mostraram promissores como candidatos a fármacos para o tratamento da COVID-19, via inibição da proteína Spike (S1) do NSARS-CoV-2¹⁰¹.

2.7.2 Inteligência artificial e machine learning na descoberta de novos fármacos

Nos últimos anos, houve um aumento drástico na digitalização de dados no setor farmacêutico. No entanto, esta digitalização traz consigo o desafio de adquirir, examinar e aplicar esse conhecimento para resolver problemas clínicos complexos¹⁰². Isso motiva o uso da IA, pois ela pode lidar com grandes volumes de dados com automação aprimorada¹⁰³. A IA é um sistema baseado em tecnologia que envolve várias ferramentas e redes avançadas que podem imitar a inteligência humana. Ao mesmo tempo, não ameaça substituir completamente a presença física humana^{104,105}. A IA utiliza sistemas e *software* que podem interpretar e aprender com os dados de entrada para tomar decisões independentes para atingir objetivos específicos. Suas aplicações estão sendo continuamente ampliadas na área farmacêutica. De acordo com o McKinsey Global Institute, os rápidos avanços na automação guiada por IA provavelmente mudarão completamente a cultura de trabalho da sociedade^{106,107}.

O vasto espaço químico, compreendendo mais de 10 moléculas, promove o desenvolvimento de muitas moléculas de medicamentos¹⁰⁸. No entanto, a falta de tecnologias avançadas limita o processo de desenvolvimento de medicamentos, tornando-o uma tarefa demorada e dispendiosa, que pode ser resolvida através da utilização de IA. A IA pode reconhecer protótipos (*hit*) e compostos líderes e fornecer uma validação mais rápida do alvo do medicamento e otimização do projeto da

estrutura do medicamento ^{108,109}. Diferentes aplicações de IA na descoberta de medicamentos são representadas na **Figura 2** ¹¹⁰.

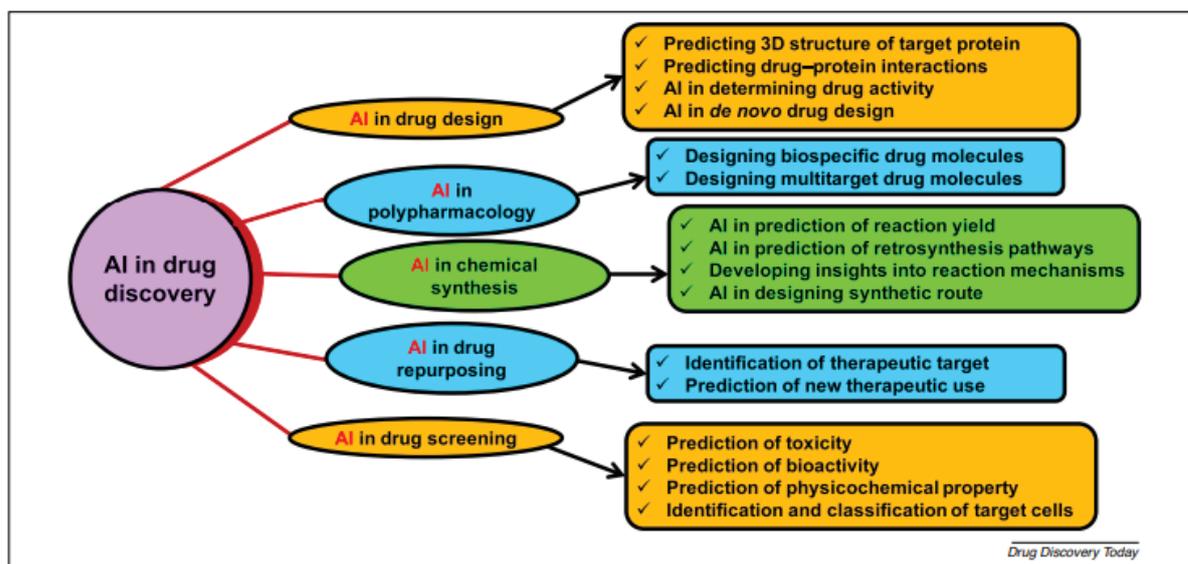


Figura 2. Papel da inteligência artificial (IA) na descoberta de medicamentos¹¹⁰

Apesar das suas vantagens, a IA enfrenta alguns desafios de dados significativos, tais como a escala, o crescimento, a diversidade e a incerteza dos dados. Os conjuntos de dados disponíveis para o desenvolvimento de medicamentos nas empresas farmacêuticas podem envolver milhões de compostos, e as ferramentas tradicionais de ML podem não ser capazes de lidar com estes tipos de dados ^{111,112}. O modelo computacional baseado na QSAR pode prever rapidamente muitos compostos ou parâmetros físico-químicos simples, como logP ou logD. No entanto, esses modelos estão um pouco distantes das previsões de propriedades biológicas complexas, como a eficácia e efeitos adversos dos compostos. Além disso, os modelos baseados em QSAR também enfrentam problemas como pequenos conjuntos de treinamento, erros de dados experimentais em conjuntos de treinamento e falta de validações experimentais. Para superar esses desafios, abordagens de IA desenvolvidas recentemente, como *Deep Learning* (DL) e estudos de modelagem relevantes, podem ser implementadas para avaliações de segurança e eficácia de moléculas de medicamentos com base em modelagem e análise de *big data*. Em 2012, a Merck apoiou um desafio QSAR ML para observar as vantagens da DL no processo de descoberta de medicamentos na indústria farmacêutica. Os modelos DL mostraram previsibilidade significativa em comparação com abordagens tradicionais

de ML para 15 conjuntos de dados de absorção, distribuição, metabolismo, excreção e toxicidade (ADMET) de candidatos a medicamentos ^{111,112}.

O espaço químico virtual é enorme e sugere um mapa geográfico de moléculas, ilustrando as distribuições das moléculas e suas propriedades. A ideia por trás da ilustração do espaço químico é coletar informações posicionais sobre moléculas dentro do espaço para procurar compostos bioativos e, assim, a triagem virtual (VS) ajuda a selecionar moléculas apropriadas para testes adicionais. Vários espaços químicos são de acesso aberto, incluindo PubChem, ChemBank, DrugBank e ChemDB ¹¹³.

Numerosos métodos *in silico* para selecionar compostos virtuais a partir de espaços químicos virtuais, juntamente com abordagens baseadas na estrutura do ligantes (LBDD), fornecem uma melhor análise de perfil, eliminação mais rápida de compostos não-líderes e seleção de moléculas de fármacos, com gastos reduzidos [19]. Algoritmos de projeto de medicamentos, como matrizes de Coulomb e reconhecimento de impressão digital molecular, consideram os perfis físicos, químicos e toxicológicos para selecionar um composto líder ¹¹³.

Vários parâmetros, como modelos preditivos, a similaridade de moléculas, o processo de geração de moléculas e a aplicação de abordagens *in silico* podem ser usados para prever a estrutura química desejada de um composto ¹¹⁴. Pereira et al. apresentou um novo sistema, DeepVS, para o docking de 40 receptores e 2.950 ligantes, que apresentou desempenho excepcional quando 95.000 compostos candidatos foram testados contra esses receptores ¹¹⁵. Outra abordagem aplicou um algoritmo de substituição automatizado multiobjetivo para otimizar o perfil de potência de um inibidor de quinase-2 dependente de ciclina, avaliando sua similaridade estrutural, atividade bioquímica e propriedades físico-químicas ¹¹⁶.

As ferramentas de modelagem QSAR têm sido utilizadas para a identificação de potenciais candidatos a medicamentos e evoluíram para abordagens QSAR baseadas em IA, como *Linear Discriminant Analysis (LDA)*, *Support Vector Machines (SVMs)*, *Random Forest (RF)* e *Decision Tree (DT)*, que podem ser aplicados para acelerar a análise QSAR ^{117,118}. King (1995) encontrou uma diferença estatística insignificante quando a capacidade de seis algoritmos de IA para classificar compostos anônimos em termos de atividade biológica foi comparada com a das abordagens tradicionais ¹¹⁹.

2.7.3 Inteligência artificial na triagem de novos fármacos

O processo de descoberta e desenvolvimento de um medicamento pode levar mais de uma década e custa, em média, 2,8 mil milhões de dólares. Mesmo assim, nove em cada dez moléculas terapêuticas falham nos ensaios clínicos de fase II e na aprovação por parte das entidades reguladoras ¹²⁰. Algoritmos, como classificadores *K-Nearest Neighbors (KNN)*, RF, máquinas de aprendizagem extrema, SVM e *Deep Neural Networks (DNNs)*, são usados para VS com base na viabilidade de síntese e podem prever atividade e toxicidade *in vivo* ¹²¹. Várias empresas biofarmacêuticas, como *Bayer*, *Roche* e *Pfizer*, uniram-se a empresas de Tecnologia de Informação (TI) para desenvolver uma plataforma para a descoberta de terapias em áreas como imuno-oncologia e doenças cardiovasculares ¹²². Os aspectos do VS aos quais a IA foi aplicada são discutidos abaixo.

2.7.4 Inteligência artificial na previsão das propriedades físico-químicas

Propriedades físico-químicas, como solubilidade, coeficiente de partição (logP), grau de ionização e permeabilidade intrínseca do fármaco, afetam indiretamente suas propriedades farmacocinéticas e sua família de receptores alvo e, portanto, devem ser consideradas ao projetar um novo fármaco ¹²³. Diferentes ferramentas baseadas em IA podem ser usadas para prever propriedades físico-químicas. Por exemplo, ML usa grandes conjuntos de dados produzidos durante a otimização composta feita anteriormente para treinar o programa ¹²⁴. Algoritmos para projeto de medicamentos incluem descritores moleculares, como strings SMILES, medições de energia potencial, densidade eletrônica ao redor da molécula e coordenadas de átomos em 3D, para gerar moléculas viáveis via algoritmo DNN e, assim, prever suas propriedades ¹²⁵.

Zang et al. criou um fluxo de trabalho de relação quantitativa estrutura-propriedade (QSPR) para determinar as seis propriedades físico-químicas de produtos químicos ambientais obtidos da Environmental Protection Agency (EPA), chamado Suíte de Interface de Programa de Estimativa (EPI) ¹²⁵. Redes neurais baseadas no preditor ADMET e no programa ALGOPS têm sido usadas para prever

a lipofilicidade e solubilidade de vários compostos ¹²⁶. Métodos Deep Learning (DL), como UnDirected Graph Recursive Neural Networks e Convolutional Variational Neural Networks (CVNN), têm sido usados para prever a solubilidade de moléculas ¹²⁶.

Kumar et al. desenvolveu seis modelos preditivos (SVMs, ANNs, KNN, LDAs, algoritmos de redes neurais probabilísticas e mínimos quadrados parciais - PLS) baseados em abordagem QSPR utilizando 745 compostos para treinamento; estes foram usados posteriormente em 497 compostos para prever sua absorvidade intestinal com base em parâmetros incluindo área de superfície molecular, massa molecular, contagem total de hidrogênio, refratividade molecular, volume molecular, logP, área de superfície polar total, soma dos índices de estados eletrônico de energia mais baixa, índice de solubilidade (log S) e ligações rotativas ¹²⁷. Na mesma linha, modelos *in silico* baseados em RF e DNN foram desenvolvidos para determinar a absorção intestinal humana de uma variedade de compostos químicos ¹²⁸. Assim, a IA tem um papel significativo no desenvolvimento de um medicamento, para prever não só as suas propriedades físico-químicas desejadas, mas também a bioatividade desejada ¹¹⁹.

2.7.5 Inteligência artificial na predição da biodisponibilidade

A eficácia das moléculas do fármaco depende da sua afinidade pela proteína ou receptor alvo. Moléculas de fármaco que não apresentam qualquer interação ou afinidade com a proteína alvo não serão capazes de fornecer a resposta terapêutica. Em alguns casos, também pode ser possível que moléculas de medicamentos desenvolvidas interajam com proteínas ou receptores não intencionais, levando à toxicidade. Conseqüentemente, a afinidade de ligação ao alvo do medicamento (DTBA) é vital para prever as interações medicamento-alvo. Os métodos baseados em IA podem medir a afinidade de ligação de um medicamento considerando as características ou semelhanças do medicamento e do seu alvo. As interações baseadas em características reconhecem as partes químicas da droga e do alvo para determinar os vetores de características. Por outro lado, na interação baseada em similaridade, a semelhança entre o medicamento e o alvo é considerada, e presume-se que medicamentos semelhantes irão interagir com os mesmos alvos ¹²⁹.

Aplicativos Web, como ChemMapper e a abordagem de conjunto de similaridade (SEA), estão disponíveis para prever interações medicamento-alvo ¹³⁰. Muitas estratégias envolvendo ML e DL têm sido utilizadas para determinar a afinidade do medicamento pela proteína alvo, como KronRLS, SimBoost, DeepDTA e PADME. Abordagens baseadas em ML, como os mínimos quadrados regulados por Kronecker (KronRLS), avaliam a similaridade entre medicamentos e moléculas de proteína para determinar a afinidade do medicamento pela proteína alvo. Da mesma forma, o SimBoost utilizou modelos de regressão de árvore de decisão para prever a afinidade do medicamento pela proteína alvo e considera interações baseadas em features (descritores moleculares) e em similaridade. Características de drogas do SMILES, subestrutura comum máxima do ligante (LMCS), impressão digital de conectividade estendida ou uma combinação delas também podem ser consideradas ¹²⁹.

DeepAffinity e Protein And Drug Molecule interaction prEdiction (PADME) são semelhantes às abordagens descritas anteriormente ¹²⁹. DeepAffinity é um modelo DL interpretável que usa RNN e CNN e dados rotulados e não rotulados. Considera o composto no formato SMILES e sequências proteicas nas propriedades estruturais e físico-químicas ¹³¹. PADME é uma plataforma baseada em DL que utiliza redes neurais *feed-forward* para prever interações entre medicamentos e alvos (DTIs). Considera a combinação das características do medicamento e da proteína alvo como dados de entrada e prevê a força da interação entre os dois. Para o fármaco e o alvo, a representação SMILES e a composição da sequência proteica (PSC) são utilizadas para ilustração, respectivamente ¹³². Técnicas de ML não supervisionadas, como MANTRA e PREDICT, podem ser usadas para prever a eficácia terapêutica de medicamentos e proteínas alvo de produtos farmacêuticos conhecidos e desconhecidos, o que também pode ser extrapolado para a aplicação de reposicionamento de medicamentos e interpretação do mecanismo molecular da terapêutica. MANTRA agrupa compostos com base em perfis de expressão gênica semelhantes usando um conjunto de dados CMap e agrupa os compostos previstos como tendo um mecanismo de ação comum e uma via biológica comum ¹³⁰. A bioatividade de um medicamento também inclui dados da ADME. Ferramentas baseadas em IA, como XenoSite, FAME e SMARTCyp, estão envolvidas na determinação dos pontos de metabolismo do medicamento. Além disso, *softwares* como CypRules, MetaSite, MetaPred, SMARTCyp e WhichCyp foram usados para identificar isoformas específicas do CYP450 que medeiam o metabolismo de um

medicamento específico. A via de depuração de 141 medicamentos aprovados foi realizada por preditores baseados em SVM com alta precisão ¹³³.

2.7.6 Inteligência artificial na predição da toxicidade

A previsão da toxicidade de qualquer molécula de medicamento é vital para evitar efeitos tóxicos. Ensaios *in vitro* baseados em células são frequentemente usados como estudos preliminares, seguidos de estudos em animais para identificar a toxicidade de um composto, aumentando o custo da descoberta de medicamentos. Várias ferramentas baseadas na web, como LimTox, pkCSM, admetSAR e Toxtree, estão disponíveis para ajudar a reduzir o custo ¹³⁴. Abordagens avançadas baseadas em IA procuram semelhanças entre compostos ou projetam a toxicidade do composto com base nas características de entrada. O Tox21 Data Challenge organizado pelos Institutos Nacionais de Saúde, EPA e Food and Drug Administration (FDA) dos EUA foi uma iniciativa para avaliar diversas técnicas computacionais para prever a toxicidade de 12.707 compostos ambientais e medicamentos ¹³⁴; um algoritmo de ML chamado DeepTox superou todos os métodos, identificando características estáticas e dinâmicas dentro dos descritores químicos das moléculas, como peso molecular (PM) e volume de Van der Waals, e poderia prever com eficiência a toxicidade de uma molécula com base em características predefinidas de 2.500 toxicóforos ¹³⁵.

A SEA foi usada para avaliar a previsão da meta de segurança de 656 medicamentos comercializados contra 73 alvos não intencionais que podem produzir efeitos adversos ¹³⁰. Desenvolvido usando uma abordagem baseada em ML, o eToxPred foi aplicado para estimar a toxicidade e a viabilidade de síntese de pequenas moléculas orgânicas e mostrou precisão de até 72% ¹³³. Da mesma forma, ferramentas de código aberto, como TargeTox e ProCTOR, também são usadas na previsão de toxicidade ¹³³. TargeTox é um método de previsão de risco de toxicidade de medicamentos baseado em alvo de rede biológica que usa o princípio de culpa por associação, pelo qual entidades que possuem propriedades funcionais semelhantes compartilham semelhanças em redes biológicas ¹³⁶. Ele pode produzir dados de redes de proteínas e unir propriedades farmacológicas e funcionais em um classificador ML para prever a toxicidade do medicamento ¹³⁷. O ProCTOR foi treinado usando um modelo de RF e levou em consideração propriedades de probabilidade de medicamento, características moleculares, características baseadas em alvos e

propriedades dos alvos proteicos para gerar uma 'pontuação ProCTOR', que previu se um medicamento falharia em ensaios clínicos devido a sua toxicidade. Também reconheceu medicamentos aprovados pela FDA que posteriormente relataram eventos adversos a medicamentos ¹³⁸. Em outra abordagem, o Tox_(R)CNN envolvendo um método CVNN profundo avaliou a citotoxicidade de medicamentos que foram expostos a células coradas com DAPI ¹³⁹.

1 CAPÍTULO I - FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 E O IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO AUMENTO DE INTERNAÇÕES HOSPITALARES

Publicado em:

1. Cobre AF, Böger B, Fachi MM, Vilhena RO, Domingos EL, Tonin FS, Pontarolo R. Risk factors associated with delay in diagnosis and mortality in patients with COVID-19 in the city of Rio de Janeiro, Brazil. *Cien Saude Colet*. 2020 Oct;25(suppl 2):4131-4140. doi: 10.1590/1413-812320202510.2.26882020.
2. Cobre AF, Böger B, Vilhena RO, Fachi MM, Dos Santos JMMF, Tonin FS. A multivariate analysis of risk factors associated with death by Covid-19 in the USA, Italy, Spain, and Germany. *Z Gesundh Wiss*. 2022;30(5):1189-1195. doi: 10.1007/s10389-020-01397-7.
3. Cobre AF, Stremel DP, Böger B, Fachi MM, Borba HHL, Tonin FS, Sarti FM, Pontarolo R. The impact of COVID-19 vaccine rejection on hospital admission and variants spread worldwide: implications for healthcare policy. *Research, Society and Development*, v. 11, n. 11, p. e189111133435-e189111133435, 2022. doi: <https://doi.org/10.33448/rsd-v11i11.33434>.

CAPÍTULO I - FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 E O IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO AUMENTO DE INTERNAÇÕES HOSPITALARES

1.1 RESUMO

Este primeiro capítulo da tese de doutorado apresenta descobertas cruciais realizadas no início da pandemia de COVID-19, em 2020, quando o conhecimento sobre os fatores de risco associados à mortalidade pela doença era escasso. Por meio de três investigações distintas, foram abordados aspectos fundamentais para o entendimento e enfrentamento da crise sanitária. Inicialmente, uma análise realizada no Rio de Janeiro-Brasil, investigou os atrasos no diagnóstico e sua relação com a mortalidade, revelando associações significativas entre o desfecho de interesse e variáveis independentes como o tempo de diagnóstico, sexo masculino, faixas etárias mais jovens e residência em áreas com índices de desenvolvimento social mais baixos. Em seguida, uma pesquisa multinacional, abrangendo os Estados Unidos, Itália, Espanha e Alemanha, destacou a necessidade premente de infraestrutura hospitalar e dispositivos de ventilação para mitigar as taxas de mortalidade por COVID-19. Por fim, um modelo preditivo baseado em regressão Poisson identificou o impacto da rejeição à vacinação nas internações em unidades de terapia intensiva (UTIs) e na propagação de variantes do SARS-CoV-2, enfatizando a urgência das campanhas de vacinação para combater os efeitos devastadores da pandemia em escala global. Essas descobertas ressaltam a importância histórica deste capítulo, que contribuiu significativamente para a compreensão inicial da pandemia e para a formulação de estratégias eficazes de intervenção.

Palavras-chave: COVID-19, fatores de risco, mortalidade

1.2 INTRODUÇÃO

No início da pandemia de COVID-19, em 2020, enfrentamos um desafio sem precedentes, com a rápida disseminação do vírus e um conhecimento limitado sobre seus efeitos e fatores de risco associados ¹⁴⁰. Nesse período inicial, havia uma lacuna significativa no entendimento dos elementos que contribuíam para a mortalidade relacionada ao COVID-19, tornando a formulação de estratégias de prevenção e tratamento um verdadeiro dilema ¹⁸. Em meio a essa incerteza, este capítulo se tornou crucial, pois buscou explorar e investigar os fatores de risco potenciais para a mortalidade por COVID-19, fornecendo informações valiosas em um momento em que a informação era escassa. Embora hoje tenhamos um entendimento mais robusto dos fatores de risco envolvidos ^{141,142}, é importante reconhecer a relevância e o impacto deste capítulo durante os estágios iniciais da pandemia, quando as evidências eram limitadas e a urgência em compreender e enfrentar o vírus era imensa. Assim, este capítulo não apenas contribuiu para o acúmulo de conhecimento essencial, mas também desempenhou um papel fundamental no combate à pandemia, destacando a importância de investigações e pesquisas contínuas para enfrentar desafios de saúde pública emergentes.

Este primeiro capítulo da tese de doutorado compreende três estudos científicos independentes. O primeiro estudo investiga os fatores de risco associados ao atraso no diagnóstico e à mortalidade por COVID-19 em nível nacional (Brasil). O segundo estudo foca na análise dos fatores de risco relacionados ao atraso no diagnóstico e à mortalidade por COVID-19 em escala internacional, abrangendo os Estados Unidos, Alemanha, Espanha e Itália. O terceiro e último estudo tem como objetivo avaliar o impacto da rejeição à vacinação no aumento do número de internações hospitalares, tanto em nível nacional quanto internacional. Portanto, a seção de materiais e métodos, resultados e discussão foi subdividida em três subseções, correspondentes aos três estudos conduzidos.

1.3 OBJETIVOS

1.3.1 *Objetivo geral*

- Analisar e identificar os fatores determinantes relacionados à mortalidade pela COVID-19, investigando paralelamente o impacto direto da resistência à vacinação na elevação das internações hospitalares, visando contribuir para estratégias eficazes de prevenção e controle da doença.

1.3.2 *Objetivos específicos*

- Identificar e analisar os fatores de risco associados à mortalidade causada pela COVID-19 em contextos nacionais e internacionais, com foco na idade, gênero, condições socioeconômicas e acesso a serviços de saúde.
- Avaliar o período de diagnóstico e sua relação com as taxas de mortalidade, investigando a influência do atraso no diagnóstico na progressão da doença e na letalidade.
- Analisar a relação entre a aceitação da vacinação contra a COVID-19 e o aumento das internações hospitalares, especificamente em contextos em que a resistência à vacinação se tornou um obstáculo significativo para o controle da doença.
- Explorar estratégias para mitigar a rejeição à vacinação, como campanhas educativas, políticas de obrigatoriedade vacinal, restrições de acesso a serviços públicos e sanções para disseminadores de informações anti-vacinação.
- Propor recomendações e diretrizes para ações efetivas de saúde pública, visando reduzir a mortalidade pela COVID-19, incluindo medidas de prevenção, fortalecimento de infraestruturas hospitalares e promoção da aceitação da vacinação como estratégia fundamental de controle da pandemia.

1.4 MATERIAL E MÉTODOS

1.4.1 ESTUDO I: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS AO ATRASO NO DIAGNÓSTICO E MORTALIDADE POR COVID-19 NO ÂMBITO NACIONAL

1.4.1.1 Desenho do estudo e coleta de dados

Este estudo de coorte retrospectivo incluiu pacientes com diagnóstico de COVID-19 no Rio de Janeiro, localizado na região sudoeste do Brasil, o que representa aproximadamente 3,0% da população brasileira. Os dados foram obtidos no banco de dados público e de acesso aberto chamado Painel Rio COVID-19, que é mantido pela Prefeitura do Rio de Janeiro.

Foram incluídos todos os pacientes (independentemente de idade ou sexo) com diagnóstico positivo para COVID-19 no período de 27 de fevereiro de 2020 a 26 de abril de 2020. O índice de desenvolvimento social (IDS) dos pacientes foi classificado de acordo com os endereços do Rio de Janeiro em dois grupos: baixo índice social e alto índice social. O tempo para diagnóstico (em dias) foi estimado pela diferença entre a data do início dos primeiros sintomas da COVID-19 e a data do diagnóstico. Para análise dos preditores de atraso no diagnóstico, as variáveis independentes foram faixa etária, sexo, IDS e evolução da doença. A variável dependente foi o tempo até o diagnóstico. Na análise dos fatores de risco associados à mortalidade por COVID-19, as variáveis independentes foram faixa etária, sexo, IDS e tempo até o diagnóstico. Neste caso, a variável dependente, mortalidade, foi binária e categorizada como 'morto' ou 'vivo'. Os pacientes com doença ativa foram considerados vivos.

1.4.1.2 Análise estatística

Foi realizada análise descritiva utilizando mediana e intervalo interquartil (IIQ) para variáveis contínuas com distribuição de probabilidade não normal, e frequência para variáveis categóricas.

O método Kaplan-Meier foi adotado para estimar o atraso no diagnóstico. O teste log-rank foi utilizado para comparar as curvas de sobrevida de Kaplan-Meier das

variáveis independentes ¹⁴³. Análise univariada e multivariada de regressão de Cox covariável tempo-dependente foram utilizados para investigar os fatores de risco associados ao atraso no diagnóstico. Os *Hazard Ratio* (HR) e os seus respectivos intervalos de confiança de 95% foram utilizados para quantificar o tamanho dos efeitos dos fatores de risco ¹⁴⁴.

Em relação aos dados de mortalidade, o objetivo inicial foi investigar os fatores prognósticos para o tempo até o óbito por COVID-19 utilizando o método de análise de sobrevivência por regressão de Cox, porém, dada a falta de informações sobre o tempo de seguimento dos pacientes (por exemplo, data do diagnóstico da doença até óbito, cura ou perda de seguimento) no banco de dados original, não foi possível utilizar o método de análise de sobrevivência. Portanto, para avaliar os fatores de risco de mortalidade por COVID-19, foi realizada análise univariada utilizando a função Custom Table do *Software* SPSS (IBM, EUA), que foi usada como uma fase preparatória da análise multivariada, que permitiu identificar possíveis variáveis independentes que podem estar associados significativamente com o desfecho da doença. Na sequência, uma análise multivariada de regressão logística binária multivariada foi realizada para investigar os fatores de risco associados à mortalidade por COVID-19. Além disso, foi realizada análise de sensibilidade (univariada e multivariada), excluindo pacientes com doença ativa. O *odds ratio* (OR) e os seus respectivos IC 95% foram utilizados para quantificar o tamanho dos efeitos dos fatores de risco. O teste qui-quadrado foi utilizado para investigar associações entre as variáveis independentes e dependentes. Todas as análises estatísticas foram realizadas no SPSS (Nova York, EUA) versão 20 e o limite de significância foi estabelecido em 5,0% ($p < 0,05$).

1.4.2 ESTUDO II: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 NO ÂMBITO INTERNACIONAL

1.4.2.1 Desenho do estudo e coleta de dados

Foi realizado um estudo epidemiológico utilizando o banco de dados de projeções de Coronavírus do *Institute for Health Metrics and Evaluation* (IHME) da Universidade de Washington (Estados Unidos da América - EUA). Os dados coletados são correspondentes ao período entre 3 de janeiro e 4 de agosto de 2020. Foram avaliadas apenas informações dos Estados Unidos da América (EUA), Itália, Espanha

e Alemanha, por serem os países com maior número de casos e mortalidade de COVID-19 na época. Dados de outros países europeus foram excluídos. O número de leitos necessários/dia; número de leitos de unidade de terapia intensiva (UTI) necessários/dia; número de dispositivos de ventilação; número de novas internações/dia; e número de novos pacientes na UTI/dia foram considerados variáveis independentes para pacientes acometidos pela COVID-19. A variável dependente foi óbito por COVID-19, que foi dicotomizada em óbito ou vivo.

1.4.2.2 Análise estatística

Os dados quantitativos com distribuição normal foram apresentados como média e intervalo de confiança (IC) de 95%. As variáveis categóricas (qualitativas) foram apresentadas como frequência. A avaliação dos fatores de risco associados à mortalidade por COVID-19 foi realizada por meio do modelo multivariado de regressão logística. A quantificação do tamanho do efeito dos fatores de risco de mortalidade foi realizada por meio do *odds ratio* (OR) com IC 95%. O modelo foi construído utilizando o critério hierárquico baseado em fundamentações teóricas. O coeficiente de determinação R^2 de Nagelkerke foi utilizado para avaliar a capacidade preditiva do modelo. A qualidade do ajuste do modelo foi avaliada pelos testes de Hosmer-Lemeshow e qui-quadrado. As análises estatísticas foram realizadas utilizando o *software* SPSS versão 20 (Nova York, EUA). O nível de significância de $p < 0,05$ foi considerado estatisticamente significativo.

1.4.3 ESTUDO III: IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO BRASIL E NO MUNDO SOBRE NO AUMENTO DO NÚMERO DE INTERNAÇÕES POR COVID-19

1.4.3.1 Desenho do estudo e coleta de dados

Os dados sobre vacinação contra COVID-19 foram selecionados do banco de dados 'Our World in Data', que é um banco de dados público da *The University of Oxford* financiado pelo Departamento de Saúde e Assistência Social do Reino Unido.

Neste estudo foram utilizados apenas dados de vacinação da população adulta e jovem, a faixa etária com maior cobertura vacinal. A população pediátrica não foi

considerada neste estudo porque até o período em que os dados foram coletados (5 de janeiro de 2022), nem todos os países haviam aprovado o uso da vacina para esta população. O banco de dados contém informações de mais de 200 países. Os dados são padronizados nas seguintes categorias: (i) dados de vacinação contra a COVID-19; (ii) testes; números de casos; (iii) internação; (iv) políticas de resposta às doenças e mortalidade. Para este estudo, utilizamos as informações disponíveis desde 13 de dezembro de 2020 (ou seja, o início da vacinação no mundo) até 5 de janeiro de 2022.

Os dados sobre a "percentagem de população não vacinada que se recusa a receber a primeira dose da vacina contra a COVID-19" foram coletados do *Our World in Data* e podem ser encontrados neste link: [Atitudes em relação às vacinações contra COVID-19](#). Da mesma forma, os dados sobre o "número de internamentos na UTI devido às diversas variantes de COVID-19 durante 2021" também foram coletados do *Our World in Data* e estão disponíveis neste link: [Pacientes COVID-19 em UTI](#).

Foram utilizados dados referentes a 15 países (Estados Unidos - EUA, Reino Unido - Reino Unido, Austrália, Dinamarca, França, Alemanha, Itália, Japão, Holanda, Noruega, Singapura, Coreia do Sul, Espanha e Suécia). Estes países foram selecionados porque foram os únicos que até a data de coleta dos dados (janeiro de 2022) apresentavam dados sobre a percentagem da população que se recusa a ser vacinada. De cada país também foram recolhidos dados sobre o número total de casos confirmados de diversas variantes identificadas (ômicron, beta, épsilon, gama, kappa, iota, eta, delta, alfa, lambda, miu).

1.4.3.2 Análise estatística

Quatro diferentes distribuições de probabilidade foram testadas para verificar qual delas melhor descreve a variável de desfecho, ou seja, número de internações por COVID-19: distribuição normal, distribuição gama, distribuição de Poisson e distribuição tweedie. Os coeficientes de informação dos critérios de Akaike (AIC) foram utilizados para comparar os modelos. Quanto menor o coeficiente AIC, melhor será o ajuste dos dados desta distribuição¹⁴⁵.

Na etapa seguinte, foi utilizado um modelo linear generalizado para avaliar o efeito da rejeição da vacina COVID-19 no número de casos de variantes do SARS-CoV-2. O modelo linear generalizado (regressão gama) foi ajustado considerando o tamanho da população, devido à grande diferença populacional entre os países. Os

resultados foram relatados como coeficientes β (beta). As análises estatísticas foram realizadas no *software* SPSS 20 (Nova York, EUA) e $p < 0,05$ foi considerado significativo.

1.5 RESULTADOS

1.5.1 ESTUDO I: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS AO ATRASO NO DIAGNÓSTICO E MORTALIDADE POR COVID-19 NO ÂMBITO NACIONAL

Entre fevereiro e abril de 2020, o número de pacientes cadastrados com resultado positivo para COVID-19 no Rio de Janeiro foi de 3.656; a maioria eram homens (53,1%) com idade superior a 40 anos (23,2%). A maioria dos pacientes (cerca de 75,0%; $n = 2.738$) vivia em regiões com baixo IDS, incluindo favelas (isto é, Complexo do Alemão, Rocinha, Jacarezinho, Cose Barros). Os demais pacientes ($n = 918$) eram provenientes de regiões com alto IDS (ou seja, Barra da Tijuca, Botafogo, Copacabana, Ipanema, Leblon e Tijuca). Cerca de 50,0% dos pacientes ($n = 1.775$) foram cobertos no período avaliado, 43,1% ($n = 1.577$) ainda apresentavam a doença ativa e 8,3% ($n = 304$) faleceram (**Tabela 1.1**).

Tabela 1.1 Características dos pacientes com COVID-19 no Rio de Janeiro, Brasil (fevereiro-abril de 2020).

Características	Frequência (n)	Porcentagem (%)
Sexo		
Feminino	1940	53,1
Masculino	1677	45,9
Não informado	39	1,1
Idade (anos)		
0-9	22	0,6
10-19	16	0,4
20-29	314	8,6
30-39	825	22,6
40-49	850	23,2
50-59	607	16,6
60-69	413	11,3

70-79	268	7,3
80-89	220	6,0
90-99	29	0,8
Não informado	92	2,5
IDS		
Baixo IDS	2738	74,9
Alto IDS	918	25,1
Tempo entre sintomas e diagnósticos (dias)		
<8	3279	89,7
>8	377	10,3
Evolução da doença		
Recuperado	1775	48,6
Morte	304	8,3
Doença ativa	1577	43,1

Nota: IDS, Índice de desenvolvimento social. **Fonte:** O Autor (2024)

Incluimos todos os 3.656 pacientes em nossa análise do tempo até o diagnóstico (dados sem eventos censurados nem truncados). O tempo mediano para o diagnóstico foi de oito dias (IIQ, 7,236-8,997), com diferenças significativas entre homens (8,0 dias [IIQ, 7,487-8,598]) e mulheres (7,0 dias [IIQ, 6,557-7,595]) (log-rank, $p = 0,027$). A idade dos pacientes teve um efeito significativo no tempo até o diagnóstico (log rank, $p = 0,009$). Pacientes jovens com idade entre 10-19 anos (18,3 dias [IQ, 15,6-20,9]), 20-29 anos (13,5 dias [IIQ, 11,2-15,8]), 30-39 anos (12,3 dias [IQ, 10,6-14,0]), tiveram tempos medianos mais longos até o diagnóstico do que pacientes com idade avançada entre 40-49 anos (7,0 dias [IIQ, 6,0-7,9]), 50-59 anos (6,0 dias [IIQ, 4,9-4,9]), 60-69 anos (6,0 dias [IC 95%, 4,9-7,0]), 70-79 anos (6,0 dias [IIQ, 4,8-7,1]), 80-89 anos (4,0 dias [IIQ, 1,4-6,5]) e 90-99 anos (2,0 dias [IIQ, 1,0-5,2]). Também foram observadas diferenças entre o SDI (log-rank, $p < 0,001$). Pacientes oriundos de regiões com baixo IDS apresentaram maior atraso no diagnóstico (mediana = 7,97 dias [IIQ, 7,530-8,410]) quando comparados aos provenientes de áreas mais desenvolvidas (mediana = 6,42 dias [IIQ, 5,67-7,18]) (**Figura 1**).

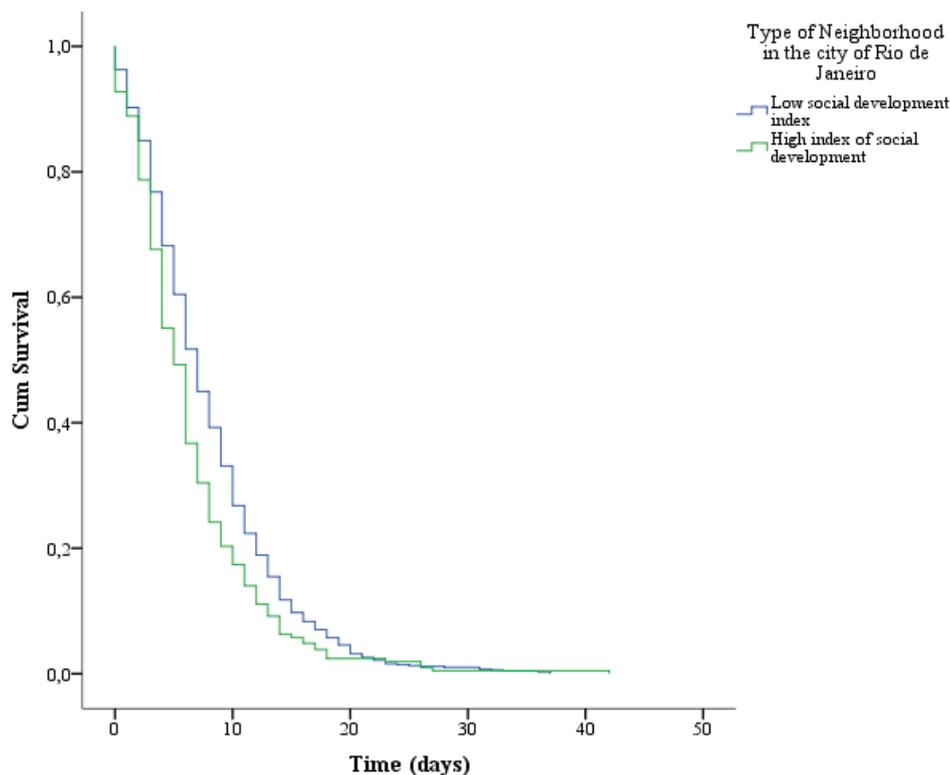


Figura 1.1. Curva de Kaplan-Meier pelo Índice de Desenvolvimento Social do tempo desde o início dos sintomas até o diagnóstico no Rio de Janeiro, Brasil (fevereiro-abril de 2020). **Fonte:** O Autor (2024)

A evolução da doença também teve efeito significativo no tempo até o diagnóstico (log rank, $p < 0,001$). Pacientes recuperados (6,1 dias [IIQ, 5,2-6,9]) tiveram um tempo mediano curto até o diagnóstico do que pacientes falecidos (17,3 dias [IIQ, 14,9-19,7]) e pacientes ainda com a doença ativa (19,7 dias [IIQ, 18,3-21,1]).

Os fatores associados ao tempo até o diagnóstico foram investigados utilizando o modelo de regressão de Cox covariável tempo-dependente, porque a análise prévia das covariáveis demonstrou que a 'idade' não apresentava riscos proporcionais. Os resultados das análises multivariadas de regressão de Cox revelaram que homens (HR = 0,846 [IC 95%, 0,739-0,968]; $p = 0,015$), pacientes que vivem em regiões com baixo IDS (HR = 0,721 [IC 95%, 0,614-0,847]; $p = 0,000$), pacientes com doença ativa (HR = 1,960 [IC 95%, 1,574-2,440]; $p = 0,000$) apresentaram atrasos maiores no diagnóstico. As faixas etárias associadas ao menor tempo até o diagnóstico foram: 40-49 anos (HR = 1,124 [IC 95%, 0,875-1,446]; $p = 0,036$), 50-59 anos (HR = 1,147 [IC 95%, 0,867 -1,519]; $p = 0,033$), 60-69 anos (HR = 1,266 [IC 95%, 0,932-1,720]; $p = 0,013$), 70-79 anos (HR = 1,200 [IC 95%, 0,840-1,713]; $p = 0,016$), 80-89 anos (HR

= 1,160 [IC 95%, 0,773-1,739]; $p = 0,007$) e 90-99 anos (HR = 1,762 [IC 95%, 0,771-4,027]; $p = 0,019$). Resultados semelhantes foram encontrados nas análises univariadas (**Tabela 1.2**).

Tabela 1.2. Resultados da análise multivariada e univariada da regressão de Cox com covariáveis dependentes do tempo dos fatores de risco associados ao atraso no diagnóstico da COVID-19 no Rio de Janeiro, Brasil (fevereiro-abril de 2020)

Característica	Análise univariada				Análise multivariada			
	HR	-IC 95%	+ IC95%	p	HR	-IC 95%	+ IC95%	p
Gênero								
Feminino	1,000	-	-	-	1,000	-	-	-
Masculino	0,867	0,760	0,989	0,034	0,846	0,739	0,968	0,015
Não informado	0,668	0,356	1,251	0,207	0,788	0,418	1,486	0,462
Idade (anos)								
30-39	1,000	-	-	-	1,000	-	-	-
0-9	0,566	0,250	1,282	0,017	0,642	0,283	1,458	0,290
10-19	2,183	1,665	3,160	0,019	2,559	0,777	8,429	0,112
20-29	1,209	0,871	1,679	0,025	1,184	0,852	1,647	0,015
40-49	1,021	0,797	1,309	0,008	1,124	0,875	1,446	0,036
50-59	1,022	0,776	1,347	0,037	1,147	0,867	1,519	0,033
60-69	1,140	0,846	1,536	0,009	1,266	0,932	1,720	0,013
70-79	1,170	0,828	1,654	0,034	1,200	0,840	1,713	0,016
80-89	1,206	,818	1,777	0,005	1,160	0,773	1,739	0,007
90-99	1,464	0,638	3,358	0,041	1,762	0,771	4,027	0,019
Não informado	3,081	0,587	16,162	0,183	1,426	0,340	5,987	0,628
IDS								
Alto IDS	1,000	-	-	-	1,000	-	-	-
Baixo IDS	0,759	0,649	0,888	0,001	0,721	0,614	0,847	0,000
Evolução								
Recuperação	1,000	-	-	-	1,000	-	-	-
Morte	1,307	1,115	1,532	0,001	1,372	1,158	1,625	0,000
Doença ativa	1,857	1,500	2,300	0,000	1,960	1,574	2,440	0,000

Nota: IDS, Índice de desenvolvimento social; HR, *Hazard ratio*; IC, intervalo de confiança. **Fonte:** O autor (2024)

Em relação ao desfecho da doença de acordo com o sexo dos pacientes, mais homens morreram de COVID-19 (5,0%; $n = 182$;) em comparação com mulheres (3,3%; $n = 122$). A taxa de recuperação também foi maior nas mulheres (25,1%; $n =$

919) do que nos homens (22,5%; n =182) (teste qui-quadrado, $p < 0,001$). Ainda assim, 24,6% (n = 899) das mulheres e 18,4% (n = 674) dos homens tinham a doença ativa (Tabela 1.3). Houve também associação significativa entre o desfecho da doença e a idade ($p < 0,001$). Taxas de mortalidade mais elevadas foram encontradas em faixas etárias mais avançadas: pacientes de 20 a 29 anos morreram com menor frequência (0,1%) em comparação com grupos entre 70 e 79 anos (1,8%) e 80 e 89 anos (2,0%). As taxas de recuperação mais elevadas foram registadas em doentes entre os 30 e os 39 anos de idade (11,4%), enquanto estas taxas foram extremamente baixas (0,2%) nos doentes com idades compreendidas entre os 90 e 99 anos. A faixa etária de 40 a 49 anos apresentou maior número de pacientes com doença ativa (11,5%). Foi encontrada associação estatisticamente significativa entre o desfecho da doença e o IDS dos pacientes ($p < 0,001$). Embora os pacientes que moravam em bairros com IDS mais baixo tivessem apresentado um número maior de mortes em comparação com aqueles com IDS alto (3,3% vs. 1,6%), eles se recuperaram com mais frequência (25,1% vs. 14,5%). O tempo entre o início dos sintomas até o diagnóstico também esteve significativamente associado ao desfecho da doença ($p < 0,001$). Foi registrada maior taxa de mortalidade para pacientes com atraso no diagnóstico superior a oito dias (6,0%; n = 220) em comparação com atrasos inferiores a oito dias (2,3%; n = 84). Os pacientes diagnosticados mais cedo (< 8 dias) tiveram taxas de recuperação mais altas em comparação com aqueles diagnosticados mais tarde (41,1% vs. 7,4%) (Tabela 1.3).

Tabela 1.3. Evolução da COVID-19 de acordo com as características dos pacientes no Rio de Janeiro, Brasil (fevereiro-abril de 2020)

Característica	% de desfecho da doença			Total	p*
	Recuperação	Morte	Doença ativa		
Gênero					
Feminino	919 (25,1%)	122 (3,3%)	899 (24,6%)	1.940 (53,1%)	0,00
Masculino	821 (22,5%)	182 (5,0%)	674 (18,4%)	1.677 (45,9%)	
Não informado	35 (1,0%)	0 (0,0%)	4 (0,1%)	39 (1,1%)	
Idade (anos)					
0-9	15 (0,4%)	0 (0,0%)	7 (0,2%)	22 (0,6%)	0,00
10-19	13 (0,4%)	0 (0,0%)	3 (0,1%)	16 (0,4%)	
20-29	151 (4,1%)	4 (0,1%)	159 (4,3%)	314 (8,6%)	
30-39	416 (11,4%)	10 (0,3%)	399 (10,9%)	825 (22,6%)	

40-49	398 (10,9%)	32 (0,9%)	420 (11,5%)	850 (23,2%)	
50-59	315 (8,6%)	41 (1,1%)	251 (6,9%)	607 (16,6%)	
60-69	216 (5,9%)	68 (1,9%)	129 (3,5%)	413 (11,3%)	
70-79	121 (3,3%)	65 (1,8%)	82 (2,2%)	268 (7,3%)	
80-89	67 (1,8%)	73 (2,0%)	80 (2,2%)	220 (6,0%)	
90-99	6 (0,2%)	8 (0,2%)	15 (0,4%)	29 (0,8%)	
Não informado	57 (1,6%)	3 (0,1%)	32 (0,9%)	92 (2,5%)	
IDS					
Baixo	919 (25,1%)	122 (3,3%)	899 (24,6%)	1940 (53,1%)	0,00
Alto	529 (14,5%)	58 (1,6%)	331 (9,1%)	918 (25,1%)	
Tempo entre sintomas e diagnóstico (dias)					
<8	1.504 (41,1%)	220 (6,0%)	1555(42,5%)	3279 (89,7%)	0,00
>8	271 (7,4%)	84 (2,3%)	22 (0,6%)	377 (10,3%)	

Nota: *Teste de qui-quadrado. IDS, Índice de desenvolvimento social. **Fonte:** O autor (2024)

A análise multivariada de regressão logística mostrou que pacientes do sexo masculino tiveram maior chance de morte por COVID-19 do que pacientes do sexo feminino (OR = 0,150 [IC 95%, 0,051-0,440]; $p = 0,001$). As faixas etárias que estiveram estatisticamente associadas ao óbito foram: 70-79 anos (OR = 1,495 [IC 95%, 1,121-1,994]; $p = 0,006$), 80-89 anos (OR = 3,146 [IC 95%, 2,256-4,387]); $p < 0,001$) e 90-99 anos (OR = 5,100 [IC 95%, 2,024-12,852]; $p = 0,001$). Pacientes de regiões com baixo SDI tiveram maior probabilidade de morte por COVID-19 (OR = 1,833 [IC 95%, 1,565-2,148]; $p < 0,001$). Atraso no diagnóstico superior a oito dias também foi fator de risco para óbito (OR = 3,537 [IC 95%, 2,769-4,519]; $p < 0,001$). Os resultados das análises univariada e multivariadas foram semelhantes (**Tabela 1.4**).

Tabela 1.4. Resultados das análises multivariada e univariada da regressão logística dos fatores de probabilidade associados à mortalidade por COVID-19 no Rio de Janeiro, Brasil (fevereiro-abril de 2020)

Característica	Análise univariada				Análise multivariada			
	OR	-IC 95%	+ IC95%	p	OR	-IC 95%	+ IC95%	p
Gênero								
Feminino	1,000	-	-	-	1,000	-	-	-
Masculino	0,110	0,039	0,310	0,000	0,150	0,051	0,440	0,001
Não informado	1,066	0,935	1,214	0,341	1,003	0,875	1,149	0,966
Idade (anos)								
30-39	1,000	-	-	-	1,000	-	-	-
0-9	0,475	0,192	1,176	0,108	0,566	0,224	1,431	0,230
10-19	0,235	0,066	0,830	0,024	0,281	0,078	1,005	0,051
20-29	1,098	0,847	1,424	0,481	1,091	0,838	1,422	0,518
40-49	1,155	0,954	1,399	0,141	1,211	0,995	1,474	0,056
50-59	0,943	0,764	1,163	0,582	1,032	0,832	1,279	0,777
60-69	0,928	0,732	1,175	0,534	1,115	0,872	1,425	0,385
70-79	1,236	0,937	1,629	0,134	1,495	1,121	1,994	0,006
80-89	2,323	1,690	3,192	0,000	3,146	2,256	4,387	0,000
90-99	3,899	1,571	9,674	0,003	5,100	2,024	12,852	0,001
Não informado	0,625	0,401	0,972	0,037	0,822	0,509	1,327	0,422
IDS								
Alto IDS	1,000	-	-	-	1,000	-	-	-
Baixo IDS	1,628	1,400	1,894	0,000	1,833	1,565	2,148	0,000
Tempo entre sintomas e diagnóstico (dias)								
<8	1,000	-	-	-	1,000	-	-	-
>8	3,017	2,386	3,816	0,000	3,537	2,769	4,519	0,000

Nota: IDS, Índice de desenvolvimento social; OR, *Odds ratio*; IC, intervalo de confiança. **Fonte:** O autor (2024)

Após a realização da análise de mortalidade (**Tabela 1.4**), foi então realizada a análise de sensibilidade do modelo de regressão logística dos fatores associados à mortalidade, excluindo os pacientes com doença ativa que representavam 43,1% (n = 1.577) da análise de mortalidade (**Tabela 1.4**) e mantendo apenas pacientes recuperados (48,6%, n = 1.775) e óbitos por COVID-19 (8,3%, n = 304). De modo geral, os resultados da análise de sensibilidade (**Tabela 1.5**) foram semelhantes ao modelo anterior (**Tabela 1.4**), porém foram encontrados aumentos significativos na

magnitude do efeito (aumento do *Odds Ratio*) de todos os fatores associados à mortalidade em comparação com o modelo incluindo pacientes com doença ativa (**Tabela 1.4**), conforme pode ser observado na **Tabela 1.5**. Além disso, o modelo multivariado de análise de sensibilidade mostrou que idades entre 40-49 anos (OR = 3,226 [IC 95%, 1,561-6,668]; p = 0,002), 50-59 anos (OR = 5,341 [IC 95%, 2,625-10,865]; p = 0,000) e 60-69 anos (OR = 13,280 [IC 95%, 6,662-26,474]; p = 0,000) tornaram-se fatores de risco associados à mortalidade por COVID 19, o que difere dos dados do modelo de análise de mortalidade incluindo pacientes com doença ativa (tabela 1.4), onde essas faixas etárias não foram associadas ao risco de mortalidade da doença, mas é o faixas etárias a partir dos 70 anos que têm sido associadas à mortalidade. Os resultados do modelo multivariado de análise de sensibilidade foram semelhantes aos da análise de sensibilidade univariada (**Tabela 1.5**).

Tabela 1.5. Resultados da análise de sensibilidade do modelo de regressão logística multivariada e univariada dos fatores de probabilidade associados à mortalidade por COVID-19 no Rio de Janeiro, Brasil (fevereiro a abril de 2020)

Característica	Análise univariada				Análise multivariada			
	OR	-IC 95%	+ IC95%	p	OR	-IC 95%	+ IC95%	p
Gênero								
Feminino	1,000	-	-	-	1,000	-	-	-
Masculino	0,599	0,467	0,767	0,000	0,642	0,486	0,848	0,002
Não informado	0,069	0,044	0,090	0,790	0,434	0,001	0,800	0,998
Idade (anos)								
30-39	1,000	-	-	-	1,000	-	-	-
0-9	0,431	0,200	0,766	0,998	0,030	0,001	0,070	0,999
10-19	0,200	0,002	0,340	0,999	0,130	0,050	0,450	,999
20-29	1,102	0,341	3,566	0,871	1,130	0,348	3,666	0,839
40-49	3,345	1,623	6,894	0,001	3,226	1,561	6,668	0,002
50-59	5,415	2,671	10,976	0,000	5,341	2,625	10,865	0,000
60-69	13,096	6,609	25,950	0,000	13,280	6,662	26,474	0,000
70-79	22,347	11,143	44,816	0,000	22,562	11,172	45,563	0,000
80-89	45,325	22,296	92,142	0,000	50,726	24,672	104,293	0,000
90-99	55,467	16,207	189,834	0,000	70,321	19,640	251,778	0,000
Não informado	2,189	0,585	8,193	0,244	3,656	0,954	14,007	0,059
IDS								
Alto IDS	1,000	-	-	-	1,000	-	-	-
Baixo IDS	1,801	1,329	2,440	0,000	2,366	1,684	3,324	0,000

Tempo entre sintomas e diagnóstico (dias)								
<8	1,000	-	-	-	1,000	-	-	-
>8	2,119	1,597	2,811	0,000	1,436	1,040	1,983	0,028

Nota: IDS, Índice de desenvolvimento social; HR, *Odds ratio*; IC, intervalo de confiança. **Fonte:** O Autor (2024)

1.5.2 ESTUDO II: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 NO ÂMBITO INTERNACIONAL

Nos EUA, o número estimado de leitos em hospitais foi de 174 leitos/dia (IC 95%, 1,04–3,85), enquanto na Itália, Espanha e Alemanha os números foram de 347 leitos/dia (IC 95%, 3,07–4,22), 357 leitos/dia (IC 95%, 2,96–4,93) e 96 leitos/dia (IC 95%, 6,9–1,85), respectivamente. Em relação ao número de leitos necessários em UTI para tratamento da COVID-19, os valores foram de 48 leitos/dia nos EUA, enquanto Itália e Espanha apresentaram taxas semelhantes (98 e 97 leitos/dia respectivamente), e a Alemanha apresentou os valores mais baixos (26 leitos/dia). Itália e Espanha também apresentaram os maiores números de equipamentos de ventilação (88 e 87 dispositivos/leitos/dia respectivamente), seguidos pelos EUA (43 dispositivos/leitos/dia) e Alemanha (24 dispositivos/leitos/dia). Esses mesmos padrões foram observados para internações hospitalares por COVID-19 (Itália: 47 leitos/dia, Espanha: 48 leitos/dia, EUA: 24 leitos/dia, Alemanha: 13 leitos/dia) e internações em UTI (Itália: 14 leitos/dia). /dia, Espanha: 14 camas/dia, EUA: 7 camas/dia, Alemanha: 4 camas/dia). Durante o período estudado, a previsão cumulativa de mortes por COVID-19 nos EUA foi de 620 (IC 95%, 4,08–12,53), enquanto na Itália foi de 1.428 mortes (IC 95%, 13,17–16,60), na Espanha 1.359 (IC 95%, 11,80–17,82) e na Alemanha 352 mortes (IC 95%, 2,69–6,44). Maiores detalhes são apresentados na tabela 1.6.

Tabela 1.6. Previsão do número médio de leitos, dispositivos de ventilação e internações hospitalares por dia para tratamento de COVID-19 nos EUA, Itália, Espanha e Alemanha (01/03/2020 - 08/04/2020).

Variável	EUA			Itália			Espanha			Alemanha		
	M	-95 % CI	+95 % CI	M	-95 % CI	+95 % CI	M	-95 % CI	+95 % CI	M	-95 % CI	+95 % CI
Número de leitos/dia	174	104	385	347	307	422	357	296	493	96	69	185
Número de leitos de UTI/dia	48	29	104	98	88	118	97	82	133	26	19	51
Número de dispositivos de ventilação/leito/dia	43	26	92	88	79	106	87	73	119	24	17	45
Número de hospitais	24	13	53	47	40	60	48	38	69	13	9	26
Admissões/dia	7	4	14	14	12	17	14	11	19	4	3	7
Número de UTI	7	4	14	14	12	17	14	11	19	4	3	7
Internações/dia	620	408	1253	1428	1317	1660	1359	1180	1782	352	269	644
Mortes acumuladas	620	408	1253	1428	1317	1660	1359	1180	1782	352	269	644

Nota: UTI, unidade de terapia intensiva; IC, intervalo de confiança; M: Média. **Fonte:** O Autor (2024)

Todos os fatores acima mencionados (número de leitos diários, leitos de UTI, dispositivos de ventilação, hospital e internação em UTI) estiveram associados ao risco de morte por COVID 19 nos quatro países (teste qui-quadrado de ajuste do modelo, $p < 0,05$). A predição dos modelos de regressão foi superior a 97% (R^2 Nagelkerk $> 0,85$) e os dados observados foram estatisticamente semelhantes aos dados previstos (Hosmer-Lemeshow, $p > 0,05$) (**Tabela 1.6**).

Tabela 1.6. Teste qui-quadrado para análise de coeficiente de significância, testes Nagelkerk R^2 e Hosmer-Lemeshow para análise de ajuste do modelo de regressão logística multivariada de todas as variáveis estudadas.

Modelo	Chi-square test "p-value"	Hosmer-Lemeshow test "p-value"	Nagelkerk R^2
EUA	<0,001	0,080	0,88
Italia	0,001	0,784	0,87
Espanha	0,001	0,435	0,88
Alemanha	0,001	0,052	0,86

Os fatores de risco associados à mortalidade por COVID-19 nos EUA foram: ter número de leitos hospitalares igual ou inferior a 174 por dia [OR = 286,344 (IC 95% 208,962–392,381); $p < 0,001$] ter número de leitos de UTI para COVID-19 inferior a 48

por dia [OR = 284,585 (IC 95% 206,916–391,408); $p < 0,001$]; ter menos de 43 dispositivos de ventilação disponíveis por dia [OR = 349,251 (IC 95% 243,715–500,488); $p < 0,001$]; número de internações igual ou superior a 24 pacientes por dia [OR = 120,738 (IC 95% 95,670–152,375); $p < 0,001$] e número de pacientes internados em UTI por COVID-19 superior a 7 por dia [OR = 146,838 (IC 95% 113,242–190,402); $p < 0,001$] (Tabela 1.3). Na Itália, os fatores de risco associados à mortalidade por COVID-19 foram: ter número de leitos hospitalares igual ou inferior a 347 ($p < 0,001$); possuir número de leitos de UTI para COVID-19 inferior a 98 ($p < 0,001$); ter menos de 118 dispositivos de ventilação disponíveis por dia ($p < 0,001$); número de internações igual ou superior a 47 pacientes por dia ($p < 0,001$); e número de pacientes internados em UTI superior a 14 por dia ($p < 0,001$) (Tabela 1.8).

Na Espanha, os fatores de risco associados à mortalidade por COVID-19 foram: número de leitos hospitalares inferiores a 357 disponíveis por dia [OR = 146,838 (IC 95% 113,242–190,402); $p < 0,001$]; número de leitos de UTI inferior a 98 por dia [OR = 195,673 (IC 95% 62,340–614,183); $p < 0,001$]; número de dispositivos de ventilação inferior a 87 por dia [OR = 676,618 (IC 95% 215,921–2120,278); $p < 0,001$]; número de internações igual ou superior a 48 pacientes/dia [OR = 284,449 (IC 95% 133,435–606,370); $p < 0,001$]; e número de internações na UTI superior a 14 pacientes por dia [OR = 277,808 (IC 95% 130,314–592,239); $p < 0,001$] (Tabela 1.8). Os fatores de risco associados à mortalidade por COVID-19 na Alemanha foram: número de leitos hospitalares inferior a 96 por dia ($p < 0,001$); número de leitos de UTI inferior a 26 por dia ($p < 0,001$); número de dispositivos de ventilação inferior a 24 por dia ($p < 0,001$); número de internações hospitalares superior a 13 pacientes por dia ($p < 0,001$) e número de internações em UTI superior a 4 pacientes por dia ($p < 0,001$) (**Tabela 1.7**).

Tabela 1.7-A. Análise multivariada dos fatores de risco de mortalidade por COVID-19 nos EUA, Itália, Espanha e Alemanha

Covariável	Itália				Estados Unidos da América				Espanha					
	OR	-95% IC	+95% IC	P	Covariável	OR	-95% IC	+95% IC	P	Covariável	OR	-95% IC	+95% IC	P
Número de leitos/dia [> 347 leitos]	1,0	0,00	Número de leitos/dia [> 174 leitos]	1,0	0,00	Número de leitos/dia [> 357 leitos]	1,0	0,00
[< 347 leitos]	2370,46	332,536	16897,722	[< 174 leitos]	286,34	208,962	392,381	[< 357 leitos]	286,34	208,962	392,381
Núm. de leitos na UTI/dia [> 98 leitos]	1,0	0,00	Núm. de leitos na UTI/dia [>48 leitos]	1,0	0,00	Núm. de leitos na UTI/dia [> 98 leitos]	1,0	0,00
[<98 leitos]	2315,12	324,767	16503,502	[<48 leitos]	284,585	206,916	391,408	[< 98 leitos]	284,585	206,916	391,408
Número de ventiladores/leito/dia [> 118 leitos]	1,0	0,00	Número de ventiladores/leito/dia [>43 leitos]	1,0	0,00	Número de ventiladores/leito/dia [> 87 leitos]	1,0	0,00
[< 118 leitos]	1784,16	250,217	12721,995	[<43 leitos]	349,251	243,715	500,488	[< 87 leitos]	349,251	243,715	500,488
Número de internações/dia [< 47 pacientes]	1,0	0,00	Número de internações/dia [< 24 pacientes]	1,0	0,00	Número de internações/dia [< 48 pacientes]	1,0	0,00
[>47 pacientes]	385,296	171,358	866,332	[> 24 pacientes]	120,73	95,670	152,375	[> 48 pacientes]	120,73	95,670	152,375
Número de internações na UTI/dia [< 14 leitos]	1,0	0,00	Número de internações na UTI/dia [< 7 leitos]	1,0	0,00	Número de internações na UTI/dia [< 14 leitos]	1,0	0,00
[> 14 leitos]	386,819	172,037	869,750	[> 7 leitos]	146,83	113,242	190,402	0,00	[> 14 leitos]	146,83	113,242	190,402	0,00

Nota: IC: intervalo de confiança; OR: Odds Ratio; UTI: Unidade de terapia intensiva. Fonte: Autor (2024)

Tabela 1.8-B. Análise multivariada dos fatores de risco de mortalidade por COVID-19 nos EUA, Itália, Espanha e Alemanha

Covariável	Itália			Espanha			Alemanha					
	OR	-95% IC	+95% IC	P	OR	-95% IC	+95% IC	P	OR	-95% IC	+95% IC	P
Número de leitos/dia				0,00				0,00				0,00
[> 347 leitos]	1,0	1,0
[< 347 leitos]	146,838	113,242	190,402						219,438	123,516	389,855	
Núm. de leitos na UTI/dia				0,00				0,00				0,00
[> 98 leitos]									1,0
[< 98 leitos]	195,673	62,340	614,183						254,884	140,447	462,563	
Número de ventiladores/leito/dia				0,00				0,00				0,00
[> 18 leitos]	1,0	1,0
[< 18 leitos]	676,618	215,921	2.120,278						207,223	118,831	361,364	
Número de internações na UTI/dia				0,00				0,00				0,00
[< 47 pacientes]	1,0	1,0
[> 47 pacientes]	284,449	133,435	606,370						65,990	44,949	96,879	
Número de internações na UTI/dia				0,00				0,00				0,00
[< 14 leitos]	1,0	1,0
[> 14 leitos]	277,808	130,314	592,239						77,519	51,802	116,003	

Nota: IC: intervalo de confiança; OR: Odds Ratio; UTI: Unidade de terapia intensiva. Fonte: Autor (2024)

1.5.3 ESTUDO III: IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO BRASIL E NO MUNDO SOBRE NO AUMENTO DO NÚMERO DE INTERNAÇÕES POR COVID-19

1.5.3.1 Modelo linear generalizado

A **Tabela 1.8** apresenta o número de internamentos em unidades de cuidados intensivos (UTI), o número de casos ômicron e o número de outras infecções pela variante SARS-cov-2. Os países com as maiores taxas de pessoas que recusaram receber a primeira dose da vacina COVID-19 foram França, EUA, Austrália e Reino Unido, com percentagens medianas de 40,80% (IIQ, 22,23% - 53,78%), 38,36% (IIQ, 31,47% - 41,24%), 37,01% (IIQ, 21,58% - 48,50%) e 21,86% (IIQ, 18,87% - 26,52%), respectivamente. O maior número de pacientes em UTI e de casos da variante ômicron ocorreu nos EUA, fato que também foi verificado para as demais variantes (ver **Tabela 1.9**).

A distribuição que melhor descreve os dados das variáveis de desfecho foi a distribuição de Poisson. De acordo com o modelo de regressão Poisson (**Tabela 1.10**), o percentual da população não vacinada contribuiu para um aumento significativo no número de pacientes internados em UTI por COVID-19 ($\beta = 4,581$ [IC 95% 1,986; 6,329], $p = 0,002$), e também no aumento de infecções por variantes do SARS-CoV-2, incluindo ômicron ($\beta = 13,069$ [IC 95% 10,067; 19,070], $p = 0,000$), alfa ($\beta = 5,025$ [IC 95% 2,026; 8,025], $p = 0,000$), delta ($\beta = 6,046$ [IC 95% 2,381; 8,915], $p = 0,001$) e gama ($\beta = 3,321$ [IC 95% 1,324; 6,319], $p = 0,000$) (**Tabela 1.10**).

Tabela 1.8. Percentagem média da população que se recusa a ser vacinada contra a COVID-19 em alguns países, março-dezembro de 2021.

Country	Pacientes na UTI*	Número de casos da variante SARS-CoV-2													Vaccine rejection rate		
		Beta	Epsilon	Gamma	Kappa	Iota	Eta	Delta	Alpha	Lambda	Mu	Omicron	Mediana	Intervalo interquartil			
Austrália	40.378	96	22	8	156	5	7	29128	613	1	1	1693	37,01%	21,58% - 48,50%			
Canadá	279.362	820	758	13271	417	213	1748	84.643	34.985	27	57	612	26,58%	21,03% - 40,33%			
Dinamarca	17.019	223	37	67	28	8	10	156.694	63.862	9	12	4823	25,14%	20,58% - 31,14%			
França	1.098.742	6176	9	1095	15	8	715	93.711	32.651	67	25	843	40,80%	22,23% - 53,78%			
Alemanha	1.098.280	2303	10	858	105	38	677	185.698	104.138	102	17	2270	31,21%	28,72% - 37,26%			
Itália	542.002	116	2	2488	19	10	361	39.386	26.877	14	83	526	28,36%	21,46% - 37,40%			
Japão	60.485	101	19	120	20	5	13	90.083	49.841	4	3	150	28,67%	20,02% - 40,59%			
Países Baixos	167.812	690	5	585	28	2	34	40.036	29.670	12	78	477	22,78%	20,93% - 37,55%			
Noruega	----	411	4	12	3	0	101	17.821	13.842	1	0	308	33,52%	29,51% - 41,60%			
Singapura	5.834	204	4	8	59	6	9	8.504	190	0	0	278	33,26%	18,67% - 50,12%			
Korea do sul	133.821	37	114	15	12	4	2	14.091	816	0	1	17	35,56%	29,33% - 50,43%			
Espanha	681.602	1578	6	1158	5	124	214	34.400	24.732	223	669	703	27,73%	17,04% - 42,31%			
Suécia	59.439	2639	2	184	5	4	14	50.652	68.608	4	4	634	31,82%	28,45% - 35,95%			
Reino Unido	417.160	3105	64393	28733	333	41720	1209	1.327.443	239.829	1254	6041	28536	21,86%	18,87% - 26,52%			
EUA	5.838.472	939	24	225	452	21	427	1.085.714	262.781	9	113	65137	38,36%	31,47% - 41,24%			

Nota: EUA: Estados Unidos da América; UTI, unidade de terapia intensiva. Fonte: Autor (2024)

Tabela 1.9. Modelo de regressão de Poisson do efeito da taxa de rejeição do país no aumento do número de internações em UTI por COVID-19 e no aumento do número de casos de variantes da COVID-19 do SARS-CoV-2

Variável	β	-95% IC	+95%IC	p
Tamanho da população (jovens e adultos)	3,217	1,864	9,538	0,030
COVID-19 ICU patients	4,581	1,986	6,329	0,002
Variante Beta	0,053	-0,317	0,056	0,418
Variante Epsilon	-2,305	-7,331	0,279	0,961
Variante Gamma	3,321	1,324	6,319	0,000
Variante Kappa	-0,010	-3,019	0,001	0,246
Variante Iota	-2,720	-4,759	2,682	0,000
Variante Eta	-0,058	-1,063	0,053	0,742
Variante Delta	6,046	2,381	8,915	0,001
Variante Alpha	5,025	2,026	8,025	0,000
Variante Lambda	0,339	-4,741	1,323	0,731
Variante Mu	0,571	-2,584	0,981	0,217
Variante Ômicron	13,069	10,067	19,070	0,000

Nota: IC, intervalo de confiança. **Fonte:** Autor (2024)

1.6 DISCUSSÃO

1.6.1 ESTUDO I: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS AO ATRASO NO DIAGNÓSTICO E MORTALIDADE POR COVID-19 NO ÂMBITO NACIONAL

No âmbito nacional, foi realizado um estudo de fatores de mortalidade da COVID-19 onde foram avaliados dados de mais de 3.500 pacientes diagnosticados com COVID-19 no Rio de Janeiro, Brasil, por um período de dois meses, com uma taxa de mortalidade resultante em torno de 8,0%. O tempo desde o início dos sintomas até o diagnóstico foi de aproximadamente uma semana, com pacientes do sexo masculino, pessoas mais jovens e aqueles que vivem em regiões menos desenvolvidas apresentando atrasos maiores no diagnóstico. Esses resultados são semelhantes aos apresentados em outros estudos publicados mundialmente^{146,147}.

Estudos anteriores sobre a etiopatogenia da infecção viral e o manejo clínico da doença demonstraram diferenças na prevalência e gravidade da COVID-19 de acordo com o sexo dos pacientes, o que pode ser um importante fator de risco para

mortalidade 15-17 ^{148,149}. Dados epidemiológicos sobre gênero são essenciais para compreender a distribuição de risco, infecção e doença na população, e até que ponto o sexo afeta os resultados clínicos comorbidades ^{150,151}. Um relatório incluindo 552 hospitais de 30 províncias da China revelou que a maioria (58,0%) dos pacientes com COVID-19 eram do sexo masculino e demonstrou que pacientes desse sexo têm maior probabilidade de contrair a doença ¹⁵². Isso pode ocorrer dada a maior prevalência de doenças crônicas anteriores em homens que permitem o desenvolvimento do vírus e contribuem para o aumento da gravidade da doença. Além disso, estudos que avaliam outras SARS causadas por coronavírus sugerem níveis elevados de estrogênio (mais prevalentes em mulheres) como um importante fator de proteção que pode ajudar para controlar a infecção ^{150,151}. Além disso, as mulheres costumam ter um maior autocuidado com a saúde quando comparadas aos homens ¹⁵³, o que pode contribuir para uma percepção mais rápida dos sintomas da doença e, conseqüentemente, para um diagnóstico mais precoce. Neste contexto, é importante promover medidas específicas de prevenção, vigilância e maior intervenção intensiva para homens idosos com comorbidades ^{150,154}. A falta de integração das diferenças de gênero nos inquéritos à COVID-19 pode negligenciar um fator de risco fundamental e possivelmente aumentar as desigualdades nos cuidados de saúde.

Adicionalmente, descobriu-se que o IDS dos pacientes no Rio de Janeiro foi uma variável importante para atrasos no diagnóstico e um fator de risco para mortalidade. Estudos anteriores mostraram extensas desigualdades socioeconômicas no acesso à saúde no Brasil. Em um estudo de Paim et al. (2011) ²⁰, 76,0% das pessoas com renda alta afirmaram ter consultado médico em 2008 para algum diagnóstico clínico, contra 59,0% daquelas com salário menor ¹⁵⁵. Por outro lado, estudo publicado no *The Lancet* demonstrou que 93,0% dos indivíduos que efetivamente procuraram serviços de saúde no Brasil nesse mesmo período receberam tratamento adequado. Ou seja, as desigualdades sociais na utilização dos serviços podem estar associadas, entre outros, ao comportamento dos pacientes ¹⁵⁵. indivíduos com baixos rendimentos podem adiar a decisão de procurar os serviços de saúde devido a experiências negativas passadas e à incapacidade de faltar ao trabalho ¹⁵⁶.

As diferentes regiões do Rio de Janeiro variam em características demográficas e de infraestrutura. Os residentes em regiões com menor IDS são, geralmente, menos

alfabetizados, mais jovens e têm acesso limitado aos serviços sanitários básicos quando comparados aos residentes em regiões com maior IDS¹⁵⁷. Estes determinantes estão frequentemente associados a causar diversas diferenças em várias características (por exemplo, prevalência, gravidade) de doenças entre a população, incluindo agora a COVID-19^{158,159}. Tendo isto em conta, é fundamental que o governo introduza novas medidas de emergência, tais como melhorar a infraestrutura de saúde, para evitar que o vírus se espalhe nestas áreas frágeis. O cenário também destaca a necessidade de fortalecer o acesso ao Sistema Único de Saúde, além da importância de aumentar os investimentos em pesquisa, tecnologia e inovação para combater eficazmente a pandemia no país¹⁶⁰.

Por fim, os dados deste período, demonstravam que o Brasil era o segundo maior número de casos confirmados de coronavírus no mundo, e, ao mesmo tempo, o país ocupava a 19ª posição em aplicação de testes diagnósticos para a população. Até 25 de maio de 2020, foram realizados no país cerca de 3.460 testes por milhão de habitantes, ante 45.586 testes por milhão de habitantes realizados nos Estados Unidos, e os 40 mil testes por milhão de habitantes na Alemanha e na Itália. Esses dados confirmam a falta de testes realizados no Brasil e que os casos eram subnotificados. Estima-se que um em cada 10 casos positivos era notificado^{161,162}.

Nosso estudo tem algumas limitações. Embora tenhamos realizado uma análise com uma coorte de pacientes de uma das maiores cidades do Brasil, representando diferentes níveis socioeconômicos, ela pode não refletir a realidade de outras regiões. Alguns desses dados podem estar subestimados, dada a subnotificação de casos. Outras variáveis potencialmente associadas à COVID-19 poderão ser avaliadas em estudos futuros. A inclusão de pacientes com doença ativa na análise pode representar um viés, pois alguns desses casos podem morrer devido à progressão da doença. Dadas algumas limitações do banco de dados original (por exemplo, falta de informações sobre o acompanhamento do paciente), não foi possível realizar análises adicionais de sobrevida.

1.6.2 ESTUDO II: INVESTIGAÇÃO DOS FATORES DE RISCO ASSOCIADOS A MORTALIDADE DE COVID-19 NO ÂMBITO INTERNACIONAL

Na investigação dos fatores de risco de mortalidade por COVID-19 ao nível internacional (EUA, Espanha, Itália e Alemanha), este estudo demonstrou com

projeção de 7 meses (janeiro de 2020 até agosto de 2020) que o baixo número de leitos de UTI e de dispositivos de ventilação disponíveis por dia está associado a um aumento de 100 vezes ($OR > 100$) na mortalidade por COVID 19 em todos os quatro países avaliados (EUA, Espanha, Itália e Alemanha) se nenhuma medida adicional fosse tomada. Esses quatro países foram escolhidos porque apresentavam os maiores números de casos previstos e a maior mortalidade por COVID-19 na base de dados utilizada. Na China (Wuhan), onde a pandemia começou, 49% de todos os 2.087 pacientes internados gravemente afetados pela COVID-19 morreram, atingindo uma taxa de mortalidade em torno de 62%. Nos EUA, estas taxas eram de aproximadamente 52% (Washington, DC) (Phua et al. 2020) ¹⁶³. Dado que a mortalidade de pacientes graves com pneumonia por SARS-CoV-2 era elevada e que o tempo de sobrevivência desses pacientes variava de 1 a 2 semanas em leitos hospitalares, a falta de infraestrutura para tratamento de pacientes com COVID-19 geraria impacto sobre o curso da doença. Além disso, os hospitais deveriam disponibilizar áreas e leitos com parâmetros mínimos de segurança para admitir pacientes de rotina que não estão infectados pelo vírus, mas que ainda necessitam de cuidados, evitando estender a morbimortalidade a outras doenças (Remuzzi e Remuzzi 2020) ¹⁶⁴.

O número reduzido de camas hospitalares predispõe o sistema de saúde à saturação ¹⁶⁵. A necessidade de medidas como aumentar o número de leitos de UTI já havia sido prevista em alguns estudos realizados para países europeus, como o estudo de Verelst et al. (2020). Na Itália, um dos países mais afetados pela pandemia a nível mundial, estimou-se que o número de camas de UTI necessárias para fornecer um conjunto mínimo de cuidados aos pacientes com COVID-19 atingiria vários milhares. De acordo com o estudo de Verelst et al. (2020) realizado com dados disponíveis até o início de março de 2020 na Itália, foi prevista a ocupação máxima de todos os leitos disponíveis no país ($n = 5.200$) até o final do mesmo mês, apontando a necessidade de alocação de recursos para aumentar o número de leitos a fim de contribuir no enfrentamento da pandemia ¹⁶⁶. No estudo de Remuzzi e Remuzzi (2020), foi calculada uma projeção do número de pacientes em leitos de UTI a partir de fevereiro de 2020, chegando a até 7.500 leitos de UTI em uma semana na Itália se a taxa de infecção continuasse a aumentar ¹⁶⁴.

Um estudo publicado por Rhodes et al. (2012) evidenciaram uma heterogeneidade muito considerável no número de camas de UTI entre os países

europeus, com a Alemanha, por exemplo, a ter 6,9 vezes o número de camas per capita em comparação com outras nações (Rhodes et al. 2012) ¹⁶⁷. Estas diferenças podem ser explicadas, entre outras coisas, pelas variações socioeconômicas entre as regiões e pelas características dos sistemas de saúde (Rhodes et al. 2012) ¹⁶⁷. Esta heterogeneidade entre os países também ficou evidente no que diz respeito à necessidade de estrutura durante uma pandemia de COVID-19. Enquanto a Itália e a Espanha identificaram o número de leitos de UTI de 98 por dia como um fator de risco associado à mortalidade por COVID-19, a Alemanha apresentou o número de 26 por dia. Numa posição intermediária aos países europeus, os EUA apontaram o número de 48 leitos de UTI por dia como fator de risco. A mesma classificação foi observada na comparação dos demais fatores de risco, como o número total de leitos hospitalares e o número de dispositivos de ventilação disponíveis por dia. Os piores resultados foram apontados para Itália e Espanha, seguidos dos EUA, enquanto a Alemanha apresentou o melhor cenário.

Os pacientes infectados pela COVID-19 que necessitam de internação eram geralmente classificados de acordo com as manifestações clínicas e gravidade da doença em: (1) doença leve com pneumonia ou ausência (81% dos casos), (2) doença grave com dispneia, frequência respiratória alterada, saturação de oxigênio no sangue $\leq 93\%$, alteração na relação PaO_2/FiO_2 e na porcentagem de oxigênio fornecido com possível infiltrado pulmonar (14% dos casos) e (3) doença crítica com insuficiência respiratória, choque séptico ou múltiplas disfunção orgânica (MOD) ou falência (MOF) (5% dos casos) ¹⁶⁸. Em um estudo de coorte retrospectivo de 201 pacientes com pneumonia confirmada por COVID-19 internados no Hospital Wuhan Jinyintan, na China, entre 25 de dezembro de 2019 e 26 de janeiro de 2020, 82,1% dos pacientes (n = 165) necessitaram de suporte de oxigênio no hospital. Na China continental, estimou-se que 3,2% de todos os pacientes infectados com COVID-19 receberam intubação e ventilação invasiva ¹⁶⁹. Apesar da pequena porcentagem de pacientes necessitando de ventilação invasiva, o grande número de pacientes infectados simultaneamente pode significar acesso limitado a este tipo de recurso de tratamento ¹⁷⁰. Estimou-se que nos EUA existiam cerca de 60.000 a 160.000, conforme a funcionalidade do equipamento. Porém, de acordo com o presente estudo, menos de 43 ventiladores disponíveis por dia foi um fator de risco para óbito, ou seja, o número de ventiladores no país é crítico diante do crescente aumento de casos de COVID-19. O cenário era ainda mais crítico quando considerava-se os números de Itália e

Espanha (118 por dia e 87 por dia respectivamente), que são países que já careciam destes ventiladores ¹⁷¹.

Embora a pandemia exigia um aumento significativo do número de camas nos hospitais, as circunstâncias também podiam reduzir os recursos disponíveis. O surto de SARS (do inglês, *Severe Acute Respiratory Syndrome*) em Toronto levou ao fechamento durante 10 dias de 38% dos leitos de UTI de uma universidade terciária, principalmente devido à falta de profissionais afetados pela doença ou em quarentena. Neste contexto, além de aumentar a capacidade de leitos e disponibilidade de ventiladores, os hospitais devem ter estratégias para aumentar a segurança dos funcionários ¹⁷².

Estes resultados confirmaram a afirmação da OMS de que a COVID-19 era um problema emergencial de saúde pública de interesse mundial, tanto para países desenvolvidos (como EUA, Espanha, Itália e Alemanha) quanto para países em desenvolvimento. Esta doença saturava os sistemas de saúde locais, independentemente do país. Neste contexto, era importante reforçar novas medidas públicas preventivas, como a quarentena, para o controle da doença. As pessoas em quarentena contribuíam para a redução de novas infecções virais, diminuindo consequentemente a superlotação dos hospitais. Além disso, eram necessários mais investimentos nas infraestruturas dos hospitais, bem como o desenvolvimento de dispositivos inovadores para a ventilação dos pacientes.

O presente estudo é baseado em dados de predição, o que pode representar uma limitação. As previsões sobre o número de novos casos e mortalidade podiam não refletir a realidade completa em todos os países, mas auxiliou na compreensão do que poderia acontecer até agosto de 2020. Além disso, não foi possível avaliar a associação de diferentes variáveis relacionadas ao acesso aos cuidados de saúde (por exemplo, ambientes de saúde, categorias de cuidados) nos resultados da COVID-19, uma vez que estes dados não estavam disponíveis na base de dados consultada. Mais estudos considerando essas informações eram necessários para melhor compreender o impacto dos diferentes serviços de saúde na evolução da doença.

1.6.3 ESTUDO III: IMPACTO DA REJEIÇÃO DA VACINAÇÃO NO BRASIL E NO MUNDO SOBRE NO AUMENTO DO NÚMERO DE INTERNAÇÕES POR COVID-19

Demonstrou-se por meio do modelo de regressão de Poisson que a rejeição da vacina contra COVID-19 tem impacto significativo no aumento de internações em UTI pela doença.

A percentagem média da população que recusou receber a primeira dose da vacina contra a COVID-19 foi de cerca de 20% para todos os países/regiões avaliados. Já a França (40,80%), os EUA (38,36%), a Austrália (37,01%) e o Reino Unido (21,86%) apresentaram valores acima da média. Esses países também foram aqueles que relataram as maiores taxas de internação em UTI por COVID-19 e aumento do número de infecções por novas variantes – especialmente ômicron (1.098.742, 5.838.472, 40.378 e 417.160 casos, respectivamente) durante o período avaliado (até 31/01/2022). Estudos conduzidos no período também demonstraram que os países de rendimento elevado são mais propensos a taxas de rejeição da vacinação em comparação com regiões de rendimento baixo e médio. Isto pode ocorrer, entre outros, devido à desinformação e notícias falsas espalhadas nas redes sociais, que impactam negativamente no processo de vacinação, levantando dúvidas sobre a sua eficácia e segurança ¹⁷³. As redes sociais permitem que utilizadores de qualquer nível educacional e socioeconômico criem e compartilhem informações sem rigor editorial ou processo de revisão por pares. Além do desconhecimento, outros fatores como o negacionismo, motivos políticos e ideológicos, culturais e religiosos podem afetar a aceitação das vacinas pelas diferentes populações, o que impacta diretamente na saúde pública. Como exemplo, os EUA, país que estava à frente na vacinação e estagnou pelo alto índice de rejeição, devido a diversos fatores, como etnia/raça, política, nível econômico, localização geográfica e fatores religiosos ¹⁷⁴.

A rejeição à vacina também pode favorecer a ocorrência de mutações virais levando ao surgimento de novas variantes com diferentes potenciais de virulência e infecção, como ocorreu com o SARS-Cov-2 ¹⁷⁵. Esta pode ser uma barreira importante ao controlo da pandemia. De acordo com o modelo de regressão de Poisson, o percentual da população não vacinada teve efeito significativo no aumento de internações em UTI ($p=0,002$) e esteve relacionado ao maior número de infecções por omícron ($p=0,000$), alfa ($p=0,000$), variantes delta ($p=0,001$) e gama ($p=0,000$) em diferentes regiões do mundo. Além disso, as taxas de rejeição da vacina podem afetar

as previsões dos modelos preditivos de séries temporais (modelos ARIMA e Holt), reduzindo o seu desempenho. Isso vale também para modelos já publicados que consideram apenas a população jovem e adulta elegível para receber a vacina. Possíveis medidas para minimizar a rejeição à vacinação incluem campanhas educativas, combate a informações enganosas, responsabilidade criminal daqueles que produzem e divulgam notícias falsas - especialmente figuras públicas, programas de vacinação obrigatória ^{173,176}. Outras estratégias além da retaliação devem ser exploradas para os países que detectam e divulgam novas variantes do SARS-CoV-2, como ocorreu com a África do Sul ¹⁷⁴, para minimizar, entre outras, discrepâncias na cobertura vacinal. Embora milhões de doses de vacina sejam agora doadas por alguns países de alto rendimento (por exemplo, EUA, Reino Unido) a regiões de baixo rendimento, devem ser asseguradas condições logísticas e de armazenamento (por exemplo, controlo de temperatura, prazo de validade mais longo) para garantir a qualidade da vacina ¹⁷⁷.

A associação entre a rejeição à vacinação e o aumento do número de casos em UTI deve ser interpretada com cautela, pois a disponibilidade de leitos de UTI e o significado de “cuidados intensivos” variam consideravelmente em diferentes países. Portanto, isso constituiu uma limitação do estudo, pois essas informações não estavam disponíveis nas bases de dados consultadas. Outra limitação é a possibilidade de subnotificação do número de pessoas vacinadas e do número de casos em UTI que podem ocorrer em países de baixa renda (por exemplo, na África Subsaariana) variável que não pode ser controlada ^{173,174}.

1.7 CONCLUSÃO

Este primeiro capítulo da tese ressalta a urgência de estratégias abrangentes e adaptáveis para lidar com os desafios pandêmicos, como o da pandemia de COVID-19. Tanto no contexto nacional quanto internacional, identificou-se fatores de risco cruciais para a mortalidade associada à doença. No Brasil, a vulnerabilidade socioeconômica e a falta de acesso a diagnósticos precoces foram determinantes significativos, destacando a necessidade de medidas preventivas direcionadas aos grupos mais vulneráveis e aprimoramento dos recursos de saúde.

Globalmente, a infraestrutura hospitalar deficiente emergiu como um fator crítico na luta contra a COVID-19 no primeiro ano da pandemia (2020). A escassez de leitos de UTI e equipamentos médicos essenciais ressaltou a necessidade premente de investimentos na expansão e fortalecimento dos serviços de saúde.

Além disso, o impacto direto da hesitação em relação à vacinação é inegável. O aumento das internações em UTI e a propagação das variantes virais são reflexos claros da resistência à imunização. Diante disso, é crucial implementar estratégias que promovam a conscientização, a obrigatoriedade da vacinação em determinados contextos, além de restrições e sanções para mitigar informações anti-vacinação.

Enfatiza-se a importância não apenas de identificar os fatores de risco, mas de agir proativamente para enfrentar os desafios associados à COVID-19. É imperativo um esforço conjunto entre líderes políticos, profissionais de saúde e a sociedade para fortalecer nossos sistemas de saúde, expandir recursos e promover a adesão às medidas preventivas, incluindo a vacinação, visando mitigar os impactos devastadores dessa pandemia.

2 CAPÍTULO II - AVALIAÇÃO DO EFEITO DOS ALIMENTOS E NUTRIENTES COMO ABORDAGENS COMPLEMENTARES NA RECUPERAÇÃO DE PACIENTES COM COVID-19 EM 170 PAÍSES

Publicado em:

Cobre AF, Surek M, Vilhena RO, Böger B, Fachi MM, Momade DR, Tonin FS, Sarti FM, Pontarolo R. Influence of foods and nutrients on COVID-19 recovery: A multivariate analysis of data from 170 countries using a generalized linear model. Clin Nutr. 2022 Dec;41(12):3077-3084. doi: 10.1016/j.clnu.2021.03.018.

2.1 RESUMO

Introdução: A COVID-19 é um problema de saúde pública de emergência de importância global. Este estudo foi realizado no começo de 2021, um momento em que, embora muitos artigos científicos tenham sido publicados em 2020, faltavam evidências sobre a relação entre nutrição e COVID-19. Como uma forma de resolver esse problema, em dezembro de 2020, a *European Society for Clinical Nutrition and Metabolism* (ESPEN) e o Escritório Regional da Organização Mundial de Saúde para a Europa (OMS/Europa) discutiram este tema e concordaram em abrir um convite aos pesquisadores de todo o mundo para submissão de artigos científicos sobre o estado nutricional e os cuidados nutricionais em pacientes com COVID-19, a serem publicados nas revistas ESPEN *Clinical Nutrition* e *Clinical Nutrition*, os dois periódicos científicos oficiais da ESPEN em 2021. O estudo teve como objetivo investigar o efeito dos alimentos e nutrientes como abordagens complementares na recuperação da COVID-19 em 170 países, especialmente considerando a complexidade da doença e a atual escassez de tratamentos ativos. **Metodologia:** Um estudo retrospectivo foi realizado utilizando o banco de dados do Kaggle, que relaciona o consumo de vários alimentos com a recuperação da COVID-19 em 170 países. Foram empregadas análises multivariadas baseadas em um modelo linear generalizado. **Resultados:** Os resultados mostraram que certos alimentos tiveram um efeito positivo na recuperação da COVID-19, incluindo ovos, peixes e frutos do mar, frutas, carne, leite, raízes amiláceas, estimulantes, produtos vegetais, nozes, óleo vegetal e vegetais. O consumo de níveis mais elevados de proteínas e lipídios também foi associado a uma recuperação mais favorável, enquanto o alto consumo de bebidas alcoólicas teve um efeito negativo. **Conclusão:** Em países desenvolvidos, onde a fome foi erradicada, os alimentos tiveram um impacto significativo na recuperação da COVID-19, especialmente aqueles ricos em lipídios, proteínas, antioxidantes e micronutrientes. No entanto, em países com extrema pobreza, os alimentos apresentaram pouco efeito na recuperação da doença. Esses achados destacam a importância da nutrição como uma abordagem complementar no tratamento da COVID-19, especialmente em contextos de desenvolvimento socioeconômico mais avançado.

Palavras-chave: Coronavírus, macronutrientes, nutrientes, fome, análise multivariada

2.2 INTRODUÇÃO

A pandemia de COVID-19 teve um enorme impacto na saúde, social e econômico em todo o mundo durante os últimos anos. O resultado clínico é pior em vários grupos de pacientes em risco de desnutrição. Pessoas com obesidade também apresentam maior risco clínico¹⁷⁸. Os cuidados nutricionais são, portanto, de extrema relevância em todos os pacientes com COVID-19 e podem desempenhar um papel importante também na prevenção desta doença e de outras^{79,80,178}.

No entanto, apesar da relevância da investigação nutricional na COVID-19, até o final de 2020, as evidências disponíveis eram extremamente limitadas. Das mais de 75.000 publicações sobre COVID-19 disponíveis no PubMed em 2020, apenas 1.200 eram recuperadas ao adicionar nutrição na pesquisa, e apenas 169 quando desnutrição é usada como palavra-chave⁸⁰.

O reforço das evidências sobre o potencial papel fundamental do estado nutricional e dos cuidados nutricionais na COVID-19 era, portanto, uma necessidade não satisfeita e uma prioridade urgente. Em dezembro de 2020, a *European Society for Clinical Nutrition and Metabolism* (ESPEN) e o Escritório Regional da Organização Mundial de Saúde para a Europa (OMS/Europa) discutiram este tema e concordaram em abrir um convite aos pesquisadores de todo o mundo para submissão de artigos científicos sobre o estado nutricional e os cuidados nutricionais em pacientes com COVID-19, a serem publicados nas revistas ESPEN *Clinical Nutrition* e *Clinical Nutrition*, os dois periódicos científicos oficiais da ESPEN^{79,80}. Entre os temas de interesse incluía os estudos voltados ao tratamento nutricional médico em pacientes com COVID-19 e seu impacto nos resultados clínicos em todos os ambientes clínicos⁸⁰. Em resposta a esse chamado, o presente capítulo, cujos resultados foram publicados no periódico científico *Clinical Nutrition*, teve como objetivo investigar o efeito dos alimentos e nutrientes como abordagens complementares na recuperação da COVID-19 em 170 países.

2.3 OBJETIVOS

2.3.1 *Objetivo geral:*

O objetivo geral deste estudo foi investigar o efeito dos alimentos e nutrientes como abordagens complementares na recuperação da COVID-19 em 170 países, considerando a complexidade da doença e a escassez de tratamentos ativos. Para alcançar esse objetivo, foram realizadas análises multivariadas baseadas em um modelo linear generalizado utilizando dados do consumo de vários alimentos relacionados à recuperação da COVID-19.

2.3.2 *Objetivos específicos:*

- Analisar o impacto de diferentes tipos de alimentos na recuperação da COVID-19 em dados de 170 países;
- Avaliar a relação entre o consumo de proteínas, lipídios e outros macronutrientes com a recuperação da COVID-19;
- Investigar o efeito do consumo de bebidas alcoólicas na recuperação da COVID-19;
- Identificar os alimentos que apresentaram um efeito positivo na recuperação da COVID-19, bem como aqueles que tiveram um efeito negativo;
- Comparar os resultados entre países desenvolvidos, onde a fome foi erradicada, e países com extrema pobreza, para entender como o contexto socioeconômico influencia o impacto dos alimentos na recuperação da doença.

2.4 MATERIAL E MÉTODOS

2.4.1 *Desenho do estudo*

A presente investigação compreende um estudo observacional com análise quantitativa de dados transversais de 170 países ao nível populacional.

2.4.2 Conjunto de dados

Os dados foram obtidos na plataforma pública Kaggle ¹⁷⁹, mantida pelo governo dos Estados Unidos da América (EUA), que inclui as seguintes informações de diversas bases de dados:

- A oferta alimentar disponível em cada país em 2017 (expressa como proporção de calorias), de acordo com os alimentos e o seu conteúdo nutricional de macronutrientes (expresso como a proporção de proteínas, gorduras e hidratos de carbono), da Organização para a Alimentação e Agricultura dos Estados Unidos Nações Unidas (FAO) ¹⁸⁰;
- A população de cada país, do Bureau PLB¹⁸¹;
- A proporção de indivíduos que se recuperaram da infecção por COVID-19 em cada país (última atualização em 27 de janeiro de 2021), do Centro Johns Hopkins de Ciência e Engenharia de Sistemas¹⁸² e a taxa de mortalidade de COVID-19

Além disso, informações de conjuntos de dados da FAO sobre o abastecimento de alimentos de acordo com itens alimentares e grupos de alimentos (em quilogramas per capita por ano) de cada país foram incorporadas ao banco de dados, para abordar o efeito do consumo de alimentos na recuperação da COVID-19 ¹⁸⁰, ajustado pela população de cada país. Considerando que não houve mudanças significativas no abastecimento alimentar nos últimos quatro anos em todo o mundo, estes dados foram utilizados no cenário atual de 2020 e 2021.

Também foi acessado um banco de dados contendo as seguintes variáveis: i) Pessoas que tinham acesso aos serviços mínimos de saneamento básico (% da população) em 2017; ii) Prevalência de subnutrição (% da população) em 2018; iii) Despesas correntes de saúde per capita, PPC (dólares internacionais correntes) em 2017, iv) Produto interno bruto (PIB) per capita, PPC (dólares internacionais correntes) em 2019; v) Taxas padronizadas de prevalência de diabetes (% da população com idade entre 20 e 79 anos) em 2019. Esses dados foram acessados no site oficial do Banco Mundial. Essas variáveis foram utilizadas como controle na avaliação da influência da alimentação na recuperação do COVID-19¹⁸³.

2.4.3 Variáveis do estudo

A variável dependente na análise foi a taxa de recuperação da COVID-19 em cada um dos 170 países. Uma definição clínica para pacientes “recuperados” ou “curados” não está especificamente disponível nas diretrizes internacionais para o manejo clínico de casos suspeitos ou confirmados de COVID-19 e pode variar em todo o mundo. Assim, considerando as últimas diretrizes da OMS e do *Center of Diseases Control* (CDC) dos EUA, 'recuperação' foi definida seguindo os critérios para alta dos pacientes do isolamento: (i) Casos sintomáticos: 10 dias após o início dos sintomas ou o recebimento de um resultado de teste positivo se a data de início não puder ser determinada, além de pelo menos 24 horas sem sintomas (incluindo febre e sintomas respiratórios) e medicamentos para reduzir a febre. Se um paciente apresentar doença grave devido à COVID-19, os profissionais de saúde podem recomendar períodos de isolamento mais longos após o início dos sintomas (possivelmente até 20 dias); (ii) Casos assintomáticos: 10 dias após os pacientes testarem positivo para SARS-CoV-2; (iii) Os pacientes pertencentes aos seguintes grupos são definidos como ‘recuperados’ apenas com base em testes: pacientes hospitalizados em qualquer fase da doença; cuidadores em instituições com uma população em risco de morbidade grave devido ao COVID19 (por exemplo, lares de idosos, instituições que cuidam de pacientes geriátricos, instalações de vida assistida), pacientes que sofrem de supressão do sistema imunológico. Os documentos da OMS relevantes para este estudo estão disponíveis para acesso através deste link: [Documentos da OMS](#). Por favor, clique para explorar mais detalhes.

As variáveis independentes de interesse foram a proporção de calorias, proteínas, gorduras e carboidratos provenientes de bebidas alcoólicas, gorduras animais, produtos de origem animal, cereais, ovos, peixes e frutos do mar, frutas, leite, carnes, diversos, vísceras, oleaginosas, leguminosas, temperos., raízes ricas em amido, estimulantes, açúcar e adoçantes, culturas açucareiras, frutos secos, óleos vegetais, vegetais e produtos vegetais

O efeito do abastecimento alimentar na recuperação da COVID-19 foi analisado considerando todos os países num modelo geral, bem como modelos específicos considerando países agrupados em regiões e categorias do índice global da fome (IGF) em 2019:

- Grupo 1 – Países desenvolvidos com erradicação da fome: Canadá, Estados Unidos da América (EUA), Japão, Coreia do Sul, Austrália, Nova Zelândia e alguns países da Europa Ocidental (por exemplo, França, Reino Unido e Suécia);
- Grupo 2 – Países das Américas com índice global de fome baixo a moderado: países da América Central, América do Sul e México.
- Grupo 3 – Países da Europa (por exemplo, Albânia, Ucrânia e Rússia) e da Ásia (por exemplo, Irão, Tailândia e China) com baixo índice global de fome.
- Grupo 4 – Países com índice global de fome baixo, moderado ou alto: todos os países da África, países da Oceania (Fiji, Polinésia Francesa, Ilhas Salomão e Kiribati) e alguns países da Ásia (por exemplo, Iêmen, Paquistão e Vietnã).

Para minimizar o viés nos modelos lineares multivariados generalizados (GLM) construídos para avaliar a influência dos alimentos e nutrientes na recuperação da COVID-19, as seguintes variáveis foram utilizadas como variáveis de controle: i) Pessoas que utilizam pelo menos serviços de saneamento básico (% da população) em 2017; ii) Prevalência de subnutrição (% da população) em 2018; iii) Despesas correntes em saúde per capita, PPC (dólares internacionais correntes) em 2017, iv) PIB per capita, PPC (dólares internacionais correntes) em 2019; v) Taxas padronizadas de prevalência de diabetes (% da população com idade entre 20 e 79 anos) em 2019; vi) Taxas de mortalidade da COVID-19.

2.4.4 Análise estatística

A análise estatística incluiu uma investigação prévia da distribuição da variável dependente (recuperação da COVID-19), testando três tipos diferentes de distribuição de probabilidade: a distribuição normal, a distribuição gama e a distribuição Tweedie. O Critério AIC foi o parâmetro adotado para comparação de ajuste entre as distribuições testadas, considerando que coeficientes AIC mais baixos indicam melhor ajuste da variável à distribuição em avaliação ^{184–187}.

O teste qui-quadrado ajustado ao GLM foi utilizado para avaliar os principais efeitos das covariáveis na recuperação da COVID-19, uma etapa preliminar na análise multivariada. Por fim, covariáveis (itens alimentares) com efeito significativo nas taxas de recuperação da COVID-19 foram incluídas no GLM ajustado à distribuição gama. O tamanho do efeito das covariáveis foi baseado nos coeficientes beta do GLM com

intervalo de confiança de 95%. Magnitudes de efeitos (b) maiores que 1 foram consideradas as mais substanciais. A análise estatística foi realizada no *software* SPSS, versão 20.0, adotando nível de significância $p < 0,05$.

2.5 RESULTADOS

2.5.1 Considerações gerais dos resultados

O estudo incluiu 170 países: 46 (27,05%) da África, 35 (20,58%) das Américas, 38 (22,35%) da Ásia, 42 (24,70%) da Europa e os restantes 9 (5,3%) da Oceania.

Considerando a avaliação das taxas de recuperação da COVID-19 em 170 países através do AIC, a distribuição gama apresentou os valores mais baixos dos coeficientes AIC; portanto, foram realizadas análises estatísticas ajustando os dados para a distribuição gama.

2.5.2 Abastecimento de alimentos e taxas de recuperação da COVID-19

A maioria dos 170 países estava bem abastecida com cereais (13,9% e 21,1% da oferta total de alimentos), frutas (6,4% e 11,6%), leite (6,3% e 22,8%), raízes amiláceas (4,7% e 7,4%) e hortaliças (6,7% e 33,9%). Os países do Grupo 1 apresentaram a maior oferta de 10 itens alimentares (bebidas alcoólicas, gorduras animais, carnes, leite, diversos, frutos do mar, estimulantes, açúcar, nozes e óleos vegetais), enquanto os países do Grupo 2 tiveram a maior oferta de frutas, miudezas e culturas de açúcar. Os países do Grupo 3 tinham a maior oferta de ovos, peixes, frutos do mar, raízes ricas em amido e vegetais. Por outro lado, os países do Grupo 4 tinham a maior oferta de cereais, oleaginosas, leguminosas e especiarias.

As altas taxas de recuperação da COVID-19 foram observadas nos países do Grupo 1 (ou seja, países com fome erradicada), enquanto os países do Grupo 4 (ou seja, países com uma elevada taxa de fome global) apresentaram as taxas de recuperação mais baixas da doença. No entanto, os países do Grupo 4 também apresentaram a menor prevalência de diabetes mellitus como comorbidade [6,00 (IIQ, 3,80 - 9,65] e as taxas mais baixas de letalidade associada à COVID-19 [0,0018% (IIQ, 0,0006 - 0,0077%)]. Nos países dos grupos 2 e 3, a prevalência de diabetes foi de 9,1000 (IIQ, 7,30- 11,35) e 6,50 (IIQ, 5,85 - 9,15), respectivamente, enquanto as

taxas de letalidade foram de 0,06% (IIQ, 0,0118 - 0,10), 0,0231% (IIQ, 0,0034 - 0,08), respectivamente.

Os países do Grupo 4 também tiveram o menor acesso aos serviços básicos de saúde, com uma percentagem mediana de 59,54% (IIQ, 36,22% - 91,18%), enquanto os países do grupo 1 apresentaram as taxas mais altas de 99,26% (IIQ, 99,26% - 99,89%). Os países do Grupo 3 tiveram a segunda maior taxa de acesso aos cuidados de saúde básicos [96,22% (IIQ, 96,22% - 97,97%)], seguidos pelo Grupo 2 [87,79% (IIQ, 83,46% - 93,80%)]. As maiores taxas de desnutrição foram encontradas nos Grupos 2 e 4, com percentuais medianos de 48,20% (IIQ 25% - 69%) e 28% (IIQ, 17,20% - 47%), respectivamente. Os países do Grupo 1 tiveram as taxas mais baixas deste resultado, com uma percentagem mediana de 7% (IIQ, 9,0% - 15%), enquanto no Grupo 3 as taxas foram de 26% (IIQ, 25% - 42,03%). Em relação ao PIB, os países de extrema pobreza com um elevado IGF tiveram a PPC mais baixa, com uma mediana de \$ 5.165 (IIQ, 2.357,50 - 10.920), enquanto nos países desenvolvidos com erradicação da fome, a renda per capita em PPC foi a mais alta, com uma mediana de \$ 51.010 (IIQ, 42.402,5 - 60.850). Os países do Grupo 2 e do Grupo 3 tiveram uma renda per capita moderada de US\$ 15.140 (IIQ, 9770 - 21.120) e US\$ 21.205 (IIQ, 14.675 - 33.447,5), respectivamente.

2.5.3 Análise multivariada do modelo linear generalizado

Os resultados da análise multivariada de um GLM ajustado pela distribuição gama em relação aos efeitos das covariáveis (itens alimentares) nas taxas de recuperação da COVID-19 considerando os 170 países investigados são apresentados nas Tabelas 2.1 e 2.2. Os modelos GLM foram estimados utilizando a oferta de proteínas (modelo 1), lipídios (modelo 2) e carboidratos (modelo 3) e oferta alimentar em geral (modelo 4) (Tabela 2.1). Além disso, foram estimados modelos GLM considerando cada um dos quatro grupos de países categorizados de acordo com o IGF (Tabela 2.2).

2.5.3.1 Efeito do abastecimento alimentar e dos seus macronutrientes nas taxas de recuperação da COVID-19

A análise multivariada do GLM envolvendo a oferta de proteínas mostrou diversos itens alimentares com efeito significativo na recuperação da COVID-19 (Tabela 2.1). No modelo 1, as fontes de proteína (por exemplo, produtos de origem animal, ovos, peixes e frutos do mar, carne, leite, vísceras, vegetais) impactam positivamente na recuperação dos pacientes ($p < 0,05$). Por outro lado, as bebidas alcoólicas tiveram efeito negativo significativo no desfecho ($\beta = 8,181$ [IC 95% -14,290; 2,071], $p = 0,009$), o que também foi observado nos modelos 3 e 4, confirmando a desvantagem do consumo desse item. O modelo 2 revelou que dos 20 itens lipídicos analisados, 8 (40%) beneficiam significativamente os pacientes (ovos, peixes, frutos do mar, carnes, leite, vísceras, nozes, produtos vegetais e óleo vegetal). Por outro lado, nenhum alimento do grupo de carboidratos (modelo 3) teve impacto no resultado. O modelo de abastecimento alimentar geral (modelo 4) demonstrou que apenas ovos ($\beta = 33,143$ [IC 95% 15,554; 50,732], $p < 0,0001$), peixes e frutos do mar ($\beta = 31,526$ [IC 95% 14,222; 48,830], $p < 0,0001$), frutas ($\beta = 31,388$ [IC 95% 14,199; 48,578], $p < 0,0001$), carne ($\beta = 31,491$ [IC 95% 14,211; 48,772], $p < 0,0001$), leite ($\beta = 31,741$ [IC 95% 14,444; 49,038], $p < 0,0001$), miudezas ($\beta = 31,477$ [IC 95% 13,944; 48,951], $p < 0,0001$) e produtos vegetais ($\beta = 23,769$ [IC 95% 7,480; 40,057], $p = 0,004$) podem contribuir para a recuperação da COVID-19.

Tabela 2.1. Análise multivariada de modelo linear generalizado para estimar o efeito da quantidade de proteína consumida (modelo 1), da quantidade de lipídios consumidos (modelo 2), da quantidade de carboidratos consumidos (modelo 3) e da quantidade de alimento consumidos em geral (modelo 4) na recuperação da COVID-19 em 170 países.

Covariável	Modelo 1		Modelo 2		Modelo 3		Modelo 4					
	β	[95%IC]	β	[95%IC]	β	[95%IC]	β	[95%IC]				
Bebidas alcoólicas	-8,181	-14,290 - -2,071	0,009	----	----	-2,553	-4,937 - -1,831	0,019	-9,750	-15,261 - 4,238	0,001	
Gorduras animais	7,335	-12,344 - 27,015	0,465	-0,553	-2,937 - 1,831	0,649	0,786	-46,057 - 47,628	0,974	-1,674	-9,846 - 6,499	0,688
Produtos animais	28,664	9,066 - 48,262	0,004	8,250	-5,322 - 21,822	0,233	1,053	-25,157 - 27,263	0,937	-1,771	-10,002 - 6,459	0,673
Cereais	1,900	-11,629 - 15,428	0,783	-0,783	-3,201 - 1,635	0,526	0,753	-7,613 - 9,119	0,860	-1,896	-10,128 - 6,336	0,652
Ovos	33,995	15,299 - 52,692	<0,001	29,700	10,010 - 49,390	0,003	1,393	-45,433 - 48,219	0,953	33,143	15,554 - 50,732	<0,001
Peixe-mariscos	33,347	14,746 - 51,948	<0,001	28,021	8,517 - 47,526	0,005	0,333	-46,482 - 47,148	0,989	31,526	14,222 - 48,830	<0,001
Frutas	2,099	-11,420 - 15,618	0,761	-0,538	-2,967 - 1,891	0,664	1,175	-7,209 - 9,560	0,784	31,388	14,199 - 48,578	<0,001
Carne	33,429	14,826 - 52,032	<0,001	28,403	8,842 - 47,965	0,004	0,667	-46,167 - 47,500	0,978	31,491	14,211 - 48,772	<0,001
Leite	33,501	14,906 - 52,096	<0,001	28,660	9,095 - 48,225	0,004	0,733	-46,102 - 47,568	0,976	31,741	14,444 - 49,038	<0,001
Diversas	2,562	-10,828 - 15,953	0,708	-1,076	-3,645 - 1,492	0,412	1,785	-6,524 - 10,093	0,674	-1,815	-9,877 - 6,247	0,659
Óffal	33,380	14,754 - 52,006	<0,001	28,678	8,730 - 48,626	0,005	1,249	-45,788 - 48,286	0,958	31,447	13,944 - 48,951	<0,001
Oil crops	1,634	-11,885 - 15,152	0,813	-1,146	-3,531 - 1,238	0,346	0,823	-7,517 - 9,163	0,847	-1,961	-10,070 - 6,148	0,635
Pulses	1,807	-11,717 - 15,331	0,793	-1,049	-3,470 - 1,373	0,396	0,082	-8,203 - 8,366	0,985	-2,454	-10,712 - 5,805	0,560
Especiálias	2,082	-11,631 - 15,795	0,766	-0,497	-3,362 - 2,367	0,734	1,253	-7,203 - 9,708	0,772	-0,291	-8,451 - 7,870	0,944
Raízes ricas em amido	1,829	-11,693 - 15,351	0,791	-0,814	-3,226 - 1,598	0,508	0,445	-7,778 - 8,668	0,916	-1,889	-10,102 - 6,323	0,652
Estimulantes	2,158	-11,328 - 15,643	0,754	-0,013	-2,542 - 2,515	0,992	1,229	-7,090 - 9,548	0,772	-2,000	-10,451 - 6,452	0,643
Açúcares adoçantes	2,825	-12,746 - 18,395	0,722	-0,568	-2,963 - 1,828	0,642	-0,568	-2,963 - 1,828	0,642	-1,609	-9,734 - 6,516	0,698

Nozes	-0,124	-2,527 - 2,279	0,919	2,413	1,510 - 3,025	0,024	1,036	-7,273 - 9,346	0,807	-,314	-8,078 - 7,450	0,937
Produtos vegetais	39,758	17,382 - 62,134	<0,001	21,173	6,992 - 35,355	0,003	0,876	-22,821 - 24,574	0,942	23,769	7,480 - 40,057	0,004
Oleos vegetais	1,891	-11,692 - 15,474	0,785	5,733	3,162 - 7,696	0,004	0,889	-7,462 - 9,239	0,835	-1,623	-9,786 - 6,541	0,697
Vegetais	32,898	14,260 - 51,536	0,001	-0,767	-3,199 - 1,666	0,537	0,490	-8,047 - 9,028	0,910	-1,798	-9,989 - 6,393	0,667

Nota: IC, intervalo de confiança. **Fonte:** O Autor (2024)

2.5.3.2 Efeito do abastecimento alimentar nas taxas de recuperação da COVID-19 segundo grupos de países

O efeito do abastecimento alimentar na recuperação da COVID-19 também foi avaliado de acordo com grupos de países agregados utilizando o IGF (Tabela 2.2). As variáveis de controle do modelo foram: serviços de saneamento básico; prevalência da subnutrição; despesas correntes em saúde per capita; despesas nacionais de saúde do governo geral per capita, PIB per capita, RNB per capita, prevalência de diabetes e taxa de mortalidade.

Os resultados mostraram que a prevalência de diabetes foi negativamente associada à recuperação dos pacientes nos países dos Grupos 1 ($p = 0,031$) e 3 ($p = 0,017$). A subnutrição não teve efeito sobre o resultado em quase todos os países do grupo, exceto para aqueles do Grupo 4 (ou seja, pobreza extrema), onde foi observado um impacto negativo significativo na recuperação da COVID-19 ($\beta = 6,611$ [IC 95% - 11,492; 1,730], $p = 0,008$). A disponibilidade de serviços de saneamento básico foi um fator importante associado à recuperação dos pacientes nos países do Grupo 1 ($p = 0,010$) e do Grupo 4 ($p < 0,0001$).

Nos países pertencentes aos Grupos 1 e 2, tanto a despesa per capita em saúde como o RNB per capita tiveram um impacto positivo no resultado (valores de $p < 0,05$). Por outro lado, para os países do Grupo 4, o RNB per capita em PPC teve um efeito negativo significativo na recuperação dos pacientes ($p < 0,001$). Os países do Grupo 1 foram os únicos associados a um resultado positivo em relação ao PIB per capita em PPC ($p = 0,006$).

Os modelos GLM confirmaram que o fornecimento de determinados alimentos teve efeitos positivos na recuperação da COVID-19 nos países dos Grupos 1, 2 e 3: ovos, peixe, marisco, carne, leite, produtos vegetais. Os vegetais foram positivamente associados ao resultado em todos os países. Produtos de origem animal, frutas e óleos vegetais apresentaram resultado positivo apenas nos países dos Grupos 1 e 3. Nos países do Grupo 3, as miudezas e nozes foram positivamente associadas à recuperação dos pacientes. O álcool foi o único item alimentar com efeito negativo significativo no resultado nos países dos Grupos 1 e 2.

Tabela 2.2. Coeficientes do modelo linear generalizado para a magnitude dos efeitos dos alimentos consumidos na recuperação da COVID-19, segundo grupos do índice global de fome.

Covariável	Países do grupo 1			Países do grupo 2			Países do grupo 3			Países do grupo 4		
	β	[95%CI]	p	β	[95%CI]	p	β	[95%CI]	p	β	[95%CI]	p
Taxa de mortalidade (%)	-	-31,083 - -	<0,001	-	32,831 - -	<0,001	-	-22,594 - -	<0,001	-	-15,399 -	0,435
Prevalência da diabetes (padronizada)	17,825	4,567	0,031	25,788	18,745	0,248	13,470	4,346	0,017	8,627	6,397	0,131
Pessoas que utilizam o mínimo de serviços de saneamento básico (%)	-4,159	-7,162 - -	0,010	-0,067	-0,180 -	0,226	-3,009	-5,174 -	0,064	0,084	-0,025 -	<0,001
Prevalência de desnutrição (%)	27,322	6,660 -	0,133	0,047	-1,306 -	0,630	6,051	-0,111 -	0,769	3,571	2,333 - 4,810	0,008
PIB per capita	1,212	-0,369 -	0,006	5,538	-1,297 -	0,356	-0,466	-2,424 -	0,140	-	-11,492 - -	0,132
RNB per capita*	5,710	3,101 -	<0,001	2,142	2,142	0,002	0,115	0,882	0,162	6,611	1,730	0,978
Despesas atuais com saúde per capita*	3,069	2,201 -	<0,001	0,592	-1,646 -	0,007	1,272	-0,419 -	0,162	0,876	4,773	0,988
Despesas domésticas do governo geral com saúde per capita**	3,233	1,832 -	0,470	5,886	1,151 -	0,001	0,340	-0,137 -	0,274	-	-2,860 -	0,778
Bebidas alcoólicas	0,326	4,634	0,016	2,344	5,123	0,028	0,474	0,818	0,678	0,040	2,780	0,106
Gorduras animais	-4,336	-0,908 -	0,016	1,590	-5,034 - -	0,028	0,039	-0,146 -	0,678	-	-0,186 -	0,894
Produtos animais	-0,683	6,090	0,421	3,707	-0,635 -	0,670	0,283	1,445	0,001	0,023	0,140	0,894
	4,235	-2,349 -	<0,001	0,982	8,048	0,286	2,068	1,320 -	0,001	-	-0,182 -	0,894
		1,270 -		6,342	0,166		3,167	0,159		0,012	0,159	

Cereais	-0,090	-0,531 - 0,351	0,690	-----	-----	-----	-0,076	-0,187 - 0,035	0,179	-	-0,125 - 0,018	0,143
Ovos	13,234	7,553 - 20,914	<0,001	7,911	3,477 - 12,699	0,001	9,721	2,312 - 16,870	0,043	-	-0,818 - 4,388	0,590
Peixe-mariscos	8,240	5,714 - 9,235	0,032	2,611	1,160 - 3,206	0,003	3,587	2,371 - 5,197	0,014	0,063	-0,135 - 0,262	0,531
Frutas	2,013	1,510 - 0,484	0,039	-0,001	-0,120 - 0,118	0,987	1,800	1,200 - 2,244	0,008	0,025	-0,108 - 0,159	0,709
Carne	5,694	3,769 - 8,382	0,002	2,002	1,241 - 3,246	0,009	1,914	1,506 - 2,090	<0,001	0,464	0,129 - 0,799	0,070
Leite	11,018	8,183 - 17,146	0,027	2,112	1,517 - 3,241	0,030	5,096	2,001 - 7,194	0,005	0,198	0,065 - 0,332	0,400
Miscellaneous	0,577	-0,162 - 1,315	0,126	0,348	-2,390 - 2,348	0,790	0,284	-0,971 - 1,539	0,658	-	-1,193 - - 0,070	0,270
Offal	-1,299	-7,797 - 5,199	0,695	0,714	-2,438 - 3,865	0,657	4,737	2,748 - 6,725	0,020	-	-4,245 - 0,929	0,209
culturas oleaginosas	-2,653	-4,965 - 0,341	0,250	-0,745	-1,548 - 0,058	0,069	-1,506	-3,145 - 0,133	0,072	-	-0,312 - - 0,048	0,070
leguminosas	-2,303	-5,235 - 0,629	0,124	-0,749	-1,593 - 0,094	0,082	-3,239	-7,239 - 1,239	0,900	-	-1,430 - - 0,403	0,068
Especiárias	-5,904	-32,263 - 20,456	0,661	-6,233	-9,343 - 3,122	0,200	-1,274	-3,798 - 1,251	0,323	1,555	-1,513 - 4,623	0,321
Raízes ricas em amido	0,018	-0,570 - 0,606	0,952	-0,006	-0,140 - 0,128	0,929	0,045	-3,247 - 1,158	0,240	-	-0,142 -0,059	0,890
Estimulantes	0,636	-1,058 - 2,330	0,462	0,220	-1,761 - 2,201	0,828	0,288	0,166 - 1,410	0,160	1,849	-1,198 - 4,895	0,234
Açúcares e adoçantes	0,261	-0,088 - 0,610	0,143	-0,197	-0,514 - 0,121	0,226	0,167	-0,160 - 0,493	0,317	0,567	0,265 - 0,869	0,470

Nozes	4,160	1,668 - 5,348	0,170	0,345	-3,151 - 3,841	0,847	1,927	1,266 - 2,212	0,005	2,033	-0,416 - 4,482	0,104
Produtos vegetais	3,469	1,927 - 6,990	0,029	0,025	-1,126 - 1,176	0,966	2,348	1,292 - 4,595	0,009	0,245	-1,329 - 1,725	0,500
Oleos vegetais	5,160	2,006 - 7,314	0,042	1,952	1,166 - 2,490	0,002	3,068	2,167 - 5,032	0,008	0,350	-1,342 -2,361	0,587
Vegetais	15,042	11,290 - 17,373	<0,001	3,095	2,230 - 5,034	0,014	5,023	3,121 - 8,075	0,004	6,905	3,084 - 9,326	0,001

Nota. IC, intervalo de confiança; Grupo 1: Países desenvolvidos com erradicação da fome; Grupo 2: Países do continente americano com índice global de fome baixo ou moderado; Grupo 3: Países europeus e asiáticos com baixa taxa de fome global; Grupo 4: Países com índice global de fome baixo, moderado ou alto; *Valores expressos em escala logarítmica; PIB: Produto interno Bruto; RNB: Rendimento Nacional Bruto. Fonte: O Autor (2024)

2.6 DISCUSSÃO

2.6.1 Efeitos do fornecimento de alimentos nas taxas de recuperação da COVID-19

Os resultados apresentados no estudo mostram a influência da oferta alimentar nas taxas de recuperação da COVID-19, considerando a diversidade e o conteúdo nutricional dos alimentos disponíveis nos países estudados. Os países onde a fome foi erradicada apresentaram taxas mais elevadas de recuperação da doença. Por outro lado, as regiões com efeitos reduzidos da oferta alimentar nas taxas de recuperação da COVID-19 foram as do Grupo 4. Isso pode ocorrer, entre outros fatores, porque o aumento do estresse oxidativo induzido por deficiências alimentares, especialmente no que se refere ao fornecimento de nutrientes antioxidantes, afetam a resposta imunológica do indivíduo, levando ao aumento da suscetibilidade a doenças virais como COVID-19 ¹⁸⁸. Portanto, além da coexistência de doenças crônicas, o estado nutricional de um paciente infectado pela COVID-19 também deve ser considerado, pois as deficiências nutricionais podem aumentar o risco de infecção grave ¹⁸⁸. As fontes alimentares de origem animal apresentaram contribuições importantes para a recuperação dos pacientes com COVID-19 nos modelos estimados. São fontes de alto teor calórico, proteínas e micronutrientes essenciais. Ovos, leite, carne e peixe tiveram um efeito substancial ($\beta > 1$) na recuperação da doença nos países dos Grupos 1, 2 e 3. Nestes grupos, os ovos e a carne foram consumidos em maiores quantidades. A oferta de ovos também teve efeito significativo em todos os países, sendo uma importante fonte de selênio e vitamina A pré-formada (retinol) ^{189,190}.

Alimentos de origem animal também são fontes de colesterol, sendo os ovos um dos alimentos mais ricos neste componente. Embora o colesterol esteja associado ao desenvolvimento de muitas doenças, alguns estudos demonstraram que determinados níveis de colesterol podem aumentar a resistência do organismo a infecções, atuando neste sentido na modulação da resposta imune ^{191,192}. Leite, carne e produtos lácteos são fontes de zinco e retinol^{189,192}, em alguns países do Grupo 1 (Finlândia, Noruega, Suécia, Canadá e EUA), os produtos lácteos são sistematicamente suplementados com vitamina D¹⁹³. Dentre os componentes biologicamente ativos dos peixes, os ácidos graxos poli-insaturados ômega-3 (EPA e DHA) são os mais estudados, sendo encontrados principalmente em espécies ricas

em óleo ¹⁹⁴. No entanto, peixes e frutos do mar também fornecem vitamina D e selênio, além de uma composição equilibrada de aminoácidos ¹⁹⁴.

Em nosso modelo, os alimentos de origem vegetal tiveram maior influência na recuperação dos pacientes em países com menores taxas de deficiências nutricionais no que diz respeito à oferta de calorias. Os óleos vegetais apresentaram contribuições significativas na maioria dos países, principalmente nos Grupos 1 e 2, onde houve maior consumo de óleos, que são compostos principalmente por ácidos graxos insaturados (óleo de colza, óleo de girassol). Também houve grande consumo de soja e azeites nos países do Grupo 1. Os óleos vegetais são importantes fontes de vitamina E (a-tocoferol) [36]; os óleos de soja e colza são fontes particularmente importantes de ácidos graxos insaturados ômega-3 (ácido α -linolênico) [37].

Os frutos secos (exceto as castanhas) também contribuíram de forma importante para as taxas de recuperação da COVID-19, sendo uma das fontes de lipídeos mais importantes depois dos óleos vegetais, com elevados raios de gordura insaturada e saturada. As nozes são um dos alimentos integrais com maior teor de ácido α -linolênico. As nozes também são fonte de proteína (aproximadamente 25% de energia) e geralmente possuem alto teor de L-arginina, aminoácido precursor do vasodilatador endógeno óxido nítrico (NO). Eles também fornecem antioxidantes, como vitamina E e compostos fenólicos ¹⁹⁵. Por fim, as frutas foram alimentos com oferta abundante entre os países dos Grupos 1 e 3, caracterizados pelo maior consumo de maçãs e uvas, alimentos com alto teor de polifenóis ^{196,197}. Os vegetais constituíram o único grupo de alimentos que mostrou um efeito significativo na recuperação da COVID-19 em países de todos os grupos. Esses itens são fontes de glutathione, um tripéptido antioxidante, além de vitamina C, principalmente frutas cítricas, brócolis, tomate e folhas verdes ^{198,199}. Na verdade, a ingestão de frutas e vegetais já foi investigada quanto aos benefícios potenciais associados a condições respiratórias e inflamatórias ²⁰⁰.

O único alimento que teve efeito negativo na recuperação da COVID-19 foram as bebidas alcoólicas. Além de causar danos hepáticos, o consumo de álcool tem sido associado a doenças pulmonares. O álcool perturba a função ciliar nas vias aéreas superiores, prejudica a função dos macrófagos alveolares e neutrófilos e enfraquece a função da barreira epitelial nas vias aéreas inferiores ²⁰¹.

2.6.2 Efeitos do fornecimento de macro e micronutrientes nas taxas de recuperação da COVID-19

No geral, as proteínas e os lipídios tiveram efeitos significativos nas taxas de recuperação da COVID-19. Níveis baixos de proteína podem aumentar o risco de infecções associadas, por exemplo, à baixa produção de anticorpos ²⁰². Os ácidos graxos são importantes na resposta à infecção porque podem alterar significativamente a resposta imune, incluindo alterações na organização dos lipídios celulares e nas interações com receptores nucleares. Em camundongos, o consumo de ácidos graxos tem sido associado à redução da homeostase e das funções das células imunológicas ²⁰².

Um estudo anterior mostrou que o nível de colesterol sérico era significativamente menor entre pacientes chineses com COVID-19 ²⁰³ e a hipolipidemia estava associada à gravidade da doença ²⁰². O colesterol baixo predispõe a doenças infecciosas porque o LDL-c participa do sistema imunológico aderindo e inativando microrganismos e seus produtos tóxicos, incluindo vírus ²⁰⁴. Os carboidratos, que em geral não apresentaram efeitos na recuperação da COVID-19, podem estar associados à oferta de carboidratos processados e ao alto índice glicêmico, o que está relacionado à sobrecarga da capacidade mitocondrial e ao aumento da produção de radicais livres [3]²⁰². Por outro lado, muitos alimentos ricos em selênio, zinco e vitaminas (leite, carne, peixe, frutos do mar, ovos, vegetais) foram positivamente associados à recuperação dos pacientes da COVID-19. O selênio dietético constitui principalmente selenoproteínas, como as selenoenzimas antioxidantes glutathione peroxidases (GPxs) e tioredoxina redutase (TrxRs), altamente expressas em células do sistema imunológico, como linfócitos T e macrófagos. Os efeitos benéficos do selênio foram relatados quase exclusivamente para infecções por vírus RNA [46]²⁰⁵. O zinco é um oligoelemento dietético crítico para o desenvolvimento de células imunológicas e um cofator para muitas enzimas. A deficiência deste micronutriente pode contribuir para uma mediação deficiente das células imunológicas e aumento da suscetibilidade a várias infecções, incluindo pneumonia ¹⁸⁸.

A vitamina A é um nutriente amplamente estudado no campo das funções imunológicas. Experimentos *in vitro* e estudos em animais sugerem que os retinóides são importantes reguladores da diferenciação e da função monocítica. A

imunocompetência das células T pode ser afetada pela deficiência de vitamina A, incluindo linfopoiese, distribuição, expressão de moléculas de superfície e produção de citocinas ¹⁹⁹ A vitamina D tem efeitos pleiotrópicos na via imunológica do hospedeiro e pode estar envolvida na função imunológica pulmonar e alveolar. As evidências sugerem que a suplementação com vitamina D3, a forma ativa da vitamina D, pode diminuir a suscetibilidade ou melhorar a recuperação de infecções como gripe, pneumonia recorrente e tuberculose ^{88,206}.

Na verdade, a sazonalidade de muitas infecções virais está associada a baixas concentrações de vitamina D, devido às baixas doses de UVB devido ao inverno em climas temperados e à estação chuvosa em climas tropicais. É o caso da gripe, infecção causada pelo vírus sincicial respiratório e pelo SARS-CoV ²⁰⁷. A vitamina E, uma vitamina solúvel em gordura, é um potente antioxidante e pode modular as funções imunológicas do hospedeiro. Sua deficiência prejudica a imunidade humoral e celular. Para algumas doenças crônicas, como a hepatite B, a sua suplementação está associada a resultados positivos ²⁰⁸ [42]. Finalmente, a vitamina C é mais conhecida pelas suas propriedades antioxidantes, sendo capaz de eliminar espécies reativas de oxigênio, protegendo assim as células e tecidos do corpo contra danos oxidativos e disfunções. Além disso, a vitamina C desempenha um papel no apoio às funções imunológicas. A depleção desta vitamina pode estar associada a uma gravidade crescente de infecções ²⁰⁹. Um estudo recente realizado nos EUA com 167 pacientes com SDRA relacionada à sepse indicou que a administração de ~15 g/dia de vitamina C IV por 4 dias pode diminuir a mortalidade dos pacientes ²¹⁰.

Os ácidos graxos poli-insaturados ômega-3 (EPA e DHA) servem como precursores para a produção de mediadores pró-resolução por macrófagos e neutrófilos ²¹¹. Até no fim de 2020, o EPA estava sendo testado em vários ensaios clínicos em pacientes com COVID-19. O ácido α -linolênico (ALA) exibe potentes propriedades anti-inflamatórias. Num modelo animal de lesão pulmonar aguda induzida por lipopolissacarídeos, o tratamento com ALA inibiu significativamente a secreção de citocinas pró-inflamatórias. Além disso, a diminuição das atividades da glutathiona (GSH) e da superóxido dismutase (SOD) causada pelo polissacarídeo foi revertida pelo tratamento com ALA ²¹².

No geral, os nossos resultados mostraram que o consumo de diferentes alimentos, especialmente proteínas, alguns lipídios, antioxidantes e micronutrientes, pode beneficiar significativamente a recuperação da COVID-19. Isto pode ser

influenciado, entre outros, pelas culturas alimentares geográficas, socioeconómicas e ecológicas dos países. É possível que na maioria dos países de extrema pobreza (Grupo 4), as baixas condições socioeconómicas estejam associadas a um acesso mais difícil aos alimentos - o que contribui para a desnutrição.

Além disso, fatores como a escassez de serviços de saneamento básico, os gastos extremamente baixos em saúde per capita e o baixo PIB per capita levam ao aparecimento de várias doenças diarreicas e infecciosas que enfraquecem a população e a tornam mais vulnerável à desnutrição e às infecções oportunistas (por exemplo, pneumonia) ^{213,214}. Nestes casos, a alimentação pode não desempenhar o papel mais significativo na recuperação da COVID-19, uma vez que uma dieta equilibrada já é irrealista para a maioria das pessoas. A base alimentar destes países é composta principalmente por carboidratos, sendo deficiente em óleos e proteínas de origem animal.

Finalmente, embora um rendimento mais elevado esteja frequentemente associado a taxas mais baixas de desnutrição, a sua melhoria reduz a desnutrição apenas num pequeno grau. Estas estimativas sugerem que os países não podem confiar apenas no crescimento económico para reduzir a subnutrição dentro de um prazo aceitável ¹⁸³. São necessárias mais medidas socioeconómicas, educacionais e culturais para melhorar esta métrica em todo o mundo. É importante destacar que países com maior renda e maior acesso à alimentação costumam apresentar menor prevalência de desnutrição medida através do Índice de Massa Corporal (IMC); entretanto, indivíduos com sobrepeso (IMC 25 kg/m²) ou obesidade (IMC 30 kg/m²) também podem estar desnutridos, especialmente considerando a ausência de informações adicionais sobre composição corporal e exames diagnósticos. Outros estudos demonstraram evidências significativas sobre o impacto de um mau estado nutricional geral nos resultados negativos de saúde de pacientes infectados por COVID-19 ⁸³⁻⁸⁵.

Nosso estudo tem algumas limitações. No período do estudo, não existiam critérios padrão para a definição de “recuperação” da COVID-19 – dada a complexidade da doença e a dinâmica da pandemia, por isso foi utilizada as diretrizes vigentes no momento da OMS e do CDC. Contudo, deve-se estar ciente de que estas definições não consideram as sequelas resultantes da doença, que podem ter um efeito a longo prazo na qualidade de vida dos pacientes, incluindo anomalias da

função pulmonar, comprometimento psicológico, depressão, redução da capacidade de exercício e transtorno de estresse pós-traumático ^{215,216}

A desnutrição é um conceito geral geralmente relacionado ao baixo IMC, que representa apenas uma dimensão na mensuração do estado nutricional dos indivíduos. A definição de desnutrição através da estimativa do IMC representa parcialmente o conceito de desnutrição, que pode ser influenciado por outros fatores como o acesso aos alimentos (ou seja, insegurança alimentar ou indivíduos desnutridos com sobrepeso/obesidade, considerando que micronutrientes e outros exames de diagnóstico também existem em regiões/países com alta disponibilidade de alimentos) e qualidade do ambiente de saúde familiar/comunitário. Portanto, variáveis adicionais que permitam diagnóstico adicional de desnutrição entre pacientes com COVID-19 devem ser consideradas, de acordo com a proposta da Iniciativa de Liderança Global sobre Desnutrição (GLIM), que afirma que pelo menos um critério fenotípico (seja perda de peso, baixo IMC, ou redução da massa muscular como a sarcopenia), e deve ser considerado pelo menos um critério etiológico (redução da ingestão ou assimilação de alimentos, ou inflamação ou carga de doença), particularmente considerando a idade dos pacientes especialmente nos países desenvolvidos ^{217,218}. No entanto, dada a falta de disponibilidade de dados sobre dimensões adicionais para o diagnóstico global da desnutrição, não foi possível estimar modelos adicionais que abrangessem múltiplas variáveis relativas ao estado nutricional das populações dos países.

O mesmo ocorreu com fatores de confusão clínicos adicionais, como as principais comorbidades dos pacientes (por exemplo, prevalência de hipertensão, insuficiência renal ou doença pulmonar obstrutiva crônica), cujas informações não estavam disponíveis. No entanto, os padrões de prevalência destas comorbidades são semelhantes em todo o mundo. Embora tenha-se utilizado dados sobre a oferta alimentar de 2017, as análises estatísticas não revelaram alterações importantes nesta variável nos últimos quatro anos. Por questões metodológicas, a variável acesso aos serviços de saneamento básico só foi estimada no modelo quando pelo menos 50% da população estava coberta pelo serviço. Um desenho transversal utilizando dados de nível ecológico foi selecionado para nosso estudo, o que compromete a capacidade de propor inferências sobre as relações longitudinais entre exposição (alimentos) e desfecho (taxas de recuperação do COVID 19), especialmente considerando a dinâmica da pandemia. Neste estudo foi possível fornecer

informações importantes sobre o uso de alguns alimentos como abordagens complementares ao manejo da COVID-19, no entanto, estudos longitudinais são necessários para confirmar esses resultados.

2.7 CONCLUSÃO

Neste capítulo II, os modelos lineares generalizados desenvolvidos no presente estudo permitiram a identificação de alguns macronutrientes (lipídios e proteínas) e micronutrientes (selênio, zinco) com efeito positivo significativo nas taxas de recuperação da COVID-19 nos 170 países investigados. Os alimentos com efeitos positivos nas taxas de recuperação foram ovos, peixes e frutos do mar, frutas, carne, leite, produtos vegetais, vegetais, frutas e nozes. Esses alimentos são fontes de nutrientes essenciais para promover o bom funcionamento do sistema imunológico. A maior disponibilidade de álcool teve um efeito negativo nas taxas de recuperação. Descobriu-se que os carboidratos não tiveram efeito no resultado, o que provavelmente se deve às características metabólicas da infecção por SARS-CoV-2. Nos países desenvolvidos onde a fome foi erradicada, os alimentos tiveram maiores efeitos na recuperação da doença, enquanto nos países com um índice global de fome mais elevado, o consumo de alimentos quase não teve efeito. Estas conclusões podem ser incluídas em orientações ou recomendações para a gestão da COVID-19 como abordagens complementares para melhorar a recuperação dos pacientes, especialmente considerando a complexidade da doença e a atual escassez de tratamentos farmacológicos adicionais.

**3 CAPÍTULO III - ACURÁCIA DA TÉCNICA DE ESPETROFOTOMETRIA DE
INFRAVERMELHO NO DIAGNOSTICO DE COVID-19: UM ESTUDO DE
REVISÃO SISTEMÁTICA COM META-ANÁLISE**

3.1 RESUMO

Introdução. Neste capítulo III desta tese de doutorado, o objetivo foi sintetizar as evidências disponíveis sobre o uso da espectroscopia infravermelha (FTIR) para o diagnóstico da COVID-19. **Métodos.** Para isso, foi realizada uma revisão sistemática com meta-análise (PROSPERO CRD42022360026) de acordo com o recomendado pela Cochrane, para avaliar a especificidade, sensibilidade e precisão da FTIR em comparação com um método padrão de referência para diagnosticar a COVID-19 em amostras humanas. As buscas foram conduzidas no PubMed e Embase em setembro de 2023 e após a etapa de triagem, elegibilidade e extração dos dados, meta-análises e área sob a curva ROC (AUC) foram calculadas no *software* Meta-Disc 1.4.7. A qualidade metodológica dos estudos incluídos foi avaliada usando o QUADAS-2. **Resultados.** Dezesete estudos foram incluídos. A técnica FTIR apresentou alta sensibilidade [0,912 (IC 95%, 0,878 - 0,939)] e especificidade [0,886 (IC 95%, 0,855 - 0,912)] para diagnosticar a COVID-19. Ambas as taxas foram substancialmente mais altas em pacientes que receberam uma vacina COVID-19 [0,959 (IC95%, 0,908 - 0,987) e 0,884 (IC95%, 0,819 - 0,932), respectivamente] em comparação com indivíduos não vacinados [0,625 (IC95%, 0,584 - 0,664) e 0,667 (IC95%, 0,629 - 0,704), respectivamente]. As regiões de 650–1800 cm^{-1} e 2300–3900 cm^{-1} foram consideradas as mais importantes no diagnóstico da COVID-19 de acordo com os estudos incluídos. **Conclusão.** A FTIR é uma técnica válida e confiável para detectar a infecção pelo SARS-CoV-2 em diferentes matrizes biológicas (com alta sensibilidade, especificidade e acurácia). Além disso, devido às suas características econômicas (por exemplo, menor custo em comparação com o RT-PCR) e amigáveis ao meio ambiente, a FTIR pode ser implementada na prática clínica como uma ferramenta de triagem para pacientes com suspeita de COVID-19, especialmente em países de baixa renda.

Palavras-chave: SARS-CoV-2, Coronavírus, FTIR, evidência, sensibilidade, especificidade

3.2 INTRODUÇÃO

A pandemia global de COVID-19 desencadeou uma corrida intensiva em busca de métodos diagnósticos precisos e eficazes para a detecção rápida e confiável do vírus SARS-CoV-2, sendo que o diagnóstico atualmente baseia-se essencialmente na detecção do vírus ou nos testes sorológicos (detecção de anticorpos IgG e IgM) ^{219,220}.

Devido à alta sensibilidade e especificidade, a RT-PCR é o padrão ouro para detectar infecções por COVID-19 ²²⁰. Contudo, é um procedimento demorado que requer um instrumento laboratorial sofisticado, pessoal qualificado e logística de transporte de amostras. Assim, ensaios alternativos como os testes sorológicos, foram sendo incorporados na rotina, como uma opção diagnóstica. Entretanto, limitam-se principalmente no seu desempenho (sensibilidade e especificidade) comparado ao padrão-ouro. Desta forma, ainda há uma necessidade não atendida de uma abordagem diagnóstica rápida e de baixo custo, com alta sensibilidade e especificidade, que possa ser usada para testes e monitoramento em larga escala da infecção por COVID-19 ^{219,221}.

Nos últimos anos, um avanço significativo nos métodos de instrumentação e análise espectral resultou em um impacto notável nas aplicações destas técnicas na área da saúde. Dentre as diferentes técnicas analíticas, o infravermelho destaca-se como uma técnica promissora para o diagnóstico e monitoramento desta infecção, por ser rápida, não destrutiva, não invasiva e barata para traçar o perfil das estruturas químicas e físicas de uma ampla gama de amostras ^{222,223}. Desta forma, o objetivo desta revisão sistemática e meta-análise é avaliar o desempenho dessa técnica analítica no diagnóstico de infecção por COVID-19 associado aos principais marcadores identificados em pacientes infectados.

3.3 OBJETIVOS

3.3.1 *Objetivo geral:*

- Avaliar a acurácia da técnica de espectrofotometria de infravermelho no diagnóstico da COVID-19 por meio de uma revisão sistemática com meta-análise.

3.3.2 Objetivos específicos:

- Identificar e reunir estudos que investigaram a aplicação da espectrofotometria de infravermelho no diagnóstico da COVID-19;
- Avaliar a sensibilidade e especificidade da técnica de espectrofotometria de infravermelho em comparação com os métodos de referência para o diagnóstico da COVID-19;
- Realizar uma meta-análise para quantificar a acurácia da espectrofotometria de infravermelho na detecção do SARS-CoV-2;
- Analisar os intervalos de confiança e os fatores que podem influenciar a acurácia da técnica de espectrofotometria de infravermelho no diagnóstico da COVID-19;
- Fornecer recomendações para o uso clínico da espectrofotometria de infravermelho como uma ferramenta de diagnóstico auxiliar na detecção da COVID-19.

3.4 MATERIAL E MÉTODOS

Este estudo foi conduzido de acordo com as recomendações da Colaboração Cochrane e relatado seguindo a declaração *Preferred Reporting Items for a Systematic Review and Meta-analyze of Diagnostic Test Accuracy Studies* (PRISMA-DTA)²²⁴. O protocolo do estudo está registrado no PROSPERO (CRD42022360026). Todas as etapas de seleção da literatura, extração de dados e análise foram realizadas por dois autores de forma independente; um terceiro autor foi convidado para discutir discrepâncias durante reuniões de consenso.

Uma busca sistemática sem limites de tempo e idioma e utilizando descritores relacionados a “*Spectrophotometry*” e “SARS-CoV-2” (combinado com os operadores booleanos AND e OR) foi realizada no Embase e PubMed (atualizado em 5 de setembro de 2023). Também foram realizadas buscas manuais nas listas de referências dos estudos incluídos.

Títulos e resumos dos artigos recuperados foram selecionados para elegibilidade. Os artigos relevantes foram lidos na íntegra e aqueles que atenderam aos seguintes critérios de inclusão tiveram seus dados extraídos:

- (P-população): estudos que avaliam qualquer amostra biológica humana de indivíduos suspeitos de terem COVID-19;
- (I-intervenção): estudos que avaliam FTIR como método de diagnóstico ou monitoramento para detecção de SARS-CoV-2
- (R-referência): estudos que relatam RT-PCR como comparador (ou seja, método de referência)
- (O-resultado): estudos que relatam dados sobre a precisão do teste (por exemplo, sensibilidade, especificidade)

Foram excluídos estudos publicados com caracteres não romanos (por exemplo, árabe, chinês, cirílico) e aqueles que avaliavam infecções virais combinadas.

Os seguintes dados foram extraídos de forma independente por dois pesquisadores usando planilhas do Microsoft Excel: detalhes gerais do estudo (autores, ano de publicação, país de origem, desenho do estudo, tipo de amostra, [sangue, urina, saliva], tamanho da amostra), características dos pacientes (idade, sexo, comorbidades), parâmetros analíticos relatados para os métodos FTIR (por exemplo, região espectral analisada, tipo de amostra, método de análise de dados, método de validação cruzada e biomarcadores espectrais) e resultados de testes diagnósticos (verdadeiro-positivo, VP; verdadeiro negativo, VN; falso positivo, FP; falso negativo, FN, sensibilidade, especificidade, precisão).

O risco de viés foi avaliado de forma independente pelos revisores por meio da ferramenta de avaliação da qualidade do estudo de precisão diagnóstica (QUADAS-2) ²²⁵.

3.4.1 Meta-análise

As meta-análises foram realizadas de acordo com os tipos de amostras biológicas (isto é, por subgrupos). A especificidade, a sensibilidade, a razão de verossimilhança negativa (NLR) e a razão de verossimilhança positiva (PLR) foram medidas com um intervalo de confiança de 95% com base nas taxas de VN, VP, FN, FP obtidas dos estudos incluídos. Também foram construídos gráficos resumidos de características operacionais do receptor (SROC) representados como gráficos de área sob a curva (AUC). A heterogeneidade dos estudos incluídos foi avaliada pelo teste qui-quadrado,

sendo valores de inconsistência (I^2) maiores que 50% considerados como heterogeneidade moderada, e I^2 maiores que 75% definidos como com alta heterogeneidade. A análise de sensibilidade por meio de remoção hipotética de estudos foi realizada quando os valores de I^2 foram superiores a 50%. As análises foram realizadas no Meta-Disc 1.4.7.

3.5 RESULTADOS

Identificamos um total de 1.501 artigos após a remoção de duplicatas, dos quais 1.473 foram excluídos durante a triagem. Dos 28 artigos lidos na íntegra, 11 foram excluídos (ver lista de estudos excluídos com motivos de exclusão no apêndice I). Assim, um total de 17 estudos foram incluídos para análise ²²⁶⁻²⁴³; sendo oito deles incluídos na síntese quantitativa (meta-análise) ^{227,229-233,239,243}, conforme apresentado na **Figura 3.1**.

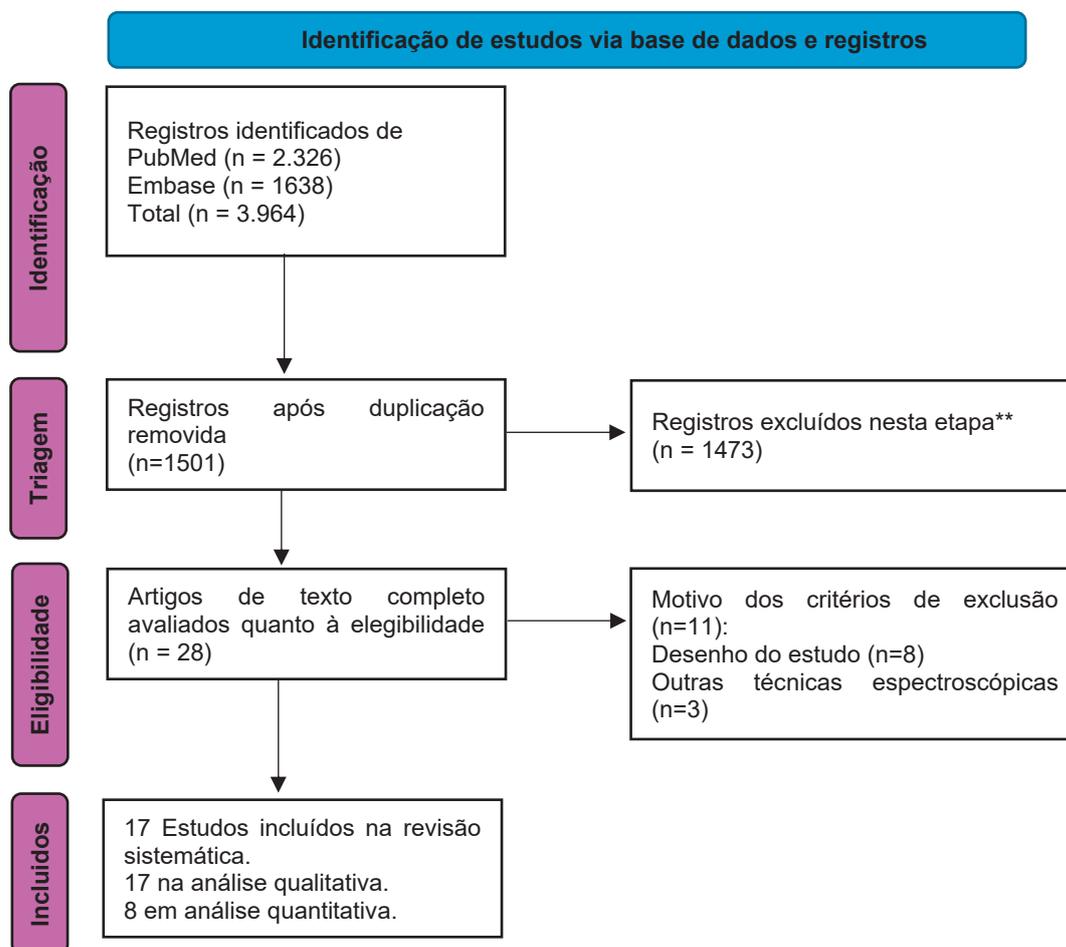


Figura 3.1. Fluxograma da revisão sistemática. **Fonte:** O Autor (2024)

Todos os estudos incluídos foram transversais, publicados entre 2021-2023 e realizados principalmente no Brasil (n=4; 23,5%), seguido pelo México, China e Rússia (n=2 estudos cada; 11,8%). Como pode ser observado na **Tabela 3.1**, o número médio de pacientes diagnosticados com Covid-19 foi de aproximadamente 123. O grupo controle negativo foi composto por cerca de 190 pacientes que testaram negativo para Covid-19. A idade média dos pacientes que foram diagnosticados com Covid-19 e dos do grupo de controle negativo foi em torno de 49 e 50 anos, respectivamente. As principais comorbidades associadas foram artrite reumatoide (11,11%), asma (11,11%), diabetes (27,22%), hipertensão (22,22%), doença pulmonar obstrutiva crônica (16,66%), obesidade (22,22%), doenças cardiovasculares (16,66%), tabagismo (11,11%) e HIV (11,11%). Além disso, também foram relatados tuberculose, colite ulcerativa, câncer gástrico, arterite de Takayasu, carcinoma pulmonar de células não pequenas, hipercortisolismo, carcinoma, hipotireoidismo e doença neurológica ou neuromuscular crônica (5,55%).

Tabela 3.1. Característica basais dos estudos e pacientes incluídos

Referência do estudo	País	Nº de pacientes COVID-19	Nº de pacientes do grupo de controle negativos para SARS-CoV-2	Nº de outros grupos de controle	Idade (anos)	Sexo (% de homens com COVID-19)	Comorbidades (% do total)
Bandeira, 2022 ²²⁶	Brasil	33	49	-----	44 ± 14 -grupo positivo 49 ± 12 -grupo negativo	23.3%	Artrite reumatóide, asma, diabetes, hipertensão arterial sistêmica, doença pulmonar obstrutiva crônica, obesidade ou hipotireoidismo
Barauna, 2021 ²²⁷	Brasil	70	111	-----	45.7 ± 19.2 - grupo positivo 39.6 ± 12.7 - grupo negativo	45.1%	Doença pulmonar crônica n=3 (1,7%) Doença cardiovascular n=9 (5%) Diabetes n=10 (5,5%) Doença neurológica ou neuromuscular crônica n=1 (0,5%)
Calvo-Gomes, 2022 ²³³	México	55	149	102 amostras pré-pandemia 322 amostras de dengue	Range – 20-25	-----	-----
Guleken, 2022 ²³⁶	Turquia	26	11	-----	30.8 ± 6.4	0% (100% mulheres grávidas)	-----
Heino, 2022 ²³¹	Finlândia	558	558	-----	32.0 ± 19.0	52.3%	-----
Karas, 2023 ²³⁵	Rússia	30	30	40 pré-pandemia Pacientes com sepsis	-----	-----	-----
Karthikeyan, 2023 ²⁴²	Índia	314	355	-----	54.9 ± 14.4	61.14%	Diabetes (46,5%) Obesidade 102 (15,2%) Hipertensão (10,4%) Fumar (4,2%) Carcinoma (1,6%) VIH (0,6%) Doença cardiovascular (0,4%) Nenhum (7,8%)
Kitane, 2021 ²³⁷	EUA	100	180	-----	Intervalo – 11-67	40.7% de total dos pacientes	-----
Laird, 2023 ²³⁸	Reino Unido	59	26	-----	56.7 - COVID-19 positivo sintomático	57.6%	Fumar (20%) Tabagismo prévio (21,2%)

Martinez-Cuazit, 2021 ²³⁴	México	255	1209	-----	66,7 - COVID-19 positivo assintomático 53.3 - Controles	62.7%	Obesidade (38%), diabetes (32,9%), hipertensão (27,5%), tabagismo (17,6%), doenças cardiovasculares (1,6%), HIV (1,2%), tuberculose (1,2%), asma (0,8%), nenhum (27,1%)
Nascimento, 2022 ²²⁹	Brasil	138	99	-----	Range - 20 - 97	33.3%	Hipertensão (24,5%) Diabetes (8,4%) Doença pulmonar obstrutiva crônica (6,3%) Obesidade (5,5%)
Nogueira, 2021 ²²⁸	Brasil	151	92	-----	46.2±15.9 - group 1 50.9±18.2 - group 2	36.1% - grupo 1 40.9% - grupo 2	-----
Shlomo, 2022 ²³⁹	Israel	96	201	-----	57,08 ± 18,86 - todos os pacientes; 55,45±17,28 - pacientes positivos; 57,85 ± 19,51 - pacientes negativos	45.3%	-----
Wood, 2021 ²³⁰ Zhang, 2021 ²⁴⁰	Alemanha China	29 41	28 20	-----	-----	-----	-----
Zhao, 2023 ²⁴¹	China	82	78	-----	-----	-----	Colite ulcerativa 3,1% Artrite reumatoide 1,0% Câncer gástrico 1,0% Arterite de Takayasu 1,0% Carcinoma pulmonar de células não pequenas 1,0% Hipercortisolismo 1,0%

Fonte: Elaboração própria

3.5.1 Parâmetros analíticos

A Tabela 3.2 apresenta um resumo dos parâmetros analíticos dos estudos incluídos²²⁶⁻²⁴¹. Os tipos de amostras biológicas foram swab nasofaríngeo (n=4 estudos, 23,53%), saliva (n=6 estudos, 35,30%), plasma (n=1 estudo, 5,88%), soro (n=4 estudos, 23,53%) e ar exalado (n=2 estudos, 11,76%).

Todos os estudos incluídos utilizaram a Transformada de Fourier acoplada ao método de Refletância Total Atenuada (ATR-FTIR)²²⁶⁻²⁴¹, com espectroscopia no infravermelho médio para análise de amostras (400-4000 cm⁻¹). Estudos relataram a região espectral 650–1800 cm⁻¹ do MIR-FTIR como essencial para discriminar entre pacientes com COVID-19 negativo e positivo. Seis estudos (35,30%) relataram adicionalmente regiões de 2.300 a 3.900 cm⁻¹ como importantes para o diagnóstico de infecção (Tabela 3.2).

Em relação às técnicas de validação cruzada de modelos de *machine learning*, o *leave-one-out* foi o mais utilizado (n=7 estudos; 41,17%), seguido pela técnica *K-Fold* (n=4 estudos; 23,53%). Os demais estudos não utilizaram nenhum algoritmo de *machine learning* para prever amostras de COVID-19 e de grupos de controle; portanto, nenhuma técnica de validação cruzada foi empregada (Tabela 3.2).

Tabela 3.2. Parâmetros analíticos reportados dos métodos por MIR-FTIR dos estudos incluídos na revisão sistemática

Referência do estudo	Período de colecta da amostra	Método de referência	Região espectral analisada (cm ⁻¹)	Tipo de amostra	Método de análise de dados	Validação cruzada	Biomarcadores espectrais
Bandeira, 2022 ²²⁶	jul-nov 2020	CLIA ELISA	-----	Soro	PCA PLS-DA	Leave-One- Out	1785–1702 cm ⁻¹ (Impressão digital de glicosilação de IgG); 850–940 cm ⁻¹ , 1015–1170 cm ⁻¹ , 1500–1570 cm ⁻¹ , 1612–1652 cm ⁻¹ , 1695–1735 cm ⁻¹ , 1755–1790 cm ⁻¹ – positivo e mistura de classes 820–890 cm ⁻¹ , 1025–1180 cm ⁻¹ , e 1685–1727 cm ⁻¹ – diferenciação entre classes negativas e positivas;
Barauna, 2021 ²²⁷	jun-set 2020	RT-qPCR	4000-650	Swab nasofaríngeo	PCA GA-LDA	-----	1800–900 cm ⁻¹ – região de impressão digital 1429 cm ⁻¹ – COVID-19 positivo 1220 cm ⁻¹ , 1084 cm ⁻¹ , 1069 cm ⁻¹ , 1041 cm ⁻¹ – COVID-19 negativo
Calvo-Gomes, 2022 ²³³	-----	RT-PCR	5000-550	Soro	MLR OLS PLS-DA por seleção de variáveis	5-Fold	1018.23 cm ⁻¹ , 1045.22 cm ⁻¹ , 1054.87 cm ⁻¹ , 1079.94 cm ⁻¹ , 1643.05 cm ⁻¹ , 1024.01 cm ⁻¹ , 1047.15 cm ⁻¹ , 1068.37 cm ⁻¹ , 1116.58 cm ⁻¹ , 1646.91 cm ⁻¹ , 1025.94 cm ⁻¹ , 1049.08 cm ⁻¹ , 1070.29 cm ⁻¹ , 1135.86 cm ⁻¹ , 1751.04 cm ⁻¹ , 1027.87 cm ⁻¹ , 1051.01 cm ⁻¹ , 1076.08 cm ⁻¹ , 1159.00 cm ⁻¹ , 1752.97 cm ⁻¹ , 1035.58 cm ⁻¹ , 1052.94 cm ⁻¹ , 1078.01 cm ⁻¹ , 1536.98 cm ⁻¹ , 2923.55 cm ⁻¹ – Separação entre infectados e não infectados por SARS-CoV-2
Guleken, 2022 ²³⁶	nov 2020-mai 2021	RT-PCR	4000-600	Soro	PLS-DA	Leave-One- Out	1392 cm ⁻¹ , 1421 cm ⁻¹ , 1460 cm ⁻¹ , 1590 cm ⁻¹ , 2925 cm ⁻¹ , 2954 cm ⁻¹ – Separação entre COVID-19 positivo e negativo
Heino, 2022 ²³¹	jan-nov 2020	PCR	4000-400	Nasopharyngeal swab	PLS-DA	5-K Fold	1800–900 cm ⁻¹ – região de impressão digital 1490–1180 cm ⁻¹ – região de diferenciação dos grupos negativos e positivos
Karas, 2023 ²³⁵	-----	-----	4000-600	Plasma	PCA-LDA	Leave-One- Out	1512 cm ⁻¹ , 1548 cm ⁻¹ , 1614 cm ⁻¹ , 1686 cm ⁻¹ and 1692 cm ⁻¹ – região de positividade da COVID-19 1525 cm ⁻¹ , 1582 cm ⁻¹ , 1611 cm ⁻¹ , 1627 cm ⁻¹ , 1670 cm ⁻¹ – região de negatividade da COVID-19 3340 cm ⁻¹ ; 3288 cm ⁻¹ ; 1645 cm ⁻¹ ; 1745 cm ⁻¹ ; 1536 cm ⁻¹ ; 1452 cm ⁻¹ ; 1241 cm ⁻¹ ; 1078 cm ⁻¹ ; 1029 cm ⁻¹ – região de pacientes recuperados da COVID-19
Karthikeyan, 2023 ²⁴²	maio 2020-jul 2021	RT-PCR	4000-400	Saliva	2DCOS	-----	3050 cm ⁻¹ – 2800 cm ⁻¹ (grupos metil e metileno de cadeias alquil);

									2055 cm ⁻¹ , 2873 cm ⁻¹ (SCN-tiocianato e lipídios em pacientes masculinos com COVID-19); 2873 cm ⁻¹ , 3065 cm ⁻¹ (amida B de proteínas NH); 2727 cm ⁻¹ , 2922 cm ⁻¹ , 2727 cm ⁻¹ , 2969 cm ⁻¹ (lipídios e ácidos graxos); 2858 cm ⁻¹ (DNA e proteínas); 2858 cm ⁻¹ em pacientes COVID-19 mulheres
Kazmer, 2021 ²³²	-----	RT-qPCR	4000-650	Saliva	PLS-DA	Leave-One-Out		1688-1658 cm ⁻¹ (deslocamento para a direita da amida I); 1430 cm ⁻¹ (diminuição alifática/RNA); 1373 cm ⁻¹ (diminuição da flexão do grupo metil); 1124 cm ⁻¹ (diminuição do grupo fosfato do RNA); 1071 cm ⁻¹ (ligação de oxigênio simétrica e ribose); 1016 cm ⁻¹ (diminuição da glicose); 1688-1658 cm ⁻¹ , 1430 cm ⁻¹ , 1373 cm ⁻¹ , 1124 cm ⁻¹ , 1071 cm ⁻¹ , 1016 cm ⁻¹ – Separação entre COVID-19 positivo e negativo	
Kitane, 2021 ²³⁷	-----	RT-PCR	4000-650	Swab nasofaríngeo	PCA PLS-DA	Leave-One-Out		600–1350 cm ⁻¹ , 1500–1700 cm ⁻¹ e 2300–3900 cm ⁻¹ – SARS-CoV-2 região e impressão digital do RNA viral	
Laird, 2023 ²³⁸	-----	PCR	8500-485	Ar exalado	SLR RF	-----		Metanol, etanol e acetaldeído regiões	
Martinez-Cuazit, 2021 ²³⁴	maio 2020- mar 2021	RT-PCR	4000-400	Saliva	MLRM	Leave-One-Out		1700-1600 cm ⁻¹ (região de impressão digital de amida I); 1100-850 cm ⁻¹ (região de impressão digital de RNA); 1560-1464 cm ⁻¹ (Região de impressão digital IgG+); 1076 cm ⁻¹ , 1037 cm ⁻¹ , 1028 cm ⁻¹ , 992 cm ⁻¹ , 986 cm ⁻¹ , 968 cm ⁻¹ , 924 cm ⁻¹ , 886 cm ⁻¹ – Separação entre COVID-19 positivo e negativo	
Nascimento, 2022 ²²⁹	-----	RT-qPCR	4000-650	Saliva	PLS <i>Unsupervised Random Forest</i> (URF) PCoA GA-LDA SPA-LDA; GA-LDA; PLS-DA; PSO-PLS-DA	<i>K Fold Venetian blinds</i>		1450-1650 cm ⁻¹ and 1100-1050 cm ⁻¹ - 1707-1792 cm ⁻¹ (lipídeos); 1650 cm ⁻¹ , 1429 cm ⁻¹ , 1400 cm ⁻¹ , 1220 cm ⁻¹ , 1200 cm ⁻¹ , 1155 cm ⁻¹ , 1069 cm ⁻¹ , 950 cm ⁻¹ - diferenças entre grupos positivos e negativos de COVID-19	
Nogueira, 2021 ²²⁸	may-jul 2020	RT-PCR	4000-650	Swab nasofaríngeo	PLS-cosine KNN	5-K Fold		2941 cm ⁻¹ , 2913 cm ⁻¹ , 2878-2880 cm ⁻¹ , 2855 cm ⁻¹ , 2838 cm ⁻¹ , 2820 cm ⁻¹ , 1552-1550 cm ⁻¹ , 1524 cm ⁻¹ , 1497 cm ⁻¹ , 1442 cm ⁻¹ , 1359-1358 cm ⁻¹ , 1299 cm ⁻¹ , 1146 cm ⁻¹ , 1088 cm ⁻¹ , 1032 cm ⁻¹ , 921-924 cm ⁻¹ , 896 cm ⁻¹ , 867 cm ⁻¹ – Principais modos vibracionais	

									presentes na região da impressão digital entre 650–1800 cm^{-1} and 2800–3000 cm^{-1} 3404.36 cm^{-1} , 2808.44 cm^{-1} and 3230.31 cm^{-1}
Shlomo, 2022 ²³⁹	fev-abr 2021	PCR	-----	Ar exalado	AI-based algorithm developed and validated	-----			
Wood, 2021 ²³⁰	-----	RT-qPCR	4000-800	Saliva	PLS-DA	MCDCV			1740 cm^{-1} , 1690 cm^{-1} , 1657 cm^{-1} , 1547 cm^{-1} , 1517 cm^{-1} , 1464 cm^{-1} , 1382 cm^{-1} , 1341 cm^{-1} , 1235 cm^{-1} , 1124 cm^{-1} , 1089 cm^{-1} , 996 cm^{-1} , 967 cm^{-1} , 934 cm^{-1}
Zhang, 2021 ²⁴⁰	feb-mar 2020	RT-PCR	4000-650	Serum	PLS-DA	Leave-One-Out			2927.5 cm^{-1} – 2925.6 cm^{-1} ; 2871.6 cm^{-1} – 2871.8 cm^{-1} ; 2853.2 cm^{-1} – 2853.0 cm^{-1} ; 1691.9 cm^{-1} – 1691.8 cm^{-1} ; 1634.0 cm^{-1} – 1632.4 cm^{-1} ; 1541.0 cm^{-1} – 1545.0 cm^{-1} ; 1469.0 cm^{-1} – 1468.9 cm^{-1} ; 1439.5 cm^{-1} – 1439.9 cm^{-1} ; 1367.9 cm^{-1} – 1367.7 cm^{-1} ; 1313.5 cm^{-1} – 1314.3 cm^{-1} ; 1104.8 cm^{-1} – 1105.0 cm^{-1} ; 1080.2 cm^{-1} – 1079.3 cm^{-1}
Zhao, 2023 ²⁴¹	-----	RT-PCR	4000-800	Saliva	GWO-SVM	Leave-One-Out			1300 – 800 cm^{-1}

Nota: AI – inteligência artificial; CLIA – imunensaio quimioluminescente; DA – análise discriminante; ELISA – ensaio imunoenzimático; GA – algoritmo genético; GWO-SVM – Regressão de Support Vector Machine aprimorada com otimizador de grey wolf; LDA – análise discriminante linear; LR – razão de verossimilhança; ML – Machine learning; MCDCV – Monte-Carlo double; PCA – análise de componentes principais; MLR: Regressão logística multinomial; PLS – regressão por mínimos quadrados parciais; OLS – regressão por mínimos quadrados ordinários; RF – Random Forest; RT-PCR - Reação em Cadeia da Polimerase com Transcrição Reversa; – SLR – regressão linear esparsa. **Fonte:** O Autor (2024)

3.5.2 Teste de acurácia diagnóstica

Meta-análises comparando o desempenho de MIR-FTIR versus RT-PCR foram realizadas, incluindo oito estudos que relataram adequadamente dados sobre parâmetros de acurácia diagnóstica (ver **Tabela 3.3**)^{227,229–233,239,243}. A sensibilidade e especificidade gerais (considerando todas as amostras de estudos) do MIR-FTIR foram 0,731 (IC 95%, 0,701-0,760, $I^2 = 93,7\%$) e 0,761 (IC 95%, 0,734-0,786, $I^2 = 95,1\%$), respectivamente. O valor de acurácia diagnóstica (AUC ROC) foi de 0,95. A análise de sensibilidade com a hipotética retirada do estudo de Heino²³¹ levou a maiores valores de sensibilidade, especificidade e acurácia de 0,912 (IC 95%, 0,878-0,939, $I^2 = 49,3\%$), 0,886 (IC 95%, 0,855 - 0,912, $I^2 = 90,2\%$) e AUC ROC=0,97, respectivamente.

Os resultados da meta-análise incluindo apenas amostras de saliva foram de 0,899 (IC 95%, 0,852-0,935), $I^2 = 45,3\%$ para sensibilidade e 0,737 (IC 95%, 0,664-0,801), $I^2 = 0,0\%$ para especificidade de MIR- Método FTIR. O valor de acurácia diagnóstica foi de 0,73. Nenhuma outra meta-análise de subgrupos de acordo com o tipo de amostra foi possível, dado o número limitado de estudos disponíveis.

Para três estudos^{230,232,239} que incluíram pacientes vacinados, a meta-análise mostrou alta sensibilidade [0,959 (IC 95%, 0,908-0,987), $I^2 = 41,9\%$] e especificidade [0,884 (IC 95%, 0,819-0,819- 0,932), $I^2 = 91,4\%$]. O valor de acurácia diagnóstica (AUC ROC) foi de 0,99. Por outro lado, as taxas de sensibilidade e especificidade de estudos envolvendo pacientes não vacinados ($n = 2$)^{227,231} foram mais baixas, de 0,625 (IC 95%, 0,584-0,664), $I^2 = 91,8\%$ e 0,667 (IC 95%, 0,629-0,704), $I^2 = 94,1\%$, respectivamente, com AUC ROC de 0,96.

Os estudos foram julgados como de qualidade metodológica moderada, conforme mostrado na **Figura 3.2**. Um dos principais aspectos considerados de alto risco foi relacionado a não evitar o desenho caso-controle, e os resultados dos testes de índice foram interpretados sem o conhecimento dos resultados do padrão de referência, impactando os domínios de seleção de pacientes e teste de índice, respectivamente.

Tabela 3.3. Meta-análise dos parâmetros de acurácia para as diferentes técnicas diagnósticas

Método	Tipo de amostra, subgrupo	Número de estudos incluídos	Sensibilidade (95% IC)	Especificidade (95% IC)	PLR (95% IC)	NLR (95% IC)
MIR-FTIR	All studies	8 ^{227,229-233,239,243}	0.731 (0.701-0.760) I ² = 93.7%	0.761 (0.734 -0.786) I ² = 95.1%	6.364 (3.303-12.259) I ² = 92.3%	0.119(0.047 - 0.304) I ² = 92.9%
MIR-FTIR	Análise de sensibilidade: todos os estudos, exceto Heino ²³¹	7 ^{227,229,230,232,233,239,243}	0.912 (0.878 – 0.939) I ² = 49.3%	0.886 (0.855 – 0.912) I ² = 90.2%	7.935 (3.948 – 15.949) I ² = 84.3%	0.105 (0.059 – 0.188) I ² = 51.4.3%
MIR-FTIR	Análise de subgrupos de acordo com tipo de amostra: saliva	3 ^{229,230,232}	0.899 (0.852-0.935) I ² = 45.3%	0.737 (0.664-0.801) I ² = 0.0%	3.33 (2.59-4.30) I ² =0.0%	0.12 (0.06-0.25) I ² =47.4%
MIR-FTIR	Análise de subgrupos de acordo com pacientes: não vacinados	2 ^{227,231}	0.625 (0.584 – 0.664) I ² = 91.8%	0.667 (0.629 – 0.704) I ² = 94.1%	3.625 (0.777 – 16.913) I ² = 94.6%	0.220 (0.019 – 2.533) I ² = 85.0%
MIR-FTIR	Análise de subgrupos de acordo com pacientes: vacinados	3 ^{230,232,239}	0.959 (0.908 - 0.987) I ² = 41.9%	0.884 (0.819 – 0.932) I ² = 91.4%	7.112 (2.092-24.182) I ² = 78.6%	0.063 (0.028-0.144) I ² = 0.0%

Nota: MIR-FTIR: espectroscopia de infravermelho com transformada de Fourier no infravermelho médio, intervalo de confiança; IC: intervalo de confiança; RPL: Razão de Verossimilhança Positiva, RNL: Razão de Verossimilhança Negativa. **Fonte:** O Autor (2024)

Study ID	Risk of Bias				Applicability Concerns			Legend
	Patient Selection	Index test	Reference Standard	Flow and Timing	Patient Selection	Index test	Reference Standard	
BANDEIRA, 2022	!	!	!	!	+	+	+	<ul style="list-style-type: none"> + Low risk ! Some concerns - High risk
BARAUNA, 2021	+	+	+	+	+	+	+	
CALVO-GOMES, 2022	-	-	+	+	!	+	+	
GULEKEN, 2022	-	-	+	+	!	+	+	
HEINO, 2021	-	-	+	-	+	+	+	
KARAS, 2022	-	-	+	!	!	+	!	
KARTHIKEYAN, 2023	-	-	+	!	+	+	+	
KAZMER, 2021	-	+	+	-	+	+	+	
KITANE, 2021	-	-	+	+	+	+	+	
LAIRD, 2023	-	-	+	!	+	+	+	
MARTINEZ-CUAZITL, 2021	-	-	+	-	+	+	+	
NASCIMENTO, 2022	+	+	+	+	+	+	+	
NOGUEIRA, 2022	+	+	+	+	+	+	+	
SHLOMO, 2022	+	+	+	+	!	+	+	
WOOD, 2021	-	-	+	!	+	+	+	
ZHANG, 2021	-	-	+	!	+	+	+	
ZHAO, 2023	-	-	+	!	+	+	+	

Figura 3.2. Qualidade metodológica dos estudos incluídos seguindo o QUADAS-2.
Fonte: O Autor (2024)

3.6 DISCUSSÃO

Acredita-se que esta seja a primeira revisão sistemática que resume as evidências disponíveis sobre a aplicação de técnicas infravermelhas para o diagnóstico de COVID-19, envolvendo diferentes amostras como saliva, swab nasofaríngeo, sangue total, plasma e soro. Foi possível avaliar qualitativamente e quantitativamente os principais parâmetros analíticos utilizados nas diversas técnicas de infravermelho empregadas.

O desenvolvimento de um novo método de diagnóstico utilizando espectroscopia infravermelha envolve a análise dos espectros de amostras do grupo de pacientes doentes e dos espectros de amostras do grupo saudável por meio de modelos quimiométricos, também conhecidos como modelos de *machine learning*. A

partir desses modelos de classificação preditiva, os dois grupos amostrais são previstos e, com base na matriz de confusão obtida, é possível calcular a acurácia diagnóstica, sensibilidade e especificidade do método diagnóstico proposto ^{241,244,245}.

Oito dos estudos incluídos empregaram o modelo quimiométrico PLS-DA na análise de seus dados espectrais MIR-FTIR para prever o diagnóstico de COVID-19. É importante notar que o PLS-DA é considerado o modelo padrão ouro para análise de dados quimiométricos através de espectroscopia (infravermelho, espectrometria de massa, ressonância magnética nuclear), incluindo ômicas (por exemplo, metabolômica, genômica, transcriptômica etc.) e PLS-DA é fortemente recomendada por especialistas na área ^{246,247}. Uma recente revisão sistemática publicada por Mendez (2019) avaliou estudos publicados entre 1990 e 2018 disponíveis na base de dados *Web of Science* e demonstrou uma tendência cada vez maior de estudos citando o algoritmo PLS-DA (n=2.242 citações), enquanto outros algoritmos de *machine learning* (por exemplo, *Random Forest*, Redes Neurais Artificiais) foram citados com menos frequência (n = 500 citações no total) ²⁴⁸. Além disso, outro estudo recente também conduzido por Mendez (2019) avaliou o desempenho preditivo de oito modelos diferentes de *machine learning* (PLS-DA LR, RF, regressão de componentes principais (PCR), kernel de função de base radial, SVM, ANN e rede neural artificial) usando dez conjuntos de dados ômicos diferentes disponíveis publicamente, descobriu que o algoritmo PLS-DA teve o melhor desempenho preditivo ²⁴⁹. Esse alto desempenho preditivo ocorre porque o algoritmo PLS-DA pode fazer previsões de dados multidimensionais em um espaço de dimensões menores, que são chamadas de variáveis latentes, que descrevem a variância entre os dados de saída (por exemplo, classes de amostra) e dos dados de entrada. dados (por exemplo, números de onda no MIR-FTIR), e isso antes mesmo da regressão da variável dependente. Esta abordagem permite analisar múltiplos conjuntos de dados com mais variáveis do que amostras, sem recorrer a variáveis de pré-filtragem. Uma segunda vantagem do PLS-DA é que suas variáveis latentes reduzem os problemas de multicolinearidade existentes entre as diversas variáveis (por exemplo, números de onda no MIR-FTIR) presentes no conjunto de dados analisado ^{249,250}. A terceira vantagem é que, uma vez otimizado, o PLS-DA é reduzido a uma regressão linear tradicional, permitindo determinar quais biomarcadores espectrais são muito importantes para o estudo. Outros modelos de *machine learning* (por exemplo, redes neurais artificiais ou

aprendizado profundo) exigem um grande volume de amostras de treinamento para obter melhores desempenhos preditivos ^{251–254}.

Em todos os estudos incluídos ^{226–243}, a região espectral de 650–1800 cm^{-1} na região do infravermelho médio foi a principal responsável pela discriminação entre indivíduos negativos e positivos para COVID-19. É importante destacar que esta região espectral é tradicionalmente conhecida como região de impressão digital da molécula de RNA do SARS-CoV-2, pois é a região onde estão localizados grupos funcionais específicos do RNA do SARS-CoV-2, por exemplo (i) Estrutura de Fosfato: A forte banda de absorção em torno de 1220-1080 cm^{-1} é devida às vibrações de estiramento dos grupos fosfato (PO_2^-) na estrutura de RNA; (ii) Vibrações dos anéis de açúcar: Bandas em torno de 1000-900 cm^{-1} estão relacionadas às vibrações dos anéis de açúcar na molécula de RNA; (iii) Bases Nitrogênicas: Bandas na região de 1700-1600 cm^{-1} são frequentemente associadas às vibrações das bases nitrogenadas (adenina, guanina, citosina e uracila) no RNA.

Seis estudos na presente revisão sistemática também relataram a região espectral de 2.300–3.900 cm^{-1} como sendo mais crucial para o diagnóstico de COVID-19. É importante notar que esta região espectral, precisamente na faixa de 3300-3400 cm^{-1} , está associada a vibrações de estiramento OH e NH, que são indicativas de interações de ligações de hidrogênio dentro da molécula de RNA.

Nossa meta-análise revelou que o diagnóstico de COVID-19 por espectroscopia infravermelha com amostras de saliva demonstrou maior sensibilidade analítica quando comparado a outros tipos de amostras (por exemplo, ar exalado e swab nasofaríngeo). Sabe-se que a escolha do tipo de amostra pode impactar na sensibilidade e especificidade do teste para COVID-19. Embora os esfregaços nasofaríngeos sejam frequentemente considerados o padrão ouro para o diagnóstico de COVID-19 por métodos de diagnóstico tradicionais (RT-PCR, teste de antígeno e detecção de anticorpos), as amostras de saliva apresentam maior sensibilidade de detecção do que as amostras de esfregaços nasofaríngeos quando o diagnóstico de COVID-19 é feito usando o método de espectroscopia infravermelha ^{229,230,232}. Resultados semelhantes foram encontrados em uma recente revisão sistemática conduzida por nosso grupo de pesquisa, onde os resultados da meta-análise também demonstraram alta sensibilidade da técnica de PCR na detecção de SARS-CoV-2 ao utilizar amostras de saliva ²⁵⁵. Quatro pontos cruciais podem explicar esta alta sensibilidade das amostras de saliva no diagnóstico de COVID-19 usando MIR-FTIR:

(i) Rica composição molecular: A saliva é um fluido biológico complexo que contém uma ampla gama de moléculas, incluindo proteínas, ácidos nucleicos, lipídios e metabólitos. No caso do COVID-19, a presença do vírus e das biomoléculas associadas pode ser detectada na saliva. Esta rica composição molecular torna as amostras de saliva adequadas para análise MIR-FTIR; (ii) Coleta não invasiva: As amostras de saliva são fáceis de coletar de forma não invasiva, o que é particularmente vantajoso para o conforto e adesão do paciente. Esta facilidade de coleta torna-o uma escolha prática para triagem e monitoramento em massa, pois elimina a necessidade de métodos de amostragem desconfortáveis ou invasivos, como swab nasofaríngeo; (iii) Biomarcadores virais: Quando um indivíduo é infectado com COVID-19, o vírus é disseminado para vários fluidos corporais, incluindo a saliva. O MIR-FTIR pode detectar alterações moleculares específicas ou biomarcadores associados à presença do vírus ou à resposta imunológica do hospedeiro. Isto torna-o uma ferramenta sensível para identificar a COVID-19 em amostras de saliva; (iv) Preparação mínima da amostra: Em comparação com algumas outras técnicas de diagnóstico, o MIR-FTIR requer uma preparação mínima da amostra. Isto simplifica o processo de teste e reduz o risco de contaminação, tornando-o eficiente e adequado para testes em larga escala ^{229,230,232}.

No presente estudo, os resultados da meta-análise mostraram que o diagnóstico de COVID-19 por FT-IR apresentou maior sensibilidade e especificidade em estudos envolvendo pacientes vacinados do que naqueles envolvendo não vacinados. Considerando que o FTIR é uma técnica utilizada para identificar e caracterizar moléculas com base nos seus padrões de absorção e vibração no espectro infravermelho, o impacto da vacinação contra a COVID-19 na precisão deste método diagnóstico pode variar dependendo de diversos fatores, tais como: (i) Alterações moleculares: o diagnóstico de COVID-19 por espectroscopia FTIR depende da detecção de alterações moleculares específicas associadas à presença do vírus SARS-CoV-2 nas secreções respiratórias. A vacinação não afeta diretamente a presença do vírus, mas ajuda o sistema imunológico a responder a ele. Portanto, as alterações moleculares em indivíduos vacinados e não vacinados podem diferir significativamente; (ii) carga viral: A quantidade de vírus nas amostras respiratórias (carga viral) pode variar de pessoa para pessoa. A frequência da infecção pode não ser necessariamente mais elevada em indivíduos não vacinados, especialmente se ambos os grupos estiverem infectados com estirpes de vírus semelhantes; (iii) O

momento da espectroscopia FTIR em relação à vacinação e ao início dos sintomas é crucial. A resposta imunológica do organismo à vacina pode variar e a janela para um diagnóstico preciso pode ser diferente para indivíduos vacinados e não vacinados. Portanto, a alta sensibilidade e especificidade do método FTIR na detecção de SARS-CoV-2 observada em nosso estudo pode ser atribuída a pelo menos esses três importantes fatores mencionados ^{229,230,232,256,257}.

Embora os resultados da meta-análise tenham comprovado boa precisão no uso da técnica de espectroscopia infravermelha para o diagnóstico de COVID-19, o presente estudo apresenta diversas limitações. Os estudos incluídos apresentaram diferenças em termos de validação externa, risco de viés e tamanho da amostra. Ainda, como limitações há as altas taxas de heterogeneidade entre os estudos, provavelmente causadas por diferenças nas características dos pacientes, no tipo de amostra utilizada, no método utilizado, no algoritmo de *machine learning* utilizado para analisar os dados espectrais MIR-FTIR e no fato de que alguns estudos utilizaram vacinados e pacientes não vacinados. Contudo, visando a contornar essas limitações, foram conduzidas diferentes meta-análises, considerando subgrupos de estudos (análise de sensibilidade), que não demonstraram diferenças significativas em relação à análise original. Os estudos incluídos na revisão sistemática apresentaram pouca preocupação quanto à aplicabilidade dos métodos na prática clínica e foram julgados de qualidade metodológica moderada.

3.7 CONCLUSÃO

Nesta revisão sistemática, a técnica FTIR mostrou-se promissora na detecção da infecção por SARS-CoV-2 em diferentes matrizes biológicas com alta sensibilidade [0,912 (IC 95%, 0,878 – 0,939)] e especificidade [IC 95%, 0,886 (0,855 – 0,912)]. Todos os estudos apontaram as regiões espectrais entre 650–1800 cm^{-1} e 2300–3900 cm^{-1} como as mais importantes na diferenciação entre pacientes negativos e positivos para COVID-19, correspondendo à região da impressão digital do RNA viral.

Dadas as inúmeras vantagens da técnica infravermelha, como baixo custo em relação ao RT-PCR, análise rápida, natureza não destrutiva, preparo mínimo da amostra, ausência de solventes na análise e respeito ao meio ambiente, facilidade de uso sem necessidade de profissionais altamente qualificados, e alta precisão e repetibilidade, as evidências encontradas nesta revisão sistemática com meta-análise sugerem que esta técnica poderia ser facilmente implementada na prática clínica como uma ferramenta de triagem em massa para pacientes com COVID-19. Isto é particularmente relevante em países de baixo rendimento onde os recursos são limitados.

4 CAPÍTULO IV - DESENVOLVIMENTO DE MODELOS PREDITIVOS E IDENTIFICAÇÃO DE BIOMARCADORES PROGNÓSTICOS EM COVID-19, HIV E TUBERCULOSE POR MEIO DE INTELIGÊNCIA ARTIFICIAL E MACHINE LEARNING

Publicado em:

1.Cobre AF, Stremel DP, Noletto GR, Fachi MM, Surek M, Wiens A, Tonin FS, Pontarolo R. Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? *Comput Biol Med.* 2021 Jul;134:104531. doi: 10.1016/j.combiomed.2021.104531.

2.Cobre AF, MORAIS A, Selege F, Stremel DP, Wiens A, Ferreira LM, Tonin F; Pontarolo R. Use of biochemical tests and machine learning in the search for potential diagnostic biomarkers of COVID-19, HIV/AIDS, and Pulmonary Tuberculosis. *Journal of The Brazilian Chemical Society*, v. 1, p. 1, 2024. <https://dx.doi.org/10.21577/0103-5053.20240020>

4.1 RESUMO

Objetivo: Este IV capítulo da tese abrangente teve como objetivo implementar e avaliar vários modelos de *machine learning* para prever o diagnóstico e a gravidade da COVID-19, juntamente com outras doenças infecciosas, como HIV/AIDS, tuberculose (TB) e suas coinfeções. **Métodos:** Os dados de várias fontes, incluindo registros hospitalares e testes farmacêuticos, foram analisados usando técnicas exploratórias como análise de componentes principais (PCA) para identificar padrões e variáveis importantes. Modelos de *machine learning*, incluindo redes neurais artificiais (RNA), árvores de decisão (AD), análise discriminante de mínimos quadrados parciais (PLS-DA), algoritmo de vizinhos mais próximos (KNN), *Light Gradient Boosting Machine*, *Extreme Gradient Boosting*, *Gradient Boosting Classifier*, *Ada Boost Classifier* e *Logistic Regression* foram treinados e validados para previsão precisa de diagnóstico. **Resultados:** O estudo abrangeu um total de 3,3 milhões de amostras de pacientes em vários conjuntos de dados. Para a previsão de diagnóstico de COVID-19, os modelos ANN, DT, PLS-DA, KNN, *Light Gradient Boosting Machine*, *Extreme Gradient Boosting*, *Gradient Boosting Classifier*, *Ada Boost Classifier* e *Logistic Regression* alcançaram precisões superiores a 84%, com variáveis importantes incluindo hiperferritinemia, hipocalcemia e vários sintomas clínicos. Para uma previsão mais ampla de doenças infecciosas, o PLS-DA apresentou alto desempenho, com precisões de 94%, 97%, 95% e 96% para COVID-19, HIV/AIDS, TB e coinfeção HIV/TB, respectivamente. Biomarcadores como cálcio, desidrogenase láctica e diferentes contagens de células sanguíneas foram associados a várias doenças. **Conclusão:** O estudo destaca a eficácia dos modelos de *machine learning* no diagnóstico da COVID-19 e de outras doenças infecciosas com base em diversos fatores clínicos e demográficos. Essas descobertas oferecem *insights* valiosos para melhorar os processos de triagem e diagnóstico em ambientes de saúde e destacam o potencial do *machine learning* no gerenciamento de saúde pública.

Palavras-chave: Exames de sangue; COVID 19; Diagnóstico; *Machine learning*; Gravidade; Teste de urina

4.2 INTRODUÇÃO

A pandemia de COVID-19, juntamente com as persistentes ameaças representadas pelo HIV e pela tuberculose, tem gerado desafios significativos para a saúde pública global ^{258–260}. Diante da complexidade dessas doenças e da necessidade de abordagens mais eficazes de prevenção, diagnóstico e tratamento, a aplicação de técnicas avançadas de inteligência artificial (IA) e *machine learning* tem despertado interesse crescente na comunidade científica e médica ^{261–263}.

No contexto do atual cenário de saúde, o desenvolvimento de modelos preditivos e a identificação de biomarcadores prognósticos têm o potencial de revolucionar a maneira como abordamos o manejo clínico e a tomada de decisão terapêutica em pacientes afetados por essas doenças ^{264,265}. A integração de dados clínicos e laboratoriais com algoritmos de IA e *machine learning* permite a análise abrangente de múltiplos fatores de risco, sintomas, respostas terapêuticas e desfechos, proporcionando *insights* para a prática clínica ^{264,265}.

Este capítulo tem como objetivo explorar o papel da inteligência artificial e do *machine learning* no desenvolvimento de modelos preditivos e na identificação de biomarcadores prognósticos em pacientes com COVID-19, HIV e tuberculose. Por meio de uma abordagem interdisciplinar que integra conhecimentos da medicina, estatística, ciência de dados e computação, busca-se elucidar padrões e correlações ocultas nos dados que podem fornecer informações cruciais para aprimorar a estratificação de risco, personalizar o tratamento e melhorar os desfechos clínicos.

Ao analisar as mais recentes descobertas e abordagens metodológicas nesta área emergente de pesquisa, este capítulo visa contribuir para o avanço do conhecimento científico e o desenvolvimento de ferramentas de apoio à decisão clínica mais precisas e eficientes

4.3 OBJETIVOS

4.3.1 Objetivo geral

Utilizar técnicas avançadas de inteligência artificial e *machine learning* para analisar dados clínicos e laboratoriais de pacientes diagnosticados com COVID-19, HIV e tuberculose, com o propósito de desenvolver modelos preditivos para

diagnóstico preciso das doenças e identificação de potenciais biomarcadores prognósticos

4.3.2 *Objetivos específicos*

- Coletar e compilar conjuntos de dados abrangentes de pacientes diagnosticados com COVID-19, HIV e tuberculose, incluindo informações clínicas e resultados laboratoriais relevantes;
- Realizar uma análise exploratória detalhada dos dados coletados para identificar padrões, correlações e características distintas associadas a cada doença;
- Desenvolver e implementar algoritmos de inteligência artificial e *machine learning* adequados para criar modelos preditivos de diagnóstico, utilizando os dados clínicos e laboratoriais disponíveis;
- Avaliar a eficácia dos modelos preditivos desenvolvidos, utilizando métricas de desempenho apropriadas, como acurácia, sensibilidade e especificidade;
- Identificar os biomarcadores potenciais associados ao prognóstico das doenças estudadas, utilizando abordagens estatísticas e de *machine learning* para seleção e validação de características relevantes;
- Gerar relatórios e *insights* interpretáveis para profissionais de saúde, destacando os resultados dos modelos preditivos e biomarcadores prognósticos identificados, facilitando sua aplicação clínica.

4.4 MATERIAL E MÉTODOS

Neste capítulo IV da tese foram realizados três estudos independentes que resultou em três artigos científicos. Portanto o presente capítulo foi subdividido em três tópicos:

- (i) Análise de dados dos exames bioquímicos, hematológicos e de urinálise de pacientes COVID-19 atendidos no Hospital Israelita Albert Einstein (Brasil) visando a predição do diagnóstico e investigação de potenciais Biomarcadores prognósticos.

- (ii) Análise de dados dos exames bioquímicos e hematológicos pacientes COVID-19, HIV, Tuberculose e co-infectados HIV/TB atendidos em um Hospital Regional de referência do Norte de Moçambique (Hospital Geral de Marrere, Província de Nampula) visando predição do diagnóstico e investigação de Biomarcadores associados a essas doenças.
- (iii) Predição de diagnóstico de COVID-19 usando dados clínicos de pacientes COVID-19 atendidos na rede de Farmácia *Drugstore* distribuída em todo território Nacional Brasileiro.

4.4.1 Estudo I: Análise de dados dos exames bioquímicos, hematológicos e de urinálise de pacientes COVID-19 atendidos no Hospital Israelita Albert Einstein (Brasil) visando a predição do diagnóstico e investigação de potenciais Biomarcadores prognósticos.

4.4.1.1 Fluxograma do estudo e conjunto de dados e pacientes

Na Figura 4.1 é mostrado o fluxograma usado para o desenvolvimento do estudo I. Dados da plataforma pública Kaggle²⁶⁶ sobre indivíduos que realizaram exame de reação em cadeia da polimerase com transcrição reversa (RT-PCR) para detectar infecção por síndrome respiratória aguda grave por coronavírus 2 (SARS-Cov-2) no hospital Israelita Albert Einstein (São Paulo, Brasil) foram coletados. Independentemente do resultado do RT-PCR (positivo ou negativo para COVID-19), os pacientes foram incluídos para análise quando apresentavam dados de parâmetros bioquímicos, hematológicos e urinários (**Tabela 4.1**). Dois subgrupos de amostras foram criados de acordo com os resultados do RT-PCR: (i) 5.643 amostras de pacientes, representando resultados negativos (n = 5.086) e positivos (n = 557); (ii) 557 amostras positivas de pacientes ambulatoriais assintomáticos e pacientes com COVID-19 grave, internados em unidades de terapia intensiva. Como a infecção por SARS-Cov-2 se assemelha a outras doenças respiratórias, para minimizar a possibilidade de obtenção de amostras falso-positivas, os pacientes que testaram positivo para pelo menos um outro vírus ou bactéria respiratória (ver **Tabela 4.1**) foram excluídos das análises.

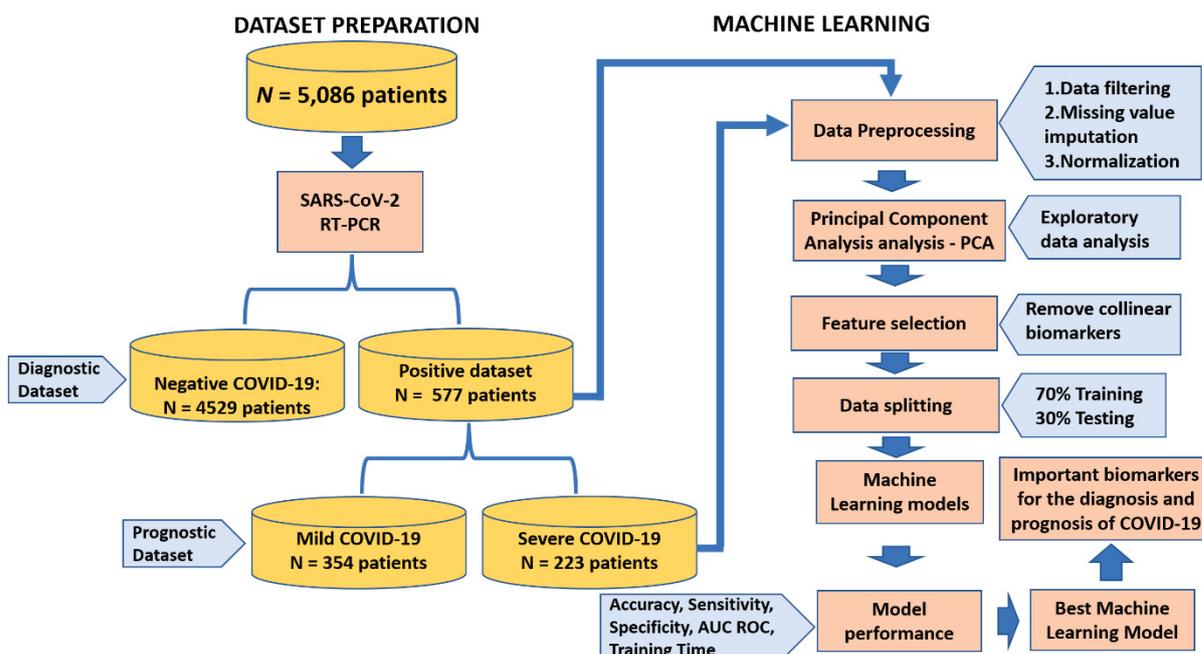


Figura 4.1. Fluxograma do estudo I: análise de dados dos exames bioquímicos, hematológicos e de urinálise de pacientes COVID-19 atendidos no Hospital Israelita Albert Einstein (Brasil) visando a predição do diagnóstico e investigação de potenciais Biomarcadores prognósticos. **Fonte:** O Autor (2024)

Tabela 4.1. Exames Bioquímicos, hematológicos, de urina, virológicos e bacteriológicos usados para o desenvolvimento de modelos de *Machine learning* para a predição de diagnóstico e prognóstico de COVID-19.

Exames	Descrição
Bioquímicos	Glicose sérica, uréia, proteína C reativa, creatinina, potássio, sódio, alanina transaminase, aspartato transaminase, gama-glutamilttransferase, bilirrubina total, bilirrubina direta, bilirrubina indireta, fosfatase alcalina, pH ionizado, sangue, análise de magnésio, HCO ₃ (sangue venoso análise de gases), lactato desidrogenase, creatina fosfoquinase, ferritina, ácido láctico arterial, dosagem de lipase, HCO ₃ (gasometria arterial), fósforo, pCO ₂ (gasometria venosa), saturação de Hb (gasometria venosa), excesso de base (gasometria venosa). gasometria arterial), pO ₂ (gasometria venosa), CO ₂ total (gasometria venosa), saturação de Hb (gasometria arterial), pCO ₂ (gasometria arterial), excesso de base (gasometria arterial), pH (gasometria arterial). gasometria arterial), CO ₂ total (gasometria arterial), pO ₂ (gasometria arterial), FiO ₂ arterial e ctO ₂ (gasometria arterial).
Hematológicos	Hematócrito, Hemoglobina, Plaquetas, Volume médio de plaquetas, Glóbulos vermelhos, Linfócitos, Concentração média de hemoglobina corpuscular, Leucócitos, Basófilos, Hemoglobina corpuscular média, Eosinófilos, Volume corpuscular médio, Monócitos, Largura de distribuição de glóbulos vermelhos
De urina	pH da urina, neutrófilos segmentados, promielócitos, metamielócitos e mieloblastos e razão normalizada internacional (INR).

Viológicos	Vírus sincicial respiratório, influenza A, influenza B, parainfluenza 1, coronavírus NL63, rinovírus/enterovírus, coronavírus HKU1, parainfluenza 3, adenovírus, parainfluenza 4, coronavírus 229E, coronavírus OC43, influenza A H1N1, influenza, teste H1N1 e teste rápido de influenza A.
Bacteriológicos	<i>Mycoplasma pneumoniae</i> , <i>Chlamydomphila pneumoniae</i> e <i>Streptococcus A</i> .

Fonte: O Autor (2024)

4.4.1.2 Pré-processamento de dados para Machine learning

O pré-processamento de dados é uma etapa importante para a análise de dados usando algoritmos de *machine learning*, e refere-se à técnica de preparação (ou seja, limpeza e organização) dos dados brutos para torná-los adequados (ou seja, legíveis) para construção e treinando modelos baseados em ML^{267,268}. Neste estudo, ambos os conjuntos de dados COVID-19 (ou seja, diagnóstico e doença gravidade) foram submetidos a diferentes métodos de pré-processamento visando selecionando aquele que melhor se ajusta aos dados: (i) Imputação: dados faltantes foram substituídos pelos valores medianos das colunas; (ii) Transformação: *absolute value*, *Log10*; (iii) *Filtering: baseline (specified points), baseline (weighted least square), derivative (Savitzky – Golay), smoothing (Savitzky – Golay), detrend, generalized least squares weighting (GLSW), orthogonal signal correction (OSC) and external parameter orthogonalization (EPO)*; (iv) Normalização: *normalize, standard normal variate (SNV) and multiplicative scatter correction (MSC-mean)*; (v) escalamento e centragem: *autoscale, group scale, Log decay scaling, mean center, median center, multiway center, multiway scale and sqrt mean scale*. Todas as análises foram realizadas no *software* SOLO (Eigenvector Research).

4.4.1.3 Machine learning

A primeira etapa para implementar qualquer modelo de ML é realizar uma análise exploratória. A análise exploratória pretende: (i) identificar a presença de possíveis outliers, (ii) reconhecer padrões de distribuição de dados no espaço multidimensional, e (iii) identificar relações entre variáveis²⁶⁹. Neste estudo, tanto os dados utilizados para construir o modelo de diagnóstico COVID-19 quanto os dados utilizados para construir o modelo de predição de gravidade foram previamente submetidos a um método de análise exploratória: análise de componentes principais

(PCA). Além disso, os outliers foram detectados e eliminados do conjunto de dados usando o método gráfico de alavancagem versus resíduos studentizados ²⁷⁰.

Na sequência, foram treinados e validados quatro algoritmos supervisionados de ML visando a predição do diagnóstico e da severidade de COVID-19: (i) redes neurais artificiais (ANN), (ii) árvores de decisão (DT), (iii) análise discriminante por mínimos quadrados parciais (PLS-DA) e (iv) o método de k -vizinhos mais próximos (KNN). Para implementação desses modelos com algoritmos de diagnóstico e predição da gravidade do COVID-19, foram utilizadas 70% das amostras para o conjunto de treinamento e 30% para o conjunto de teste, como pode ser observado na **Tabela 4.2**. Tanto para os modelos de ML de diagnóstico quanto para o modelo de gravidade, o método Kennard-Stone foi empregado para selecionar amostras do conjunto de treinamento e amostras do conjunto de teste. As amostras utilizadas para implementação dos algoritmos para o diagnóstico de COVID-19 foram divididas em classe 1 (amostras negativas para COVID-19) e classe 2 (amostras positivas para COVID-19). Para os modelos de predição de gravidade, as amostras foram classificadas em classe 1 (doença não grave; ou seja, pacientes ambulatoriais) e classe 2 (doença grave, ou seja, pacientes hospitalizados) (**Tabela 4.2**).

Tabela 4.2. Subconjunto de dados de treinamento e de teste utilizados para o desenvolvimento de modelos de *machine learning* para prever o diagnóstico e severidade de COVID-19

Dados de diagnóstico					
Grupo de pacientes		Treinamento (70%)		Teste (30%)	
Negativo - controle	Positivo	Negativo	Positivo	Negativo	Positivo
5086	557	3560	390	1526	167
Dados de severidade					
Grupo de pacientes		Treinamento (70%)		Teste (30%)	
Não severo	Severo	Não severo	Severo	Não severo	Severo
402	155	281	109	121	54

Fonte: O Autor (2024)

O número de variáveis latentes (VLs) selecionadas para os modelos ML foi realizado pelo método de validação cruzada *leave-one-out*. O número de VLs foram selecionados considerando os menores valores da raiz quadrada do erro médio de validação cruzada (RMSECV). A validação analítica dos modelos baseados nos algoritmos de *machine learning* foi realizada utilizando as seguintes métricas:

sensibilidade (**Equação 1**), especificidade (**Equação 2**) e acurácia (**Equação 3**). A acurácia dos modelos de ML foi também avaliada através do cálculo da área sob a curva característica de operação do receptor (AUC ROC).

Essas métricas de desempenho foram calculadas usando os parâmetros verdadeiro positivo (VP), verdadeiro negativo (VN), falso positivo (FP) e falso negativo (FN) ^{269,271-273}. Em ML, uma amostra é chamada de verdadeiro positivo quando pertence à classe um (1) e é corretamente classificada pelo algoritmo de ML como pertencente à classe um (1). Uma amostra é considerada falso positiva quando pertence à classe zero (0) e é classificada incorretamente pelo algoritmo ML como sendo da classe um (1). Uma amostra verdadeiramente negativa pertence à classe zero (0) e é corretamente classificada como classe zero (0). Finalmente, uma amostra é falsa negativa quando pertence à classe um (1) e é erroneamente classificada pelo algoritmo ML como classe zero (0). Sensibilidade e especificidade são definidas como a capacidade do modelo baseado em ML de classificar corretamente amostras negativas e positivas, respectivamente. Precisão é a capacidade de um modelo baseado em ML de classificar corretamente amostras negativas e positivas. Os valores de sensibilidade, especificidade e acurácia variam de zero (0) a um (1), e quanto mais próximo de 1, mais sensível, específico e acurado é o modelo, respectivamente.

$$\text{Equação 1. Sensibilidade} = \frac{VP}{VP+FN}$$

$$\text{Equação 2. Especificidade} = \frac{VN}{VN+FP}$$

$$\text{Equação 3. Acurácia} = \frac{VP+VN}{VP+VN+FP+FN}$$

Nota: VP, verdadeiro positivo; VN, verdadeiro negativo; FP, falso positivo; FN, falso negativo.

O gráfico de Importância Variável na Projeção (VIP) foi construído para o modelo que apresentou alta acurácia diagnóstica e prognóstica, para investigar potenciais biomarcadores associados ao diagnóstico de COVID-19.

4.4.2 Estudo II: Análise de dados dos exames bioquímicos e hematológicos de pacientes COVID-19, HIV, Tuberculose e co-infectados HIV/TB atendidos em um Hospital Regional de referência do Norte de Moçambique (Hospital Geral de Marrere, Província de Nampula) visando predição do diagnóstico e investigação de biomarcadores associados a essas doenças.

4.4.2.1 Fluxograma do estudo e conjunto de dados e pacientes

A Figura 4.2, mostra o passo a passo que foi utilizado para a realização do estudo.

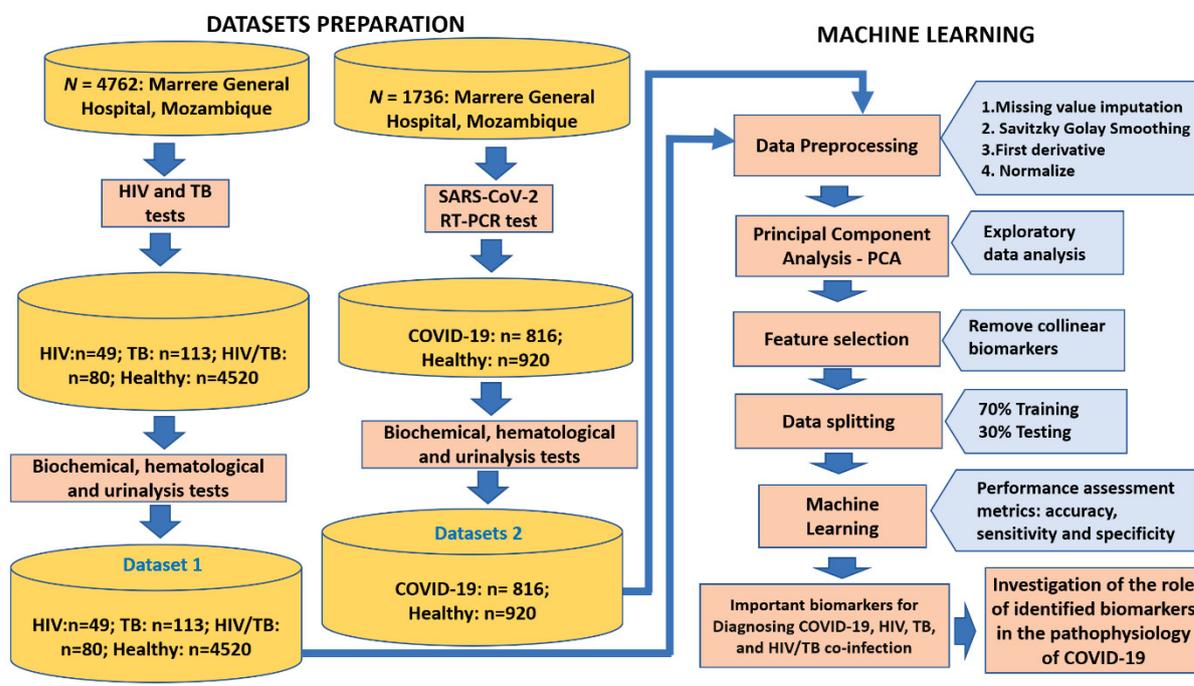


Figura 4.2. Fluxograma do estudo envolvendo dados dos exames bioquímicos e hematológicos de pacientes COVID-19, HIV, TB e co-infectados HIV/TB atendidos no Hospital Geral de Marrere (Moçambique) visando a predição do diagnóstico dessas doenças. **Fonte:** O Autor (2024)

4.4.2.2 Considerações éticas

Os conjuntos de dados de pacientes com COVID-19, HIV, tuberculose pulmonar e pacientes co-infectados HIV/TB foram obtidos do Hospital Geral do Marrere, Moçambique (Universidade Lúrio), um hospital universitário e de referência para COVID-19, HIV, e gestão da TB em Moçambique.

O estudo observou as diretrizes da Declaração de Helsinque da Associação Médica Mundial. Este estudo foi previamente aprovado pelo comitê de ética em pesquisa do Hospital Geral de Marrere (Universidade Lúrio, Moçambique) número: 36.1/Abril/CBISUL/21. O documento que fundamenta o parecer do comitê de bioética encontra-se na secção de anexos da presente tese de doutorado. Por se tratar de pesquisa com dados secundários, o comitê de ética da referida instituição dispensou a assinatura do Termo de Consentimento Livre e Esclarecido.

4.4.2.3 Conjunto de dados I: pacientes com COVID-19

Independentemente do teste para COVID-19 (RT-PCR negativo ou positivo), foram incluídos pacientes que apresentavam dados de exames hematológicos e bioquímicos (Figura 4.2). Assim, o conjunto de dados foi dividido de acordo com o resultado do RT-PCR em (i) pacientes positivos para COVID-19 (n=816 amostras); e pacientes negativos para COVID-19 (controle, n=920 amostras). Os dados foram coletados entre abril e novembro de 2021.

4.4.2.4 Conjunto de dados II: pacientes com HIV, TB e HIV/TB

Independentemente do resultado do teste de HIV/AIDS ou TB (negativo ou positivo), foram incluídos no estudo pacientes com resultados de exames hematológicos e bioquímicos (**Tabela 4.3**). Assim, o conjunto de dados foi organizado em quatro grupos: (i) pacientes com HIV (n=49); (ii) pacientes com TB pulmonar (n=113); (iii) pacientes co-infectados com HIV/TB (n=80); (iv) pacientes com teste negativo para HIV ou TB (controle, n=4.520). Os dados também foram coletados entre abril e novembro de 2021.

Tabela 4.3. Exames bioquímicos e hematológicos, de urina, virológicos e bacteriológicos usados para o desenvolvimento de modelos de *Machine learning* para a predição de diagnóstico de COVID-19, HIV/AIDS, TB e coinfeção HIV/TB.

Nome do exame	Unidade de mensuração
Glóbulos brancos	$10^9 L^{-1}$
glóbulos vermelhos	$10^{12} L^{-1}$
Hemoglobina	$g dL^{-1}$
Hematócrito	%
Volume corpuscular médio	fL
Hemoglobina corpuscular média	$g cell^{-1}$
Hemoglobina corpuscular média	$g Hb dL^{-1}$
Contagem de plaquetas	$10^9 L^{-1}$
Contagem de plaquetas	%
Contagem de plaquetas	$10^9 L^{-1}$
Contagem de neutrófilos	%
Contagem de neutrófilos	$10^9 L^{-1}$
Contagem de monócitos	%
Contagem de monócitos	$10^9 L^{-1}$
Creatinina Quinase	$U L^{-1}$
Uréia	$mg dL^{-1}$
Aspartato aminotransferase	$U L^{-1}$
Alanina aminotransferase	$U L^{-1}$
Glicose	$mg dL^{-1}$
Cálcio	$mmol L^{-1}$
Sódio	$mmol L^{-1}$
Potássio	$mmol L^{-1}$

Fonte: O Autor (2024)

4.4.2.5 Caracterização do grupo controle

O grupo controle utilizado neste estudo foi uma subdivisão do conjunto de dados bioquímicos, que incluiu pacientes atendidos no hospital que sofriam de doenças crônicas não transmissíveis, sendo as mais frequentes diabetes, obesidade e hipertensão. É fundamental destacar que todos os pacientes do grupo controle

foram negativos para infecções por HIV, *Mycobacterium tuberculosis* e SARS-CoV-2.

4.4.2.6 Combinação dos conjuntos de dados I e II: análise descritiva

Para o desenvolvimento dos modelos de *machine learning*, os conjuntos de dados I e II foram analisados separadamente. No entanto, uma análise descritiva geral dos dados (dados expressos como mediana e intervalo interquartil) foi realizada combinando os conjuntos de dados I e II em um único conjunto de dados final. Nesta combinação de conjuntos de dados, apenas os biomarcadores comuns em ambos os conjuntos de dados foram mantidos no banco de dados final (**Tabela 4.3**).

4.4.2.7 Conjunto de dados III: amostras de validação externa de modelos de *machine learning*

O conjunto de dados brasileiro foi utilizado para validação externa do modelo de *machine learning*. Ou seja, foram utilizados dados de Moçambique para o treino e validação do modelo para COVID-19, HIV, TB e co-infecção HIV/TB, ao passo que, o conjunto de dados do Brasil foi usado para testar se o modelo poderia prever amostras de pacientes de uma fonte externa (validação externa do modelo).

Os dados de pacientes no Brasil utilizados neste estudo foram obtidos do repositório público brasileiro de dados denominado “FAPESP COVID-19 *DataSharing/BR*”, que é um grande repositório de dados de pacientes com COVID-19 no estado de São Paulo (Brasil). Este repositório contém dados demográficos, dados de exames clínicos e/ou laboratoriais de pacientes e controles COVID-19 dos cinco principais hospitais do Estado de São Paulo (Brasil), a saber: (i) Hospital Israelita Albert Einstein; (ii) Hospital de Clínicas da Faculdade de Medicina da Universidade de São Paulo; (iii) Hospital Sírio-Libanês e (iv) Beneficência Portuguesa de São Paulo. A maioria desses pacientes com COVID-19 apresenta diversas comorbidades, como diabetes, hipertensão e hipotireoidismo.

Para acessar o repositório FAPESP COVID-19 *DataSharing/BR*, clique no seguinte link: <https://repositoriodatasharingfapesp.uspdigital.usp.br/>. Informações adicionais sobre o repositório FAPESP COVID-19 *DataSharing/BR* estão disponíveis na literatura ²⁷⁴. Considerando que os dados utilizados neste estudo são de acesso

aberto, para a sua utilização não foi necessário o parecer do comitê de bioética do Brasil.

4.4.2.8 *Análise univariada*

Antes de desenvolver os modelos de *machine learning* (análise multivariada), realizamos análise univariada para comparar os níveis de biomarcadores bioquímicos e hematológicos nos quatro grupos de pacientes do estudo (COVID-19, HIV, HIV/TB e grupo controle). Nesta análise, inicialmente foi testada a distribuição normal de cada biomarcador bioquímico e hematológico (variável) por meio do teste de Shapiro-Wilk. Em seguida, foram realizadas comparações dos biomarcadores dos quatro grupos de pacientes (COVID-19, HIV, HIV/TB e grupo controle) por meio do teste não paramétrico de Kruskal-Wallis, quando os biomarcadores não apresentavam distribuição normal (Teste de Shapiro Wilk, $p < 0,05$). Por outro lado, quando os biomarcadores apresentaram distribuição normal (teste de Shapiro Wilk, $p > 0,05$), as comparações entre os quatro grupos foram realizadas por meio do teste ANOVA one-way. É importante destacar também que os biomarcadores que não apresentaram distribuição normal foram relatados como mediana e intervalo interquartil (IIQ), e aqueles que apresentaram distribuição normal foram relatados como média \pm desvio padrão. Os níveis de significância $p < 0,05$ foram considerados estatisticamente significativos e todas as análises univariadas foram realizadas no software SPSS versão 2020 (IBM, EUA).

4.4.2.9 *Pré-processamento e análise exploratória dos dados*

O pré-processamento e a análise exploratória dos dados de pacientes COVID-19, HIV, TB e co-infectados HIV/TB seguiu os mesmos procedimentos descritos nos itens 3.2.2 e 3.2.3.

4.4.2.10 *Machine learning*

Após os dados terem sido pré-processados e análise exploratória ter sido conduzida, um total de sete modelos de ML [análise discriminante de mínimos quadrados parciais (PLS-DA), *Artificial Neural Network* (ANN), *eXtreme Gradient*

Boosting (XGBoosted), *K-Nearest Neighbors (KNN)*, regressão logística (LREG), *Soft independent modelling by class analogy (SIMCA)* e *Support Vector Machine (SVM)* foram treinados, validados e avaliados para prever o diagnóstico de COVID-19, HIV/AIDS, TB ou coinfeção HIV/TB. Para garantir uma representação significativa dos dados de treinamento e teste, cada uma das quatro classes de pacientes (COVID-19, HIV/AIDS, TB e HIV/TB) foi analisada individualmente, onde 70% dos dados foram usados para calibração e 30% foram utilizado para validação, conforme **Tabela 4.4**. É importante destacar que esta proporção de divisão dos dados de treinamento e teste foi realizada com base em estudos previamente publicados na literatura científica. As amostras foram selecionadas aleatoriamente usando o algoritmo de Kennard-Stone. Na **Tabela 4.4**, os dados desequilibrados observados entre os pacientes (COVID-19, HIV, TB e HIV/TB) e o grupo controle podem ser explicados devido as taxas de prevalência e incidência das doenças. Por exemplo, a prevalência do VIH em Moçambique é de 13,5% enquanto a incidência da tuberculose é de 551 casos por 100 mil pessoas ^{275,276}. Todos os modelos de ML foram otimizados considerando valores mais baixos de *Root-Mean-Square Error of Cross-Validation (RMSECV)* ²⁷⁷.

Tabela 4.4. Subconjunto de dados de calibração e validação usados para o desenvolvimento de modelos de *machine learning* para prever o diagnóstico de pacientes com COVID-19, HIV/AIDS, TB e HIV/TB

Conjunto de dados	Amostras		Treino (70%)		Teste (30%)	
	Positivo	Negativo (controle)	Positivo	Negativo	Positivo	Negativo
COVID-19	816	920	571	644	245	276
HIV/AIDS	49	4520	34	3164	15	1356
TB	113	4520	79	3164	34	1356
HIV/TB	80	4520	56	3164	24	1356

Fonte: O Autor (2024)

A validação do modelo (desempenho) foi avaliada considerando a especificidade, a sensibilidade e a acurácia (estabelecida pela ROC) ²⁷⁸. As equações 1, 2, 3, 4, 5 e 6 apresentadas abaixo foram usadas para calcular a sensibilidade (recall), especificidade, acurácia, precisão, F1 score e coeficiente de correlação de Mattew (CCM), respectivamente.

O gráfico de Importância Variável na Projeção (VIP) foi construído para o modelo que apresentou alta acurácia diagnóstica, para investigar potenciais

biomarcadores associados ao diagnóstico de COVID-19, HIV/AIDS, TB e coinfeção HIV/TB.

$$\text{Equação 4. Sensibilidade (recall)} = \frac{VP}{VP+FN}$$

$$\text{Equação 5. Especificidade} = \frac{VN}{VN+FP}$$

$$\text{Equação 6. Acurácia} = \frac{VP+VN}{VP+VN+FP+FN}$$

$$\text{Equação 7. Precisão} = \frac{VP}{VP+FP}$$

$$\text{Equação 8. F1 score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

$$\text{Equação 9. CCM} = \frac{VP \times VN - FP \times FN}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}}$$

Nota: VP: verdadeiro positivo, VN: verdadeiro negativo; FP: falso positivo e FN: falso negativo. CCM: coeficiente de correlação de Matthew.

4.4.3 Estudo III: Predição de diagnóstico de COVID-19 usando dados clínicos de pacientes COVID-19 atendidos na rede de Farmácia Drugstore distribuída em todo território Nacional.

4.4.3.1 Fluxograma do estudo e conjunto de dados dos pacientes

O fluxograma do estudo II é mostrado na Figura 4.3. Em suma, foi realizado um estudo transversal envolvendo cerca de 4,0 milhões de pacientes brasileiros com sintomas de COVID-19 que, curiosa e voluntariamente, se dirigiram às diversas farmácias privadas da rede *Drugstore*, para realizar o teste rápido de COVID-19 para confirmar as suas suspeitas, se os sintomas que apresentavam eram COVID-19 ou outras patologias. É importante destacar que *Drugstore* é uma rede de farmácias com abrangência nacional e, portanto, o presente estudo envolveu pacientes localizados em todos os estados brasileiros que foram atendidos entre janeiro-junho de 2023. Todos os pacientes que realizaram o teste rápido para COVID-19 tiveram seus dados coletados, independentemente do resultado do seu teste COVID-19 ser positivo ou

negativo. Como se trata de dados de públicos, não foi necessário o parecer do comitê de ética em pesquisa, mas todos os dados dos pacientes no estudo foram fornecidos anonimamente e o estudo está em conformidade com todos os padrões éticos da Declaração de Helsinque. (Associação Brasileira de Redes de *Drugstore*; <http://abrafarma.com.br>) (abrangência nacional).

4.4.3.2 Critérios de inclusão e exclusão

A coleta de dados para este estudo foi realizada naqueles pacientes que se deslocaram voluntariamente às farmácias para realizar o teste de COVID-19 por curiosidade, devido aos sintomas que apresentavam. Tanto os pacientes que foram negativos para COVID-19 quanto aqueles que testaram positivo para COVID-19 foram incluídos no estudo.

A decisão sobre qual tipo de teste rápido para COVID-19 utilizar (antígeno ou anticorpos) dependeu de avaliação clínica prévia do farmacêutico onde o paciente foi atendido. Para estabelecer os critérios de testagem dos pacientes foi utilizada a metodologia do *Center for Disease Control* dos EUA, a mesma adotada pelo Ministério da Saúde do Brasil (<https://www.cdc.gov/coronavirus/2019-ncov/lab/resources/antigen-tests-guidelines.html>). Como o teste de antígeno da COVID-19 é mais preciso na fase aguda da doença, foram incluídos para testagem (métodos antígeno ou anticorpo; IgG-IgM): (i) pacientes sintomáticos entre o 1º e o 7º dia após o início dos sintomas do COVID-19; (ii) pacientes assintomáticos a partir do 5º dia de contato com casos confirmados. Além disso, (iii) pacientes que apresentaram sintomas leves de coronavírus por mais de 10 dias foram incluídos para testes de antígeno; (iv) pessoas com sintomas de coronavírus há algum tempo e que desejam confirmar a causa da infecção; (v) indivíduos sem sintomas, mas que desejam saber se tiveram contato com o vírus; (vi) Foram excluídas do estudo crianças menores de 14 anos sem responsável legal para supervisionar a realização do teste. Não foi necessária solicitação médica para a realização do exame.

4.4.3.3 Testes e prestação de cuidados farmacêuticos

Cinco tipos diferentes de fabricantes de testes para COVID-19 autorizados e aprovados pelo Ministério da Saúde do Brasil estavam disponíveis durante o estudo:

Abbott (sensibilidade 98,1% [93,2%; 99,8%], especificidade 99,8% [98,6%; 100,0%], porcentagem geral concordância 99,4% [98,3%; 99,9%]); Diagnóstico ECO (sensibilidade 96,52%, especificidade: >99,9%); Lepu/Leccurate (sensibilidade 90%, especificidade IgM 100%); MedLevensohn/FlowFlex (sensibilidade 97,1%, especificidade 99,6%); Newscen (sensibilidade 96,79%, especificidade 99,28%, precisão 98,43%).

Os farmacêuticos aplicaram os protocolos de triagem de sintomas e testes rápidos utilizando a plataforma Clinicarx, garantindo o mesmo padrão de prática para todos os pacientes. Os pacientes também foram convidados a participar de anamnese, oximetria e medidas de pressão arterial. Todas as farmácias foram estruturadas com consultórios clínicos para prestar esse serviço e garantir a privacidade entre profissionais de saúde e pacientes. Protocolos rígidos de segurança e qualidade foram seguidos pelos profissionais de saúde durante todo o procedimento, conforme recomendações da Organização Mundial da Saúde (OMS). Dependendo dos resultados dos testes e rastreios da COVID-19, os pacientes foram aconselhados pelo farmacêutico a continuar a seguir as medidas preventivas (por exemplo, quarentena, isolamento) ou foram encaminhados para o médico ou hospital (por exemplo, casos mais graves).

4.4.3.4 Análise de dados

Todos os dados do estudo foram analisados em linguagem *Python*. A Figura 4.3 apresenta o fluxograma de todo o estudo. Todos os códigos *Python* utilizados em cada etapa do estudo estão disponíveis no seguinte link do GitHub: [GitHub - AlexandreCOBRE/code: Compilação dos códigos de programação em linguagem Python](#). Portanto, o passo a passo para a realização do estudo foi o seguinte:

1. Fusão dos cinco conjuntos de dados dos diferentes meses de coleta (janeiro-junho) formando um único conjunto de dados (denominado 'Dataset 1') contendo 3.325.567 pacientes e 35 variáveis;
2. Pré-processamento e engenharia de recursos, onde colunas (variáveis) e linhas (pacientes) completamente vazias foram eliminadas (ou seja, os dados brutos foram transformados em dados mais limpos, normalizando os dados e obtendo modelos de ML facilmente interpretáveis)^{279,280}. Para isso, foram aplicados dois métodos: limpeza de dados (pacientes vazios e variáveis eliminadas no Dataset

- 1) e imputação de mediana de valores faltantes da idade dos pacientes ^{281,282}, resultando em um novo conjunto de dados ('Dataset 2') incluindo 3325567 pacientes e 20 variáveis.
3. Divisão do Dataset 2 de acordo com o tipo de teste diagnóstico em três novos conjuntos: teste de antígeno (Dataset 3 formado por 287.754 pacientes e 20 variáveis), IgG (Dataset 4 formado por 173.556 pacientes e 20 variáveis) e IgM (Dataset 5 formado por 2.44.023 pacientes e 20 variáveis).
 4. Análises exploratórias utilizando o método de análise de componentes principais (PCA) visando: (i) reconhecer padrões dada a distribuição das amostras no espaço multidimensional (ou seja, discriminação entre pacientes positivos e negativos para COVID-19), (ii) detectar a presença de possíveis amostras anômalas (outliers); (iii) avaliar a relação entre variáveis ²⁸³.
 5. Desenvolvimento de modelos de IA e ML para cada conjunto de dados, separadamente, para a previsão do diagnóstico de COVID-19. Um total de 15 algoritmos baseados foram testados para cada conjunto de dados: *Decision Tree Classifier* (DT), *Extreme Gradient Boosting* (XGBOOST), *Extra Trees Classifier* (ET), *Random Forest Classifier* (RF), *Quadratic Discriminant Analysis* (QDA), *Ada Boost Classifier* (ADA), *Classificador de Gradient Boosting* (GBC), Classificador Dummy (DUMMY), *Naive Bayes* (NB), *Light Gradient Boosting Machine* (LIGHTGBM), Regressão Logística (LR), Classificador de Ridge (RIDGE), Análise Discriminante Linear (LDA), Classificador de Vizinhos K (KNN), *Kernel Linear* (SVM). Para o treinamento e teste dos modelos foram utilizados 70% e 30% dos dados, respectivamente.

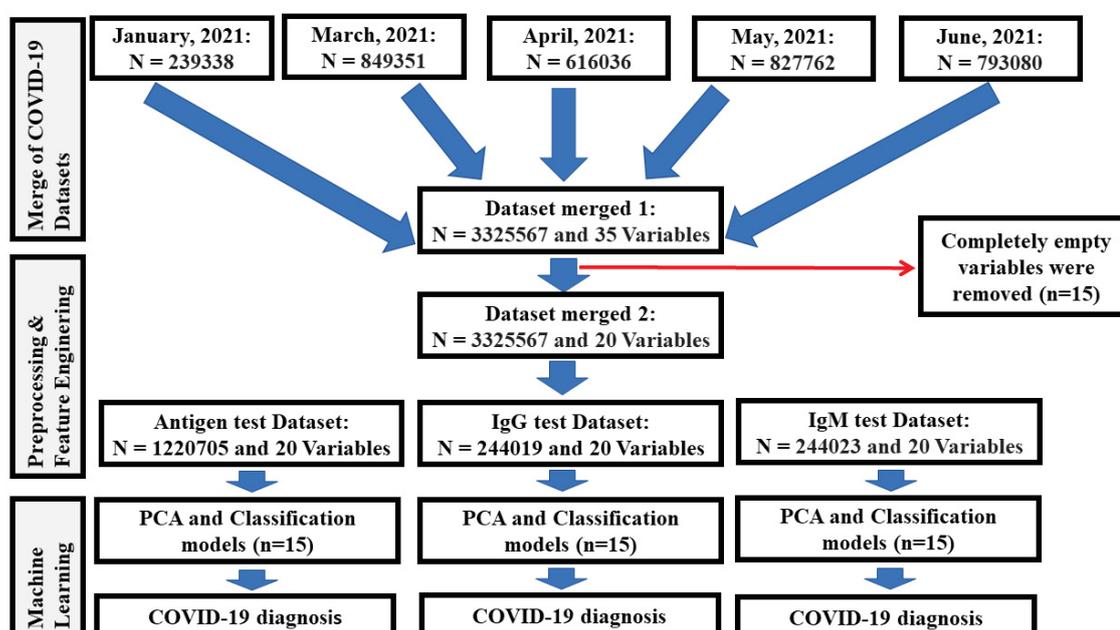


Figura 4.3. Fluxograma do estudo envolvendo análise de dados clínicos e demográficos dos pacientes atendidos nas diferentes farmácias privadas no Brasil visando o desenvolvimento de modelos de *machine learning* para predição do diagnóstico da COVID-19. **Fonte:** O Autor (2024)

O desempenho dos modelos de ML na previsão do diagnóstico de COVID-19 em cada conjunto de dados foi avaliado usando precisão (ver equação 1 mostrada abaixo) e métricas de tempo de treinamento. A partir dos 5 principais modelos de ML (ou seja, apresentando as melhores métricas) foram calculadas variáveis de predição associadas ao diagnóstico de COVID-19. Variáveis relevantes (ou seja, relatadas em pelo menos três modelos de ML) foram selecionadas para discussão. A precisão varia entre zero e um e quanto mais próximo de 1, melhor desempenho do modelo ML na previsão do diagnóstico de COVID-19. Por outro lado, quanto menor o tempo necessário para treinar o modelo, melhor será o desempenho do modelo.

A validação cruzada é uma ferramenta crucial em ML para garantir que seus modelos tenham bom desempenho e sejam capazes de generalizar para dados novos e invisíveis. É uma prática padrão no desenvolvimento e avaliação de modelos ²⁸⁴. Neste estudo, para melhorar o desempenho dos modelos de *machine learning* no diagnóstico de COVID-19, foi utilizada a validação cruzada K-Fold (k = 10).

4.5 RESULTADOS

4.5.1 Estudo I: Análise de dados dos exames bioquímicos, hematológicos e de urinálise de pacientes COVID-19 atendidos no Hospital Israelita Alberto Einstein (Brasil) visando a predição do diagnóstico e investigação de potenciais Biomarcadores prognósticos.

No estudo I, a variação de todos os biomarcadores (bioquímicos, hematológicos e urinários) para os pacientes com COVID-19 está resumida na Tabela 4.5.

Tabela 4.5. Níveis de variação dos biomarcadores bioquímicos, hematológicos e urinários de pacientes positivos e com a doença grave em uma escala normalizada de pacientes.

Biomarcadores bioquímicos, hematológicos e urinários	Amostras positivas de pacientes para COVID-19*	Amostras de pacientes graves com COVID-19**
Hematócrito	Baixo	Baixo
Hemoglobina	Baixo	Baixo
Plaquetas	Baixo	Baixo
Volume médio de plaquetas	Baixo	Baixo
Glóbulos vermelhos	Baixo	Baixo
Linfócitos	Baixo	Baixo
Concentração média de hemoglobina corpuscular (MCHC)	Baixo	Baixo
Leucócitos	Alto	Alto
Basófilos	Normal	Normal
Hemoglobina corpuscular média (MCH)	Normal	Baixo
Eosinófilos	Baixo	Baixo
Volume corpuscular médio (VCM)	Baixo	Baixo
Monócitos	Alto	Normal
Largura de distribuição de glóbulos vermelhos (RDW)	Baixo	Normal
Glicose sérica	Alto	Alto
Neutrófilos	Baixo	Baixo
Uréia	Baixo	Baixo
proteína C-reativa	Alto	Alto
Creatinina	Alto	Alto
Potássio	Baixo	Baixo
Sódio	Baixo	Baixo
Alanina transaminase	Alto	Alto
Aspartato transaminase	Alto	Alto
Gama-glutamyltransferase	Alto	Alto
Bilirrubina total	Alto	Alto
Bilirrubina direta	Alto	Alto
Bilirrubina indireta	Alto	Alto
Fosfatase alcalina	Alto	Alto
Cálcio ionizado	Baixo	Baixo
pCO ₂ (gasometria venosa)	Alto	Alto
Magnésio	Baixo	Baixo
Saturação de Hb (gasometria venosa)	Baixo	Baixo
Excesso de base (gasometria venosa)	Baixo	Baixo
pO ₂ (gasometria venosa)	Baixo	Baixo

CO ₂ total (análise de gasometria venosa)	Alto	Alto
pH (análise de gases no sangue venoso)	Baixo	Baixo
HCO ₃ (gasometria venosa)	Alto	Alto
Varas	Alto	Alto
Segmentado	Baixo	Baixo
Promielócitos	Normal	-----
Metamielócitos	Normal	-----
Mielócitos	Normal	-----
Urina - pH	Baixo	Baixo
Urina - Densidade	Normal	Baixo
Urina – Glóbulos vermelhos	Normal	Normal
Razão normalizada internacional (INR)	Alto	Alto
Desidrogenase láctica	Alto	Alto
Creatinofosfoquinase (CPK)	Normal	Baixo
Ferritina	Alto	Alto
Ácido láctico arterial	Alto	Alto
Saturação de Hb (gasometria arterial)	Baixo	Baixo
pCO ₂ (gasometria arterial)	Alto	Alto
Excesso de base (gasometria arterial)	Baixo	Baixo
pH (análise de gasometria arterial)	Baixo	Baixo
CO ₂ total (gasometria arterial)	Alto	Alto
HCO ₃ (gasometria arterial)	Alto	Alto
pO ₂ (gasometria arterial)	Baixo	Baixo
FiO ₂ arterial	Baixo	Baixo
Fósforo	Baixo	-----

Nota:* Dados de diagnóstico; ** Dados de gravidade da doença. **Fonte:** O Autor (2024)

4.5.1.1 Machine learning: análise não supervisionada

Os resultados das análises exploratórias dos dados de diagnóstico e gravidade da doença usando o modelo PCA estão representados na Figura 4.4. O modelo PCA de diferenciar entre pacientes positivos e negativos para COVID-19 (Figura 4.4-A, dados de diagnóstico) e entre pacientes com doença não grave e pacientes com doença grave (Figura 4.4-B, dados de gravidade da doença).

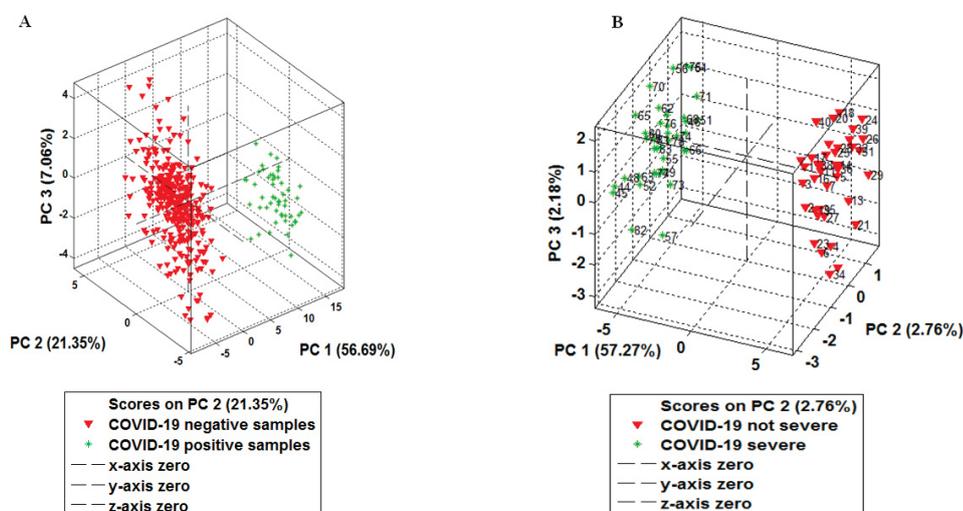


Figura 4.4. Análise exploratória. Modelo de análise de componentes principais (PCA) de discriminação de amostras negativas e positivas (A) e amostras de pacientes com doença grave e não grave (B). **Fonte:** O Autor (2024)

Na sequência, foram analisadas amostras discrepantes para dados de diagnóstico e gravidade da doença usando o gráfico de alavancagem versus resíduos estudentizados com intervalo de confiança de 95% (Figura 4.5). Neste gráfico, uma amostra é considerada outlier se e somente se tiver simultaneamente altos valores de *leverage* e de resíduos de *student*. Assim, embora algumas amostras tenham mostrado altos valores de *leverage* (eixo X), elas não podem ser consideradas outlier porque estão dentro de resíduos de ± 2.5 desvios-padrão de resíduos de *student* (eixo y), como pode ser observado na Figura 4.5.

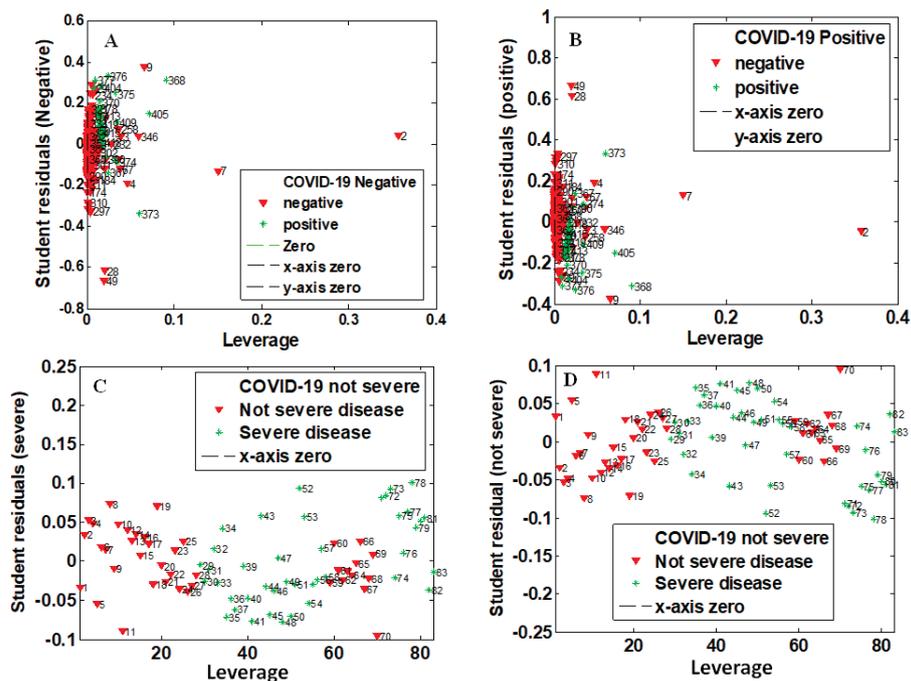


Figura 4.5. Gráfico de *leverage* versus resíduos de *student* para detecção de amostras discrepantes. Para dados de diagnóstico: análise de outliers de amostras negativas (A) e amostras positivas (B). Para dados de gravidade: análise de outliers para amostras de pacientes sem gravidade (C) e com gravidade (D).

4.5.1.2 Machine learning: análise supervisionada

Os resultados dos quatro modelos baseados em ML para os subconjuntos de dados são apresentados na **Tabela 4.6**. Todos os modelos para diagnóstico de COVID-19 e previsão da gravidade da doença foram comparados entre si usando as seguintes métricas: tempo de treinamento, erro de treinamento do modelo, erro de validação cruzada, sensibilidade, especificidade e acurácia (área sob a curva ROC). O modelo de redes neurais artificiais teve melhor desempenho preditivo por apresentar menos tempo de treinamento e menores erros de previsão, além de apresentar maior acurácia e maior valor de AUC ROC (Figura 4.6). As curvas ROC dos modelos estão representadas na Figura 4.6.

Tabela 4.6. Comparação de desempenho dos modelos de *machine learning* para COVID-19

Métrica	Modelo de diagnóstico				Modelo de severidade			
	ANN	DT	PLS-DA	K-NN	ANN	DT	PLS-DA	K-NN
Tempo de treino	21min. 43 s	27 min. 11 s	31 min. 19 s	22 min. 15 s	7min. 1 s	10 min. 19 s	18 min. 3 s	09 min. 53 s
Erro de calibração	1.0%	0.5	1.2%	0.5%	1.0%	8.4%	6.0%	0.4%
Erro de validação cruzada	0.8 %	1.0	0.9%	0.6%	0.5%	1.8%	4.0%	0.7%
Sensibilidade	0.93	0.89	0.88	0.84	0.99	0.90	0.87	0.82
Especificidade	0.94	0.89	0.90	0.83	0.97	0.94	0.88	0.88
Acurácia*	0.94	0.90	0.90	0.84	0.98	0.92	0.88	0.86

Nota: ANN: *Artificial neural network*; DT: *Decision tree*; PLS-DA: *Partial Least Squares Discriminant Analysis*; KNN: *K-Nearest Neighbors*; *Área under the ROC curve.

Fonte: O Autor (2024)

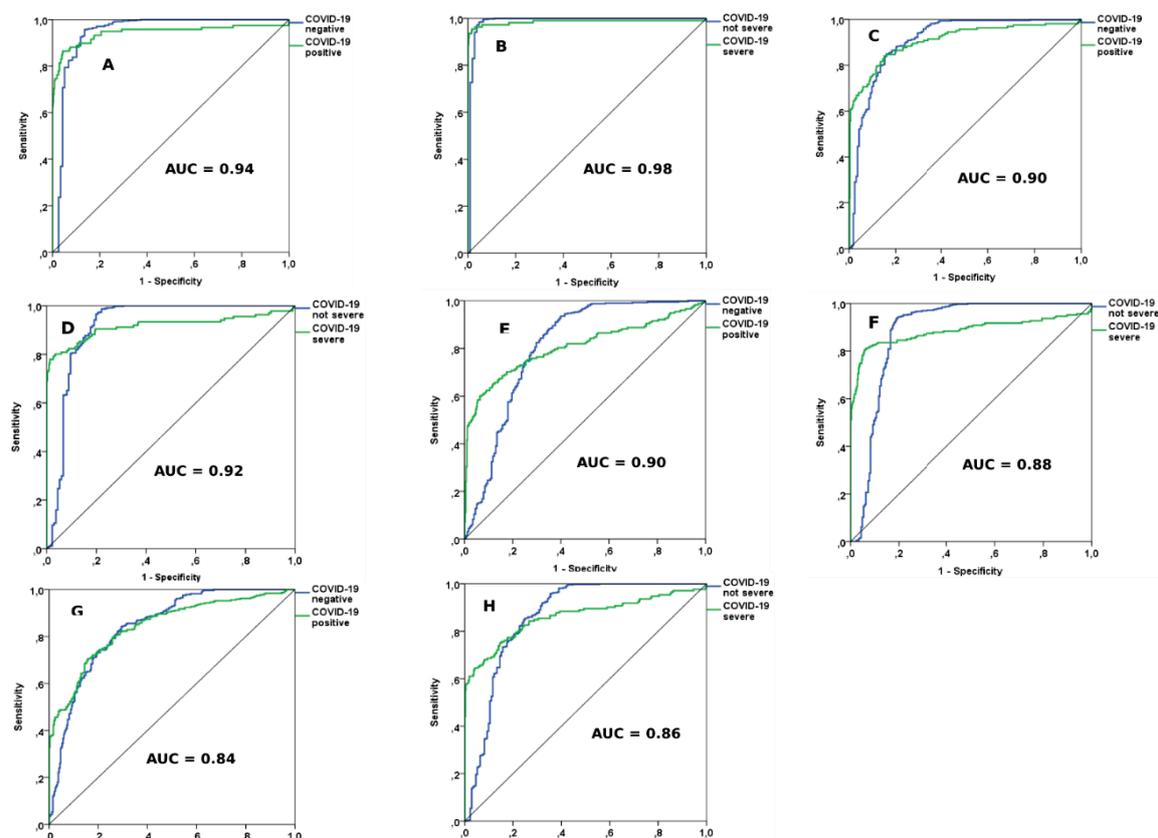


Figura 4.6. Curvas ROC de acurácia dos modelos de *machine learning*. *Artificial Neural Network* (ANN): diagnóstico (A) e gravidade (B). *Decision tree* (DT): diagnóstico (C) e gravidade (D). *Partial Least Squares Discriminant Analysis* (PLS-DA): diagnóstico (E) e gravidade (F). *K-Nearest Neighbors* (KNN): diagnóstico (G) e gravidade (H).

4.5.1.3 Identificação de biomarcadores importantes no diagnóstico e prognóstico da COVID-19

De acordo com os modelos baseados em ML, alguns biomarcadores foram considerados críticos ou importantes para prever a COVID-19 e a gravidade da doença (ver Tabela 4.7). A ferritina foi classificada como a variável mais importante de todos os modelos.

Tabela 4.7. Variáveis importantes dos modelos de *machine learning* na classificação de positividade e severidade de COVID-19

Variável	Modelo de diagnóstico				Modelo de severidade			
	ANN	DT	PLSD A	KNN	ANN	DT	PLSD A	KNN
Ferritina	++	++	++	++	++	++	++	++
Gama-glutamyltransferase	+	-	-	-	-	-	-	-
HCO ₃ (arterial)	+	-	-	+	-	-	+	+
Excesso de base (arterial)	+	-	-	-	-	-	+	+
Excesso de base (venoso)	-	-	-	-	-	-	+	-
Sódio	+	-	-	-	-	-	-	-
O ₂ total (arterial)	-	-	-	-	+	-	-	-
pO ₂ (arterial)	-	-	-	+	-	-	-	-
CO ₂ total	-	+	+	+	-	-	-	+
pCO ₂ (arterial)	+	-	-	-	-	-	+	+
pCO ₂ (venoso)	+	-	+	-	-	-	+	-
Bilirrubina indireta	+	-	-	-	-	-	-	-
Fosfatase alcalina	+	-	-	-	+	-	-	-
PH da urina	-	-	+	+	+	-	+	-
pH (venoso)	-	-	-	-	-	-	+	-
pH (arterial)	-	-	-	-	-	-	-	+
FiO ₂ (arterial)	-	-	-	-	+	-	-	-
ctO ₂ (arterial)	-	-	-	-	-	-	-	+
Bilirrubina total	-	-	-	-	+	-	-	-
Largura de distribuição de glóbulos vermelhos	-	-	-	-	+	-	-	-
Plaquetas	-	-	-	-	+	-	-	-

proteína C-reativa	-	-	-	-	+	-	+	-
Cálcio ionizado	-	-	+	+	-	-	-	-
Densidade da urina	-	-	+	+	-	-	-	-
Desidrogenase láctica	-	-	-	-	-	-	+	-
Ácido láctico arterial	-	-	-	+	-	-	-	+
Saturação de hemoglobina (arterial)	-	-	-	-	-	-	-	+
Fósforo	-	-	+	+	-	-	-	-
Dosagem de lipase	-	-	-	-	-	-	+	-
<i>Roods</i>	-	-	-	+	-	-	+	-

Nota: Variável não importante (-); Variável importante (+); variável mais importante (++) . **Fonte:** O Autor (2024)

4.4.2 Estudo II: Análise de dados dos exames bioquímicos e hematológicos de pacientes COVID-19, HIV, Tuberculose e co-infectados HIV/TB atendidos em um Hospital Regional de referência do Norte de Moçambique (Hospital Geral de Marrere, Província de Nampula) visando predição do diagnóstico e investigação de biomarcadores associados a essas doenças.

4.5.2.1 Análise univariada

Os resultados da análise univariada comparando os níveis de todos os biomarcadores bioquímicos e hematológicos entre pacientes com COVID-19, HIV/AIDS, TB pulmonar e coinfeção HIV/TB são apresentados na **Tabela 4.8**. O teste de Kruskal Wallis mostrou que houve diferenças significativas ($p < 0,05$) na viabilidade de todos esses biomarcadores nos quatro grupos estudados.

Tabela 4.8. Biomarcadores bioquímicos e hematológicos utilizados no estudo.

Variável (unidade)	COVID-19			HIV/AIDS			Tuberculose			HIV/TB			Controle		
	Mediana	IIQ*	Mediana	IIQ	Mediana	IIQ	Mediana	IIQ	Mediana	IIQ	Mediana	IIQ	Mediana	IIQ	P**
Anos de idade)	62,0	51,00 - 75,00	38,00	32,00 - 43,00	32,00	24,00 - 40,00	33,50	28,25 - 42,75	64,00	45,00 - 78,00	0,00				0,00
Globulos brancos (109 L ⁻¹)	6,58	5,01 - 9,58	4,50	3,70 - 5,60	7,30	5,15 - 9,75	7,15	4,83 - 9,98	8,80	6,75 - 11,31	0,00				0,00
Globulos vermelhos (1012 L ⁻¹)	4,67	4,27 - 5,07	3,93	3,46 - 4,75	4,48	3,99 - 4,94	3,73	3,27 - 4,33	4,50	4,03 - 4,89	0,00				0,00
Hemoglobina (g dL ⁻¹)	13,70	12,31 - 14,74	12,40	11,25 - 13,85	11,20	9,85 - 13,15	10,00	8,53 - 11,43	13,10	11,53 - 14,30	0,00				0,00
Hematócrito (%)	40,50	37,00 - 43,55	38,10	34,65 - 40,90	35,00	31,35 - 39,60	30,30	26,55 - 34,40	39,03	35,30 - 42,20	0,00				0,00
Volume corpuscular médio (fL)	87,10	84,00 - 90,10	98,90	76,95 - 107,85	73,00	-70,30 - 83,95	71,05	-73,88 - 87,78	87,70	84,26 - 91,24	0,00				0,00
Hemoglobina corpuscular média (g cell ⁻¹)	29,40	28,21 - 30,50	31,20	27,90 - 34,45	25,20	-19,30 - 27,75	26,50	-17,75 - 29,05	29,50	28,10 - 30,70	0,00				0,00
Hemoglobina corpuscular média (g Hb dL ⁻¹)	33,70	32,90 - 34,35	31,90	29,70 - 34,55	31,90	30,00 - 33,25	32,05	28,98 - 33,80	33,50	32,56 - 34,20	0,00				0,00
Contagem de plaquetas (109 L ⁻¹)	209,0	162,00 - 268,75	212,00	159,50 - 268,50	319,50	226,50 - 396,00	305,50	226,00 - 389,75	234,67	187,00 - 284,00	0,00				0,00
Contagem de plaquetas (%)	15,6	10,16 - 22,69	41,10	27,85 - 51,70	27,20	21,00 - 37,90	24,20	14,80 - 37,75	18,40	14,18 - 22,60	0,00				0,00
Contagem de plaquetas (109 L ⁻¹)	1,0	0,71 - 1,30	2,20	1,75 - 2,95	2,00	1,50 - 2,65	1,60	1,10 - 2,38	1,50	1,20 - 1,75	0,00				0,00
Contagem de neutrófilos (%)	75,6	66,83 - 83,50	31,80	20,65 - 47,35	58,50	52,45 - 66,60	57,35	47,15 - 76,75	70,40	66,40 - 76,10	0,00				0,00
Contagem de neutrófilos (109 L ⁻¹)	4,80	3,50 - 7,25	1,80	1,34 - 2,35	4,40	3,06 - 6,70	4,30	3,03 - 6,28	5,95	5,00 - 7,10	0,00				0,00
Contagem de monócitos (%)	7,30	5,20 - 9,50	10,70	8,00 - 15,15	10,30	8,30 - 12,80	9,15	7,08 - 11,88	7,78	6,80 - 8,70	0,00				0,00
Contagem de monócitos (109 L ⁻¹)	0,50	0,30 - 0,60	0,50	0,35 - 0,68	0,80	0,60 - 1,10	0,80	0,50 - 0,90	0,60	0,56 - 0,75	0,00				0,00
Creatinina Quinase (U L ⁻¹)	100,0	100,00 - 100,00	83,85	76,15 - 91,35	61,83	61,83 - 61,83	63,41	63,41 - 63,41	78,00	78,00 - 78,00	0,00				0,00
Uréia (mg dL ⁻¹)	34,83	30,00 - 42,00	3,55	3,10 - 3,80	3,40	3,40 - 3,40	4,74	4,74 - 4,74	35,00	33,00 - 38,30	0,00				0,00
Aspartato aminotransferase (U L ⁻¹)	43,00	30,00 - 63,00	31,00	26,50 - 34,50	29,30	29,30 - 29,30	23,40	23,40 - 23,40	26,00	21,00 - 33,00	0,00				0,00
Alanina aminotransferase (U L ⁻¹)	33,85	23,00 - 54,46	31,00	27,00 - 38,35	21,00	21,00 - 21,00	25,20	25,20 - 25,20	22,00	16,31 - 31,00	0,00				0,00
Glicose (mg dL ⁻¹)	107,0	95,00 - 129,00	4,41	4,41 - 4,41	4,70	4,70 - 4,70	4,82	4,82 - 4,82	100,75	89,00 - 118,00	0,00				0,00

Nota: IIQ: Intervalo interquartil; HIV: Vírus de imunodeficiência humana; AIDS: síndrome de imunodeficiência adquirida. **Fonte:** O Autor (2024)

4.5.2.2 Análise multivariada: Análise de Componentes Principais-PCA

Os resultados da análise exploratória (modelo PCA) dos dados de diagnóstico de COVID-19 são mostrados na Figura 4.7, enquanto a análise exploratória para dados de pacientes com HIV/AIDS, TB pulmonar e coinfeção HIV/TB é mostrada na Figura 4.9. No PCA análise, uma amostra é considerada um potencial outlier se, e somente se, tiver simultaneamente altos valores de Q-resíduos e altos valores de Hotelling T^2 . Para ambos os conjuntos de dados, embora algumas amostras tenham valores Q-resíduos elevados, elas não podem ser consideradas outliers porque estão dentro do intervalo de confiança de 95% do Hotelling T^2 (figuras 4.7-B e 4.9-B). É importante destacar que ambos os modelos PCA foram otimizados utilizando a combinação dos seguintes métodos de processo: normalização + autoescalamento. Para discriminar pacientes do conjunto de dados 1 (COVID-19 vs controle) e do conjunto de dados 2 (HIV vs TB vs HIV/TB vs controle) pelo modelo PCA, dois componentes principais (90,42% de variância explicada) e três componentes principais (96,89% de variância explicado), respectivamente. A partir desta análise, não foram detectadas amostras outliers nos dados de diagnóstico de COVID-19 (Figura 4.9-B) e nos dados de diagnóstico de HIV/AIDS, TB pulmonar e coinfeção HIV/TB (Figura 4.9-B).

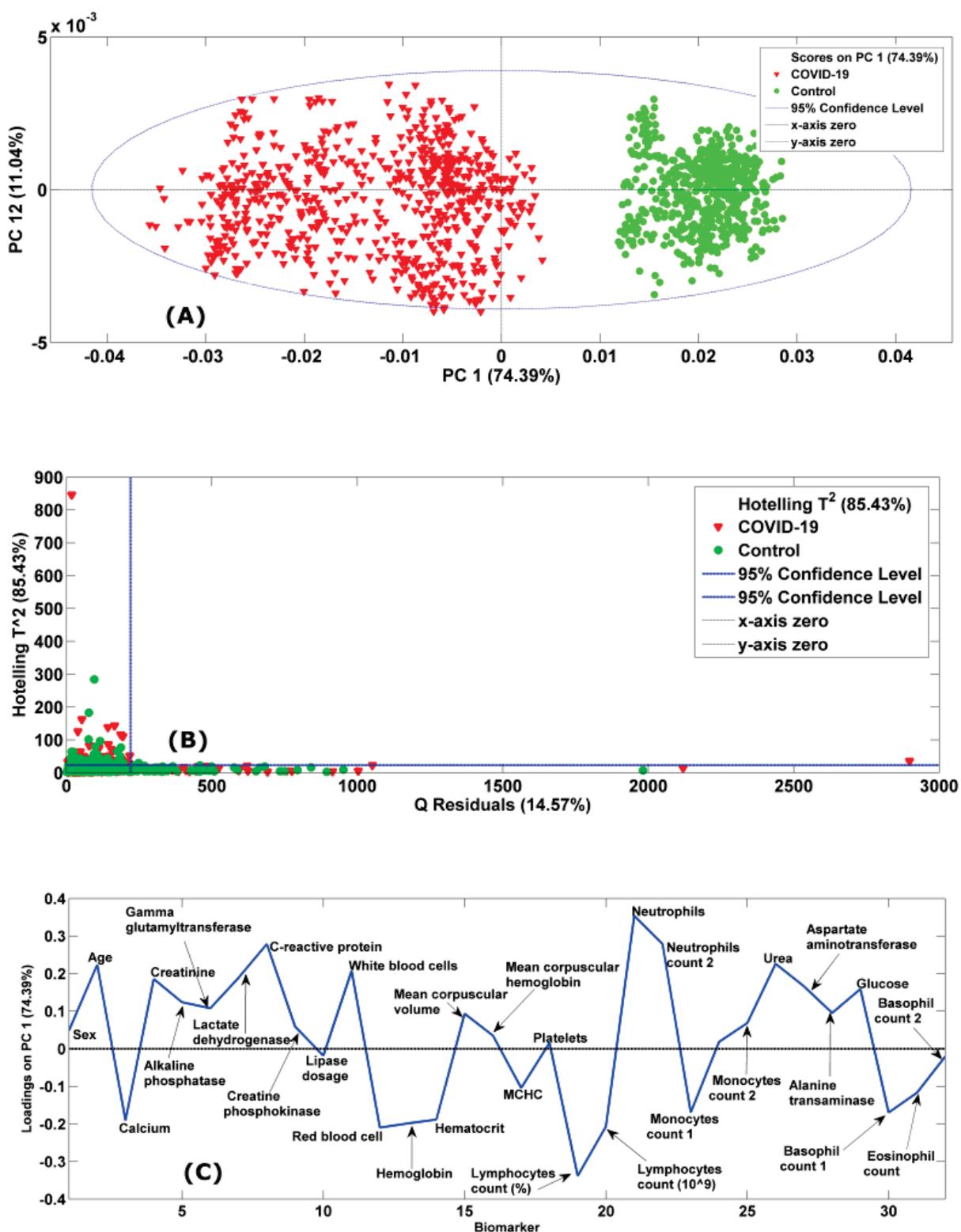


Figura 4.7. Análise exploratória do conjunto de dados COVID-19. Em (A) é mostrado o modelo PCA das amostras de sangue de 816 pacientes com COVID-19 diagnosticados por RT-PCR são representadas pelos triângulos vermelhos e as amostras de sangue de 920 controles com RT-PCR negativo são representadas por círculos verdes. Em (B) é mostrado o gráfico de Hotelling T^2 versus Q-resíduos do modelo PCA para detectar valores discrepantes em dados de amostras de pacientes

com COVID-19. Neste gráfico, uma amostra é considerada outlier se e somente se apresentar simultaneamente altos valores de Hotelling T^2 e altos valores de Q-residual. De acordo com o gráfico, apesar de algumas amostras apresentarem valores elevados de Q-Resíduos, elas não podem ser consideradas outliers porque estão dentro do intervalo de confiança de 95% do Hotelling T^2 . Em (C) é mostrado o gráfico de cargas que representa as variáveis mais importantes na discriminação de amostras do grupo COVID-19 e controles. Apenas o gráfico de cargas do primeiro componente principal é mostrado por ser o que capturou a maior variância explicada dos dados originais, que foi de 74,39%. **Fonte:** O Autor (2024)

Na Figura 4.7-A, pode-se observar que as $n=4520$ amostras do grupo controle estão sobrepostas, dificultando a visualização. Para entender a origem dessa alta sobreposição realizamos uma análise de variância no qual construímos um score plot do modelo PCA usando apenas as amostras do grupo controle ($n=4520$) aplicando a mesma combinação dos mesmos métodos de pré-processamento (normalização + escala automática) usado para construir o modelo PCA mostrado na Figura 4.7-A do manuscrito, e observamos que uma alta sobreposição permaneceu, como pode ser observado na Figura 4.8-A. Na sequência, também construímos um novo score plot do modelo PCA das mesmas amostras, mas sem usar qualquer método de pré-processamento e observamos uma alta dispersão das amostras, tornando bem visível que as $n=4520$ amostras do grupo controle (Figura 4.8-B). Com base nessas análises, concluímos que a alta sobreposição entre as amostras de controle no grupo controle da Figura 4.7-A ($n=4520$) se deve à combinação de métodos de pré-processamento utilizados para otimizar o modelo PCA, ou seja, a combinação normaliza +escala automática.

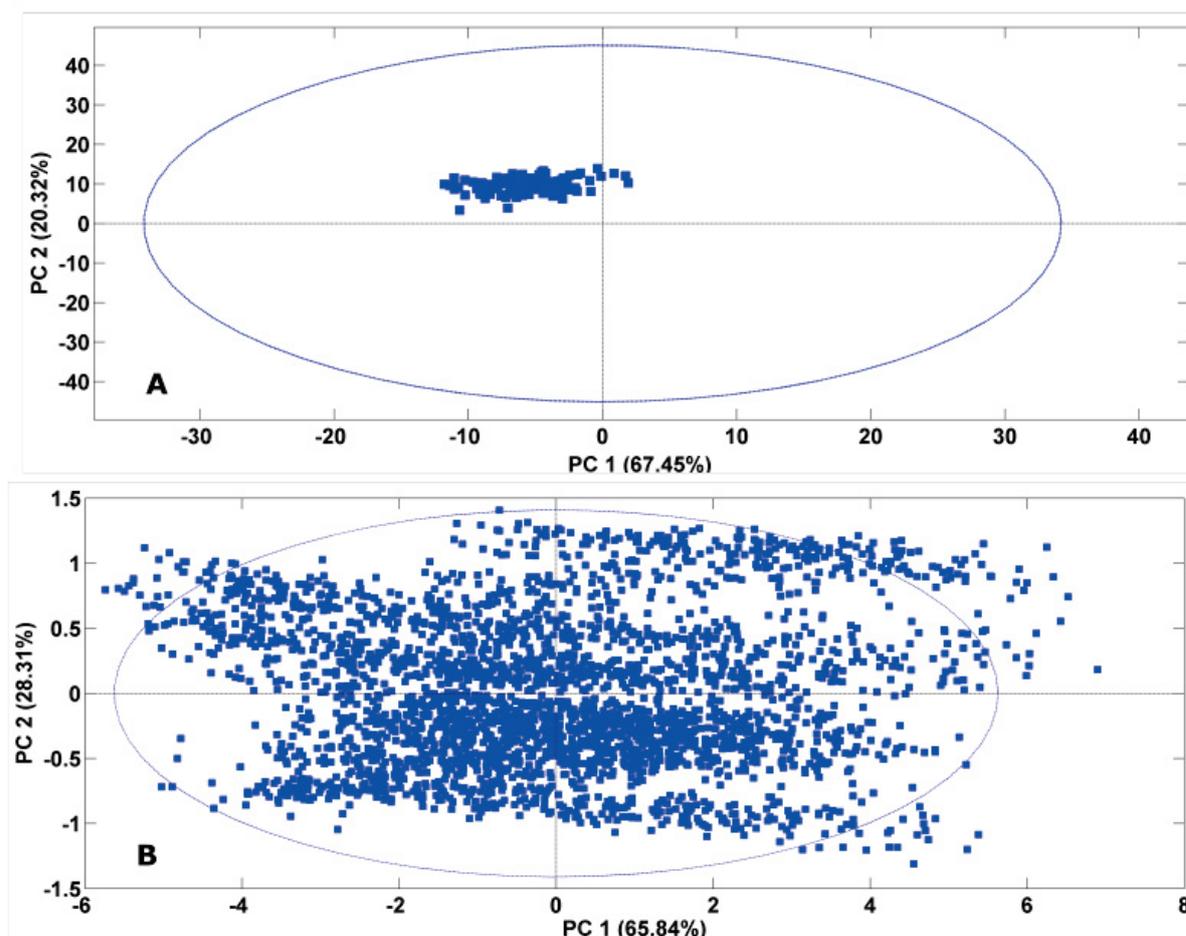


Figura 4.8. Modelo PCA dos pacientes do grupo controle ($n=4.520$). Cada triângulo azul representa uma amostra do grupo controle. em (A), o modelo PCA foi construído usando a seguinte combinação de pré-processamento: normalização + escala automática. Pelo gráfico, percebe-se que as amostras do grupo controle se sobrepõem, dando a falsa impressão de que foi utilizado um pequeno número de tamanhos de amostra, quando na verdade foram utilizadas 4.520 amostras. Em (B), o modelo PCA foi construído sem qualquer tipo de pré-processamento de dados. Pelo gráfico percebe-se que as amostras do grupo controle estão espalhadas. Portanto, conclui-se que a alta sobreposição de amostras 4.520 amostras observada em (A) foi causada pelos métodos de pré-processamento de dados. **Fonte:** O Autor (2024)

Na Figura 4.9 (Figura 4.9-A), o modelo PCA foi capaz de discriminar amostras de pacientes com HIV/AIDS, TB pulmonar, coinfeção HIV/TB e controle. De acordo com o gráfico de *loadings* de PC1 (Figura 4.9-B), os biomarcadores mais importantes na discriminação dos quatro grupos de pacientes (COVID-19, HIV, HIV/TB e controle) foram glóbulos vermelhos, hematócrito, concentração de volume corpuscular médio, plaquetas, linfócitos, neutrófilos, monócitos, creatinina, alanina transaminase, largura de distribuição de glóbulos vermelhos e glicose sérica.

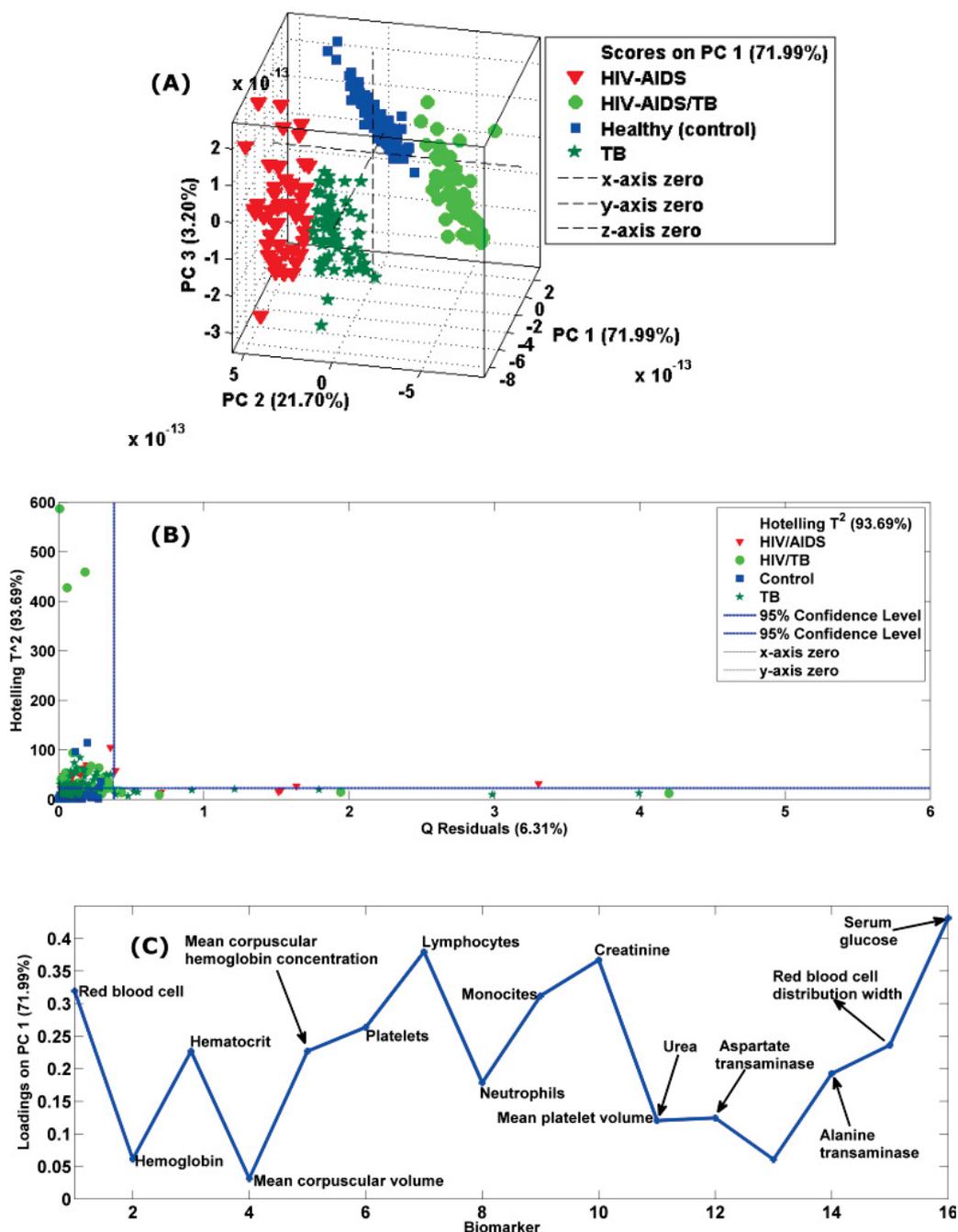


Figura 4.9. Análise exploratória do conjunto de dados de imunodeficiência humana (HIV)/AIDS, tuberculose (TB) e coinfeção HIV/TB. Em (A) é mostrado o modelo PCA das amostras de sangue de 49 pacientes com HIV representados pelos triângulos vermelhos; 113 pacientes com TB pulmonar estão representados pelas estrelas verdes; 80 pacientes co-infectados com HIV/TB estão representados pelos círculos verdes e 4.520 controles que testaram negativo para HIV ou TB estão representados pelos quadrados azuis (controle). Em (B) é mostrado o gráfico de Hotelling T^2 versus Q-resíduos do modelo PCA para detectar valores discrepantes em dados de amostras de pacientes com HIV, TB, co-infecção HIV/TB. Neste gráfico, uma amostra é

considerada outlier se e somente se apresentar simultaneamente altos valores de Hotelling e altos valores de Q-residual. De acordo com o gráfico, apesar de algumas amostras apresentarem valores elevados de Q-Resíduos, elas não podem ser consideradas outliers porque estão dentro do intervalo de confiança de 95% do Hotelling T^2 . Em (C) é mostrado o gráfico de cargas que representa as variáveis mais importantes na discriminação de amostras de HIV, TB, HIV/TB e controle. Apenas o gráfico de cargas do primeiro componente principal é mostrado por ser aquele que capturou a maior variância explicada dos dados originais, que foi de 71,99%.

4.5.2.3 Análise multivariada: modelos de machine learning de classificação

Antes de desenvolver os modelos de *machine learning* de classificação das amostras COVID-19, HIV, TB, co-infecção HIV/TB, foram construídos gráficos de *leverage* versus resíduos studentizados (Figura 4.10). Tanto para o conjunto de dados 1 (COVID-19 vs controle) quanto para o conjunto de dados 2 (COVID-19 vs HIV vs TB vs HIV/TB vs controle), embora existam algumas amostras que apresentaram valores de alavancagem elevados, nenhuma delas foi considerada como outliers porque todas as amostras estão dentro de $\pm 2,5$ desvios padrão dos resíduos dos alunos. Portanto, todas as amostras presentes nos dois conjuntos de dados foram utilizadas para desenvolver modelos de *machine learning* para prever o diagnóstico de COVID-19, HIV, TB e a detecção de coinfeção HIV/TB.

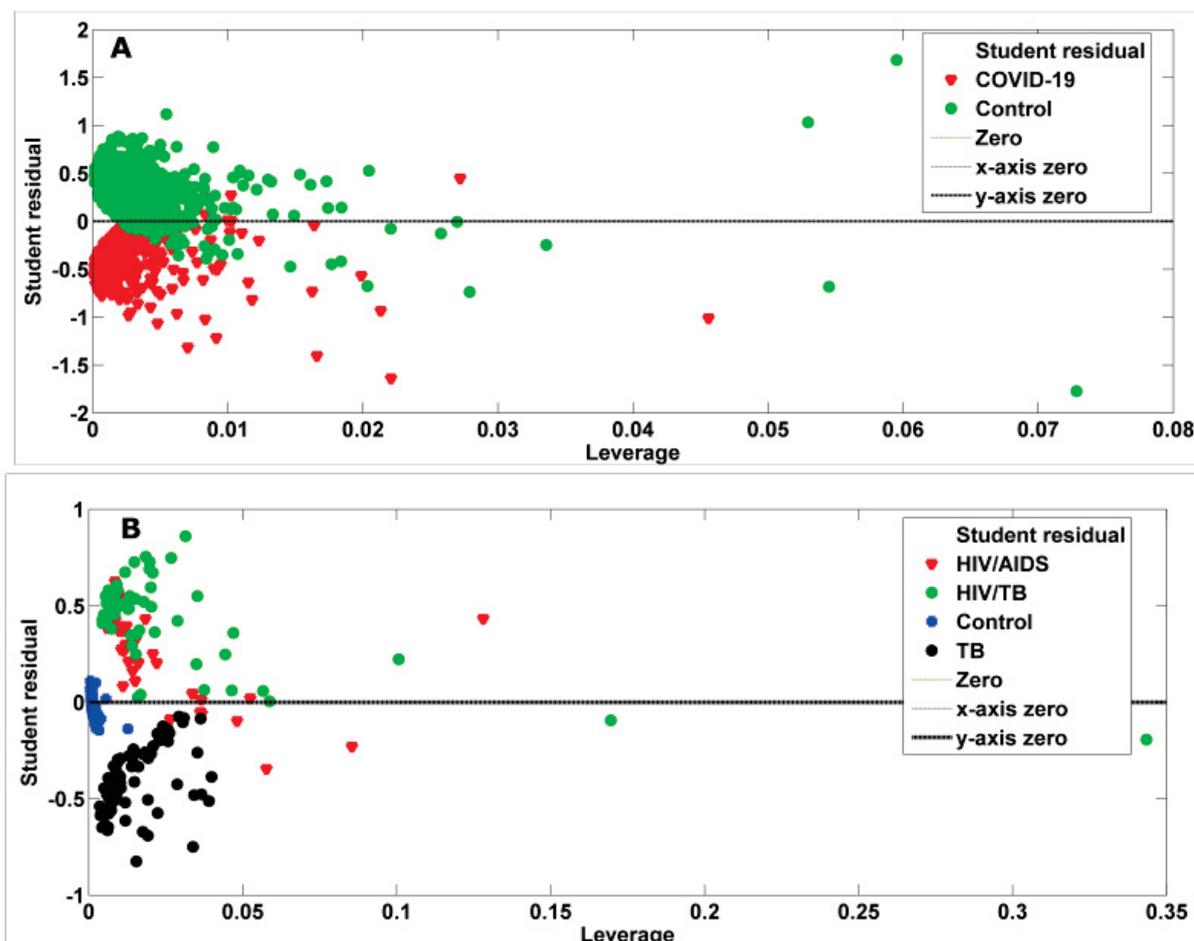


Figura 4.10. Gráfico de *leverage* versus resíduos de *student* para detecção de amostras outliers. Em (A) é referente ao conjunto de dados de pacientes com COVID-19 e no grupo de controle. Em (B) é referente ao conjunto de dados de pacientes HIV, TB, HIV/TB e pacientes de controle. Em (A), as amostras de COVID-19 e de controle são representadas por triângulos vermelhos e círculos verdes, respectivamente. Em (B), as amostras de HIV, TB, HIV/TB e controles são representadas por triângulos vermelhos, círculos verdes, círculos azuis e círculos pretos, respectivamente. Neste gráfico, uma amostra é considerada outlier se apresentar simultaneamente valores elevados de *leverage* e de resíduos de *student*. Embora algumas amostras tivessem valores de *leverage* elevados, todas as amostras estavam dentro do intervalo de $\pm 2,5$ desvios padrão dos resíduos dos alunos, pelo que não foram detectadas amostras discrepantes. **Fonte:** O Autor (2024).

Todos os modelos de ML de classificação de diferentes grupos de pacientes (COVID-19, HIV, TB, HIV/TB e controle) foram treinados e testados usando os mesmos métodos de processamento usados nas análises de PCA mencionadas acima (normalização + escala automática). Os resultados de desempenho de todos os sete modelos de *machine learning* (PLS-DA, ANN, XGBoosted, KNN, LREG, SIMCA e SVM) usados para prever o diagnóstico de COVID-19, HIV/AIDS, TB

pulmonar e coinfeção por HIV/ TB são apresentados na **Tabela 4.9**. O modelo PLS-DA foi o que apresentou maior desempenho na predição do diagnóstico de COVID-19, HIV, TB e HIV/TB devido à sua alta sensibilidade, especificidade, acurácia, precisão, escore F1 e valores de MCC, variando entre 88-96%. (Tabela 4.9). É importante ressaltar que os valores de sensibilidade (recall), especificidade, exatidão, precisão, escore F1 e coeficiente de correlação de Matthew (MCC) foram calculados através das equações 1, 2, 3, 4, 5 e 6 apresentadas no material e métodos seção. Esses modelos foram treinados considerando valores mais baixos de RMSECV conforme mostrado na **Figura 4.11** e **Figura 4.12**.

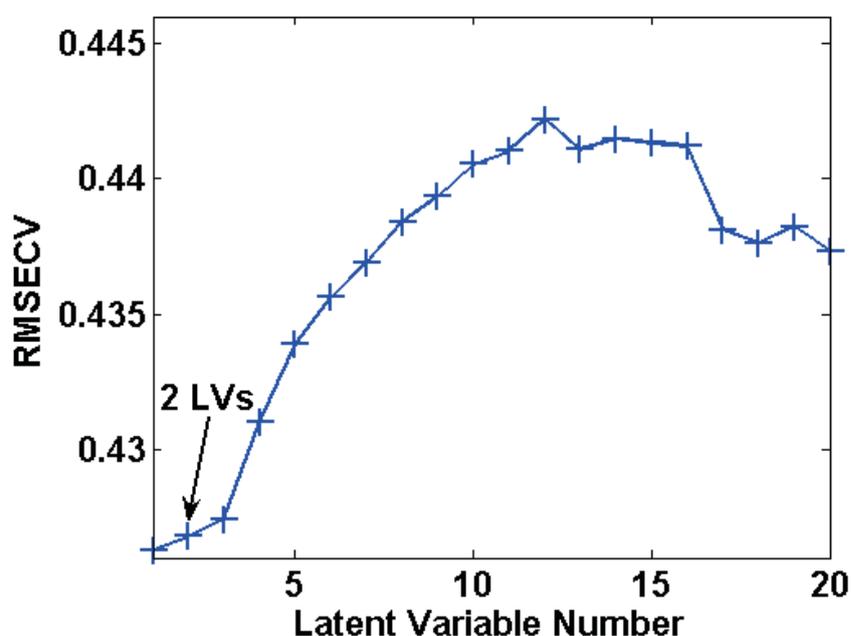


Figura 4.11. Raiz quadrática do erro médio de validação cruzada (RMSECV) versus número da variável latente. Um total de 2 variáveis latentes (LV) foram selecionadas para a construção do modelo PLS-DA para predição do diagnóstico de COVID-19, por apresentarem menores valores de RMSECV. **Fonte:** O Autor (2024).

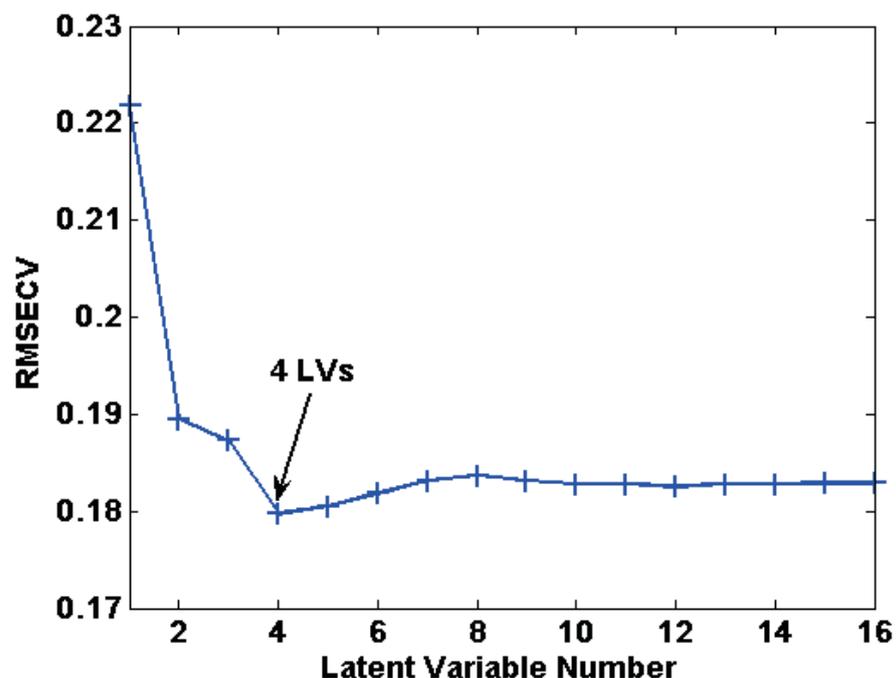


Figura 4.12. Raiz quadrática do erro médio de validação cruzada (RMSECV) versus número de variáveis latentes. Foram selecionadas um total de 4 variáveis latentes para a construção do modelo PLS-DA para predição do diagnóstico de HIV/AIDS, TB e coinfeção HIV/TB, por apresentarem menores valores de RMSECV. **Fonte:** O Autor (2024)

O modelo PLS-DA teve o melhor desempenho (maiores valores de sensibilidade e especificidade conforme tabela 4.9). As curvas ROC dos modelos diagnósticos PLS-DA de HIV/AIDS, TB pulmonar e coinfeção HIV/TB são apresentadas na Figura 4.13.

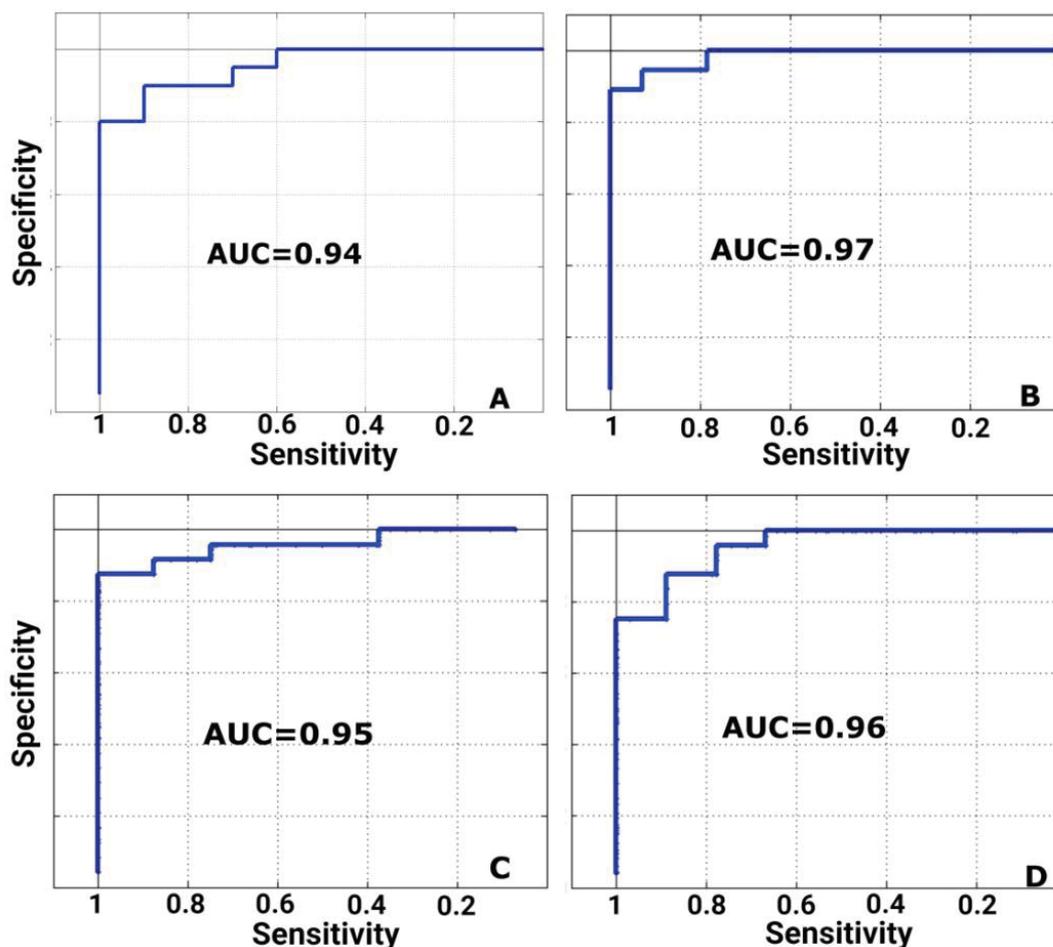


Figura 4.13. Área sob a curva ROC do desempenho do modelo PLS-DA para a perda de diagnóstico de pacientes COVID-19, HIV, TB, e co-infectados HIV/TB. A área sob as curvas (AUC) reflete a precisão dos modelos PLS-DA na previsão de pacientes de diferentes classes de pacientes [(COVID-19, vírus da imunodeficiência humana (HIV)/síndrome da imunodeficiência adquirida (AIDS), tuberculose (TB) e (coinfecção HIV/TB)] e controle. As curvas incluem ambos os conjuntos de amostras (amostras de treinamento e de teste). Os valores de AUC para prever a coinfecção por COVID-19, HIV/AIDS e HIV/TB foram 94 (A), 97 (B), 95 (C) e 96% (D), respectivamente.

Tabela 4.9. Comparação de desempenho dos modelos de *machine learning* para COVID-19, HIV/AIDS, Tuberculose e coinfeção HIV/TB.

Classe	Modelo	VP	FN	VN	FP	Sensibilidade	Especificidade	Accurácia	Precisão	F1 score	CCM	
COVID-19	PLS-DA	226	19	265	11	0,92	0,96	0,94	0,95	0,94	0,88	
	ANN	224	21	248	28	0,91	0,90	0,91	0,89	0,90	0,81	
	KNN	219	26	259	17	0,89	0,94	0,92	0,93	0,91	0,83	
	SVM	205	40	259	17	0,84	0,94	0,89	0,92	0,88	0,78	
	SIMCA	238	7	244	32	0,97	0,88	0,93	0,88	0,92	0,85	
	XGboost	210	35	258	18	0,86	0,93	0,90	0,92	0,89	0,80	
	LREG	213	32	264	12	0,87	0,96	0,92	0,95	0,91	0,83	
	PLS-DA	15	0	1,310	46	1,00	0,97	0,97	0,25	0,39	0,49	
	ANN	12	3	1,273	83	0,80	0,94	0,94	0,13	0,22	0,30	
	KNN	12	3	1,250	106	0,80	0,92	0,92	0,10	0,18	0,27	
HIV	SVM	11	4	1,261	95	0,73	0,93	0,93	0,10	0,18	0,26	
	SIMCA	13	2	1,213	143	0,87	0,89	0,89	0,08	0,15	0,25	
	XGboost	15		1,255	101	1,00	0,93	0,93	0,13	0,23	0,35	
	LREG	11	4	1,236	120	0,73	0,91	0,91	0,08	0,15	0,23	
	PLS-DA	31	3	1,251	105	0,91	0,92	0,92	0,23	0,36	0,43	
	ANN	29	5	1,247	109	0,85	0,92	0,92	0,21	0,34	0,40	
	KNN	29	5	1,247	109	0,85	0,92	0,92	0,21	0,34	0,40	
	SVM	31	3	1,220	136	0,91	0,90	0,90	0,19	0,31	0,39	
	SIMCA	25	9	1,207	149	0,74	0,89	0,89	0,14	0,24	0,29	
	XGboost	29	5	1,251	105	0,85	0,92	0,92	0,22	0,35	0,41	
TB	LREG	24	10	1,257	99	0,71	0,93	0,92	0,20	0,31	0,34	
	PLS-DA	22	2	1,343	13	0,92	0,99	0,99	0,63	0,75	0,75	
	ANN	21	3	1,255	101	0,88	0,93	0,92	0,17	0,29	0,37	
	KNN	20	4	1,238	118	0,83	0,91	0,91	0,14	0,25	0,33	
	SVM	18	6	1,243	113	0,75	0,92	0,91	0,14	0,23	0,30	
	HIV/TB	PLS-DA	22	2	1,343	13	0,92	0,99	0,99	0,63	0,75	0,75
		ANN	21	3	1,255	101	0,88	0,93	0,92	0,17	0,29	0,37
		KNN	20	4	1,238	118	0,83	0,91	0,91	0,14	0,25	0,33
		SVM	18	6	1,243	113	0,75	0,92	0,91	0,14	0,23	0,30

SIMCA	11	13	1,214	142	0,46	0,90	0,89	0,07	0,12	0,15
XGboost	20	4	1,252	104	0,83	0,92	0,92	0,16	0,27	0,35
LREG	19	5	1,273	83	0,79	0,94	0,94	0,19	0,30	0,36

Nota: PLS-DA: *partial least squares discriminant analysis* (PLS-DA); ANN: *artificial neural network*; XGBoosted: eXtreme Gradient Boosting, KNN: *K-Nearest Neighbors*, LREG: Regressão logística; SIMCA: *Soft independent modeling by class analogy*, SVM: *Support vector machine*; TB: tuberculose, HIV: Vírus de imunodeficiência humana; AIDS: síndrome de imunodeficiência adquirida; TP: verdadeiro positivo, TN: verdadeiro negativo; FP: falso positivo e FN: falso negativo. MCC: coeficiente de correlação de Matthew.

Fonte: O Autor (2024).

A Figura 4.14 e Figura 4.15 demonstram os gráficos VIP com os biomarcadores que mais contribuíram para o diagnóstico. O eixo X representa cada biomarcador utilizado no estudo. O eixo Y mostra a pontuação VIP que corresponde à importância de cada biomarcador na predição do diagnóstico de COVID-19, HIV e TB. Valores VIP superiores a 1 (valores acima da linha horizontal vermelha tracejada) são considerados significativamente importantes na previsão do diagnóstico. Assim, os seguintes biomarcadores foram associados ao diagnóstico de COVID-19: cálcio, lactato desidrogenase (LDH), glóbulos brancos (leucócitos), glóbulos vermelhos (hemácias), hemoglobina, hematócrito, contagem de neutrófilos, aspartato aminotransferase, contagem de basófilos, e contagem de eosinófilos (Figura 4.14). As diferenças nos níveis de biomarcadores entre o grupo de pacientes com COVID-19 e os controles são mostradas na **Tabela 4.8**. É importante destacar que esses biomarcadores também foram identificados pelo modelo PCA como importantes na discriminação dos dois grupos de pacientes (COVID-19 vs. ao controle).

Os parâmetros como volume corpuscular médio (VCM), plaquetas, neutrófilos e volume plaquetário médio (VPM) foram associados à infecção pelo HIV (Figura 4.15). As plaquetas, neutrófilos, largura de distribuição de glóbulos vermelhos (RBCD), ureia e glicose sérica foram relacionadas à infecção por *Mycobacterium tuberculosis* (Figura 4.15). Os seguintes biomarcadores foram associados à coinfeção HIV/TB: glóbulos vermelhos, hemoglobina, hematócrito, plaquetas, linfócitos, neutrófilos, largura de distribuição de glóbulos vermelhos (RBCD), aspartato transaminase (AST), alanina transaminase (ALT) e soro glicose (Figura 4.15). As diferenças nos níveis de biomarcadores entre o grupo de pacientes com coinfeção HIV, TB, HIV/TB e os controles são mostradas na **Tabela 4.8**. É importante destacar que esses biomarcadores também foram identificados pelo modelo PCA como importantes na discriminação dos quatro grupos de pacientes.

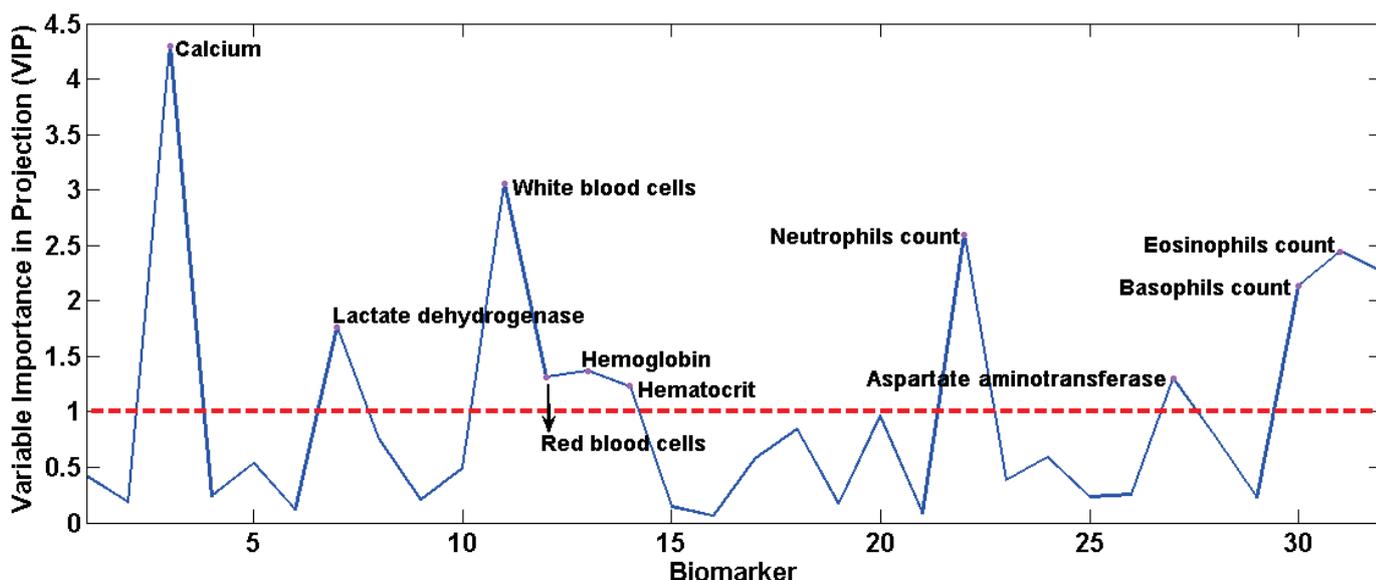


Figura 4.14. Gráfico de Importância da Variável na Projeção (VIP) dos biomarcadores mais importantes para diagnóstico de COVID-19. O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na predição das diferentes classes das amostras (COVID-19 e controle). A linha tracejada vermelha paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP = 1). Os metabólitos que contribuem significativamente para prever o diagnóstico de COVID-19 estão acima do limite (pontuação VIP > 1). **Fonte:** O Autor (2024).

4.5.2.4 Aplicabilidade prática do modelo de machine learning

A robustez do modelo PLS-DA foi avaliada minuciosamente usando 1.228 amostras externas provenientes de pacientes no Brasil. Esses indivíduos tiveram resultados negativos para COVID-19, HIV e TB, mas apresentaram um espectro de outras condições, incluindo diabetes, hipercolesterolemia, hipertrigliceridemia, hipotireoidismo e obesidade. O modelo demonstrou desempenho excepcional, classificando com precisão esses casos com uma sensibilidade de 98% (1.205 de 1.228 amostras) e uma taxa de erro de apenas 2%. Isto sublinha a elevada eficácia preditiva do modelo em cenários clínicos do mundo real (**Tabela 4.10**).

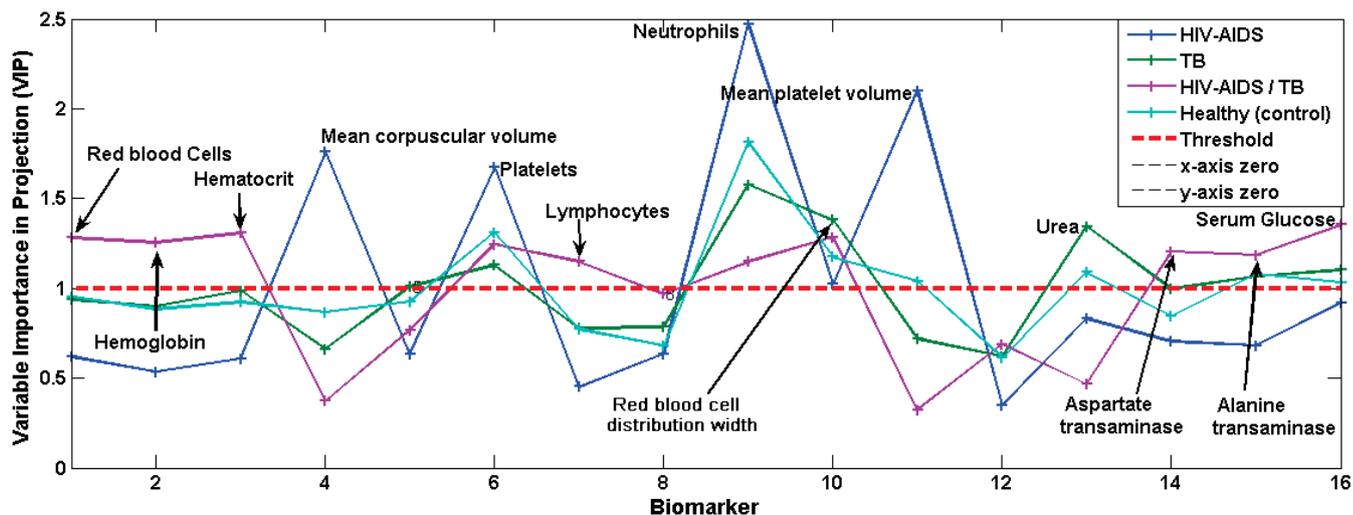


Figura 4.15. Gráfico de importância variável na projeção (VIP) dos biomarcadores mais importantes para o diagnóstico de HIV, TB e coinfeção HIV/TB. O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na predição das diferentes classes das amostras (HIV representado pela cor azul, TB representada pela cor verde, coinfeção HIV/TB representada pela cor rosa e saudável representado pela cor azul claro). A linha tracejada vermelha paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP = 1). Os metabólitos que contribuem significativamente para a previsão das diferentes classes das amostras estão acima do limite (pontuação VIP > 1). **Fonte:** O Autor (2024).

Tabela 4.10. Avaliação do desempenho preditivo do modelo PLS-DA na predição de amostras externas de pacientes sem nenhuma das doenças estudadas (COVID-19, HIV e TB), mas com outras comorbidades.

Grupo de Pacientes	Total	VP	FN	Taxa de não erro	Taxa de erro
Diabetes	86	85	1	0,988	0,012
Dislipidemias*	282	275	7	0,975	0,025
hipotireoidismo	184	181	3	0,984	0,016
Obesidade	676	664	12	0,982	0,018
Total	1228	1205	23	0,981	0,019

Nota:*Pacientes com hipercolesterolemia ou hipertrigliceridemia; VP: verdadeiro positivo; FN: falso negativo. **Fonte:** O autor (2024)
Fonte: O Autor (2024)

4.5.3 Estudo III: Predição de diagnóstico de COVID-19 usando dados clínicos de pacientes COVID-19 atendidos na rede de Farmácia Drugstore distribuída em todo território Nacional.

No estudo III, as análises exploratórias utilizando a abordagem PCA foram capazes de discriminar amostras de pacientes com resultados negativos e positivos para COVID-19 tanto nos testes de antígeno (Figura 4.16), IgG (Figura 4.17) quanto IgM (Figura 4.18). Nenhuma amostra discrepante foi detectada.

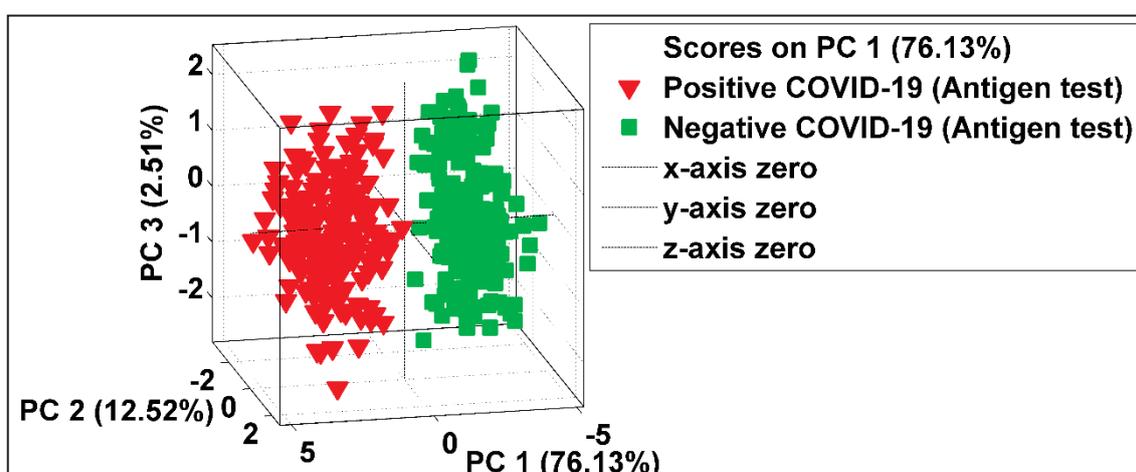


Figura 4.16. Análise exploratória dos dados. Modelo PCA de pacientes com COVID-19 testados pelo método de antígeno. O PCA foi capaz de discriminar entre pacientes positivos (triângulos vermelhos) e negativos (quadrados verdes) com COVID-19.

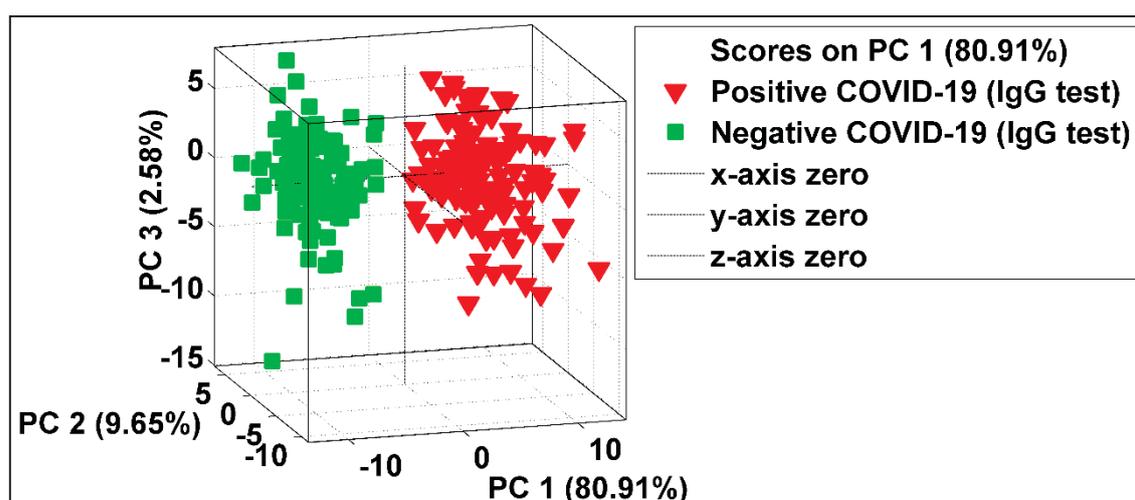


Figura 4.17. Análise exploratória dos dados. Modelo PCA de pacientes com COVID-19 testados pelo método anti-IgG. O PCA foi capaz de discriminar entre pacientes positivos (triângulos vermelhos) e negativos (quadrados verdes) com COVID-19.
Fonte: O Autor (2024)

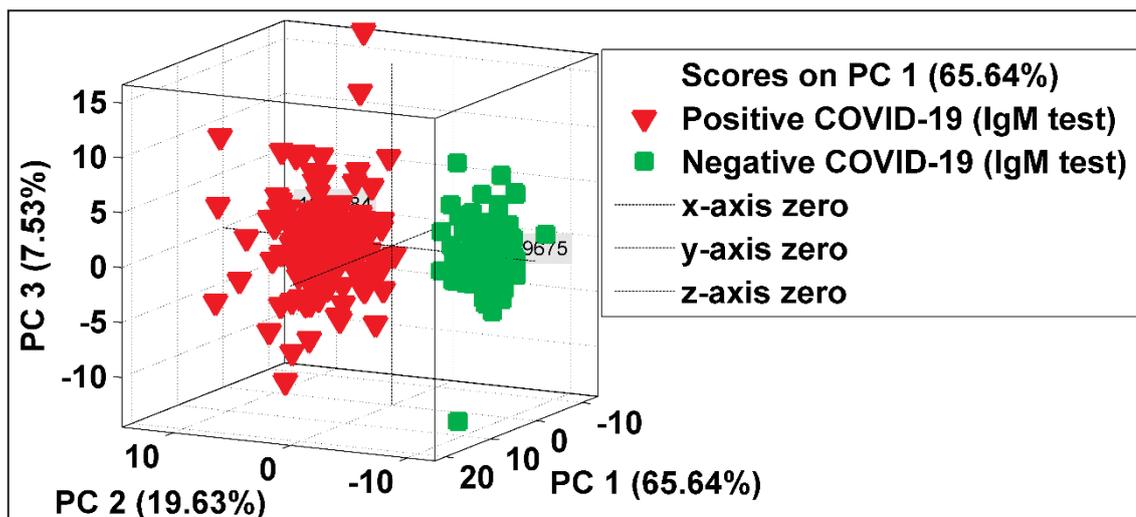


Figura 4.18. Análise exploratória dos dados. Modelo PCA de pacientes com COVID-19 testados pelo método anti-IgM. O PCA foi capaz de discriminar entre pacientes positivos (triângulos vermelhos) e negativos (quadrados verdes) com COVID-19. **Fonte:** O Autor (2024)

Os cinco melhores modelos de ML para dados de pacientes testados com o método de antígeno (Dataset 3) foram: LIGHTGBM, EGB, GBC, ADA e LR com valores de precisão diagnóstica de 0,76, 0,76, 0,76, 0,76 e 0,76, respectivamente. O tempo de treinamento desses modelos variou entre 12 e 57 segundos (Tabela 4.11). Os resultados da validação cruzada também confirmaram a robustez destes dados (valores CV <0,1%).

Tabela 4.11. Desempenho de modelos de Machine Learning para predição do diagnóstico de COVID-19

Teste de antígeno COVID-19 (Conjunto de dados 3)			Teste anti-IgG COVID-19 (Conjunto de dados 4)		
Modelo de <i>Machine Learning</i>	Acurácia	Tempo de treinamento (segundos)	Modelo de <i>Machine Learning</i>	Acurácia	Tempo de treinamento (segundos)
<i>Light Gradient Boosting Machine</i>	0,76	14,01	<i>Gradient Boosting Classifier</i>	0,71	6,99
<i>Extreme Gradient Boosting</i>	0,76	56,82	<i>Light Gradient Boosting Machine</i>	0,71	6,75
<i>Gradient Boosting Classifier</i>	0,76	45,88	<i>Quadratic Discriminant Analysis</i>	0,71	0,31
<i>Ada Boost Classifier</i>	0,76	14,52	<i>Ada Boost Classifier</i>	0,71	2,59
<i>Logistic Regression</i>	0,76	12,31	<i>Dummy Classifier</i>	0,71	0,22
<i>Ridge Classifier</i>	0,76	2,05	<i>Ridge Classifier</i>	0,71	0,23
<i>Linear Discriminant Analysis</i>	0,76	2,58	<i>Linear Discriminant Analysis</i>	0,71	0,49
<i>Quadratic Discriminant Analysis</i>	0,76	1,17	<i>Extreme Gradient Boosting</i>	0,71	9,12
<i>Dummy Classifier</i>	0,76	0,85	<i>Extra Trees Classifier</i>	0,70	6,04
<i>SVM - Linear Kernel</i>	0,76	13,32	<i>Decision Tree Classifier</i>	0,70	0,49
<i>Decision Tree Classifier</i>	0,76	2,66	<i>Random Forest Classifier</i>	0,70	6,95
<i>Extra Trees Classifier</i>	0,76	39,60	<i>Random Forest Classifier</i>	0,70	6,95
<i>Random Forest Classifier</i>	0,76	51,62	<i>Naive Bayes</i>	0,69	0,28
<i>K Neighbors Classifier</i>	0,71	58,21	<i>K Neighbors Classifier</i>	0,64	3,68
<i>Naive Bayes</i>	0,64	1,66	<i>SVM - Linear Kernel</i>	0,63	1,66
Teste anti-IgM COVID-19 (Conjunto de dados 5)		
<i>Machine Learning model</i>

<i>Quadratic Discriminant Analysis</i>
<i>Ada Boost Classifier</i>	0,85	2,82
<i>Gradient Boosting Classifier</i>	0,85	6,99
<i>Dummy Classifier</i>	0,85	0,23
<i>Naive Bayes</i>	0,85	0,28
<i>Light Gradient Boosting Machine</i>	0,85	4,84
<i>Logistic Regression</i>	0,85	2,56
<i>Ridge Classifier</i>	0,85	0,25
<i>Linear Discriminant Analysis</i>	0,85	0,46
<i>Extreme Gradient Boosting</i>	0,85	9,29
<i>Extra Trees Classifier</i>	0,84	6,00
<i>Random Forest Classifier</i>	0,84	6,97
<i>Decision Tree Classifier</i>	0,84	0,59
<i>K Neighbors Classifier</i>	0,83	4,01
<i>SVM - Linear Kernel</i>	0,72	1,44

Fonte: O Autor (2024)

4.6 DISCUSSÃO

4.6.1 Estudo I: Análise de dados dos exames bioquímicos, hematológicos e de urinálise de pacientes COVID-19 atendidos no Hospital Israelita Albert Einstein (Brasil) visando a predição do diagnóstico e investigação de potenciais Biomarcadores prognósticos.

No estudo I, usamos quatro modelos baseados em ML (ANN, DT, PLSDA e KNN) com mais de 5.000 amostras de RT-PCR e dados sobre parâmetros bioquímicos, hematológicos e urinários dos pacientes que previram efetivamente o diagnóstico e a doença de COVID-19 gravidade no Brasil. Considerando a elevada probabilidade de mutações existentes no SARS-Cov-2 (B.1.1.7, P.1 e P.2) na amostra, a complexidade destes modelos é ainda maior ^{285,286}.

Vários modelos baseados em ML, incluindo abordagens não supervisionadas (por exemplo, PCA e análise de cluster hierárquica) e modelos supervisionados (por exemplo, rede neural artificial, PLS-DA, DT, KMC e KNN), estão disponíveis na literatura científica ²⁸⁷. O desempenho desses modelos depende de vários fatores, incluindo o tamanho da amostra e o tipo de dados. Modelos baseados em ML construídos com amostras maiores são geralmente mais precisos e eficientes para previsão de séries; por exemplo, as redes neurais profundas que requerem uma grande quantidade de dados de treinamento ^{288–290}. Quanto maior a arquitetura da rede, mais dados serão necessários para obter modelos mais robustos^{291–293}.

Em relação à COVID-19, modelos anteriores foram implementados com o objetivo de prever o comportamento e a gravidade da doença. No entanto, a maioria desses estudos utilizou um tamanho de amostra pequeno, o que pode impactar diretamente o desempenho do modelo ^{294–296}. Banerjee (2020) desenvolveu um algoritmo de ML para prever o diagnóstico de COVID-19 utilizando uma base de dados pública com 598 pacientes, dos quais apenas 39 foram positivos para SARS-CoV-2. Os autores obtiveram um modelo com boa especificidade (91%), mas baixa sensibilidade (43%), o que pode impedir a utilização do modelo na prática para diagnóstico precoce da doença ²⁹⁷. Da mesma forma, Joshi (2020) implementou um modelo de regressão logística previamente treinado com 390 amostras, das quais apenas 33 foram positivas para COVID-19, comprovando valores de sensibilidade e especificidade de 93% e 43%, respectivamente ²⁹⁸. Além disso, a maioria dos estudos

implementou modelos baseados em ML usando apenas exames de sangue de rotina^{299,300}. Em nosso estudo, além do hemograma completo, também foram incluídos dados de exames bioquímicos, urinários, bacteriológicos e virológicos dos pacientes, visando identificar outros biomarcadores associados à COVID-19.

Os modelos alcançaram acurácia de 84% a 98%. Os biomarcadores que mais contribuíram para este resultado e predizem o diagnóstico e gravidade da COVID-19 incluíram: hiperferritinemia, hipocalcemia (níveis baixos de cálcio ionizado), hipoxemia (pressão arterial de oxigênio baixa, pO₂), hipóxia pulmonar (baixa fração inspirada de oxigênio arterial, FiO₂), acidose respiratória (níveis elevados de CO₂ total e pCO₂), acidose metabólica (níveis elevados de ácido láctico e pH venoso baixo), pH urinário baixo e níveis elevados de lactato desidrogenase (LDH). Valores de acurácia semelhantes foram relatados recentemente por Zhou (2021) (94%) e Wu (2021) (90%) após a implementação de modelos baseados em ML para diagnóstico de COVID-19 e gravidade da doença, respectivamente^{301,302}. Contudo, diferentes variáveis foram destacadas pelos autores como importantes para a classificação dos dados. De acordo com Zhou (2021), essas eram as taxas de linfócitos circulantes, enquanto Wu (2021) relatou as taxas de neutrófilos e linfócitos, a proporção neutrófilos/linfócitos e a proporção plaquetas/linfócitos^{301,302}. Esta variação pode estar associada aos diferentes tamanhos de amostra (n = 357 vs n = 51) e tipo de modelos (modelo de árvore de decisão vs modelo de máquina de vetores de suporte).

Recentemente, a maioria dos pacientes com COVID-19 grave que necessitam de hospitalização em unidades de terapia intensiva desenvolve uma forma atípica de síndrome do desconforto agudo, que geralmente é acompanhada por um volume preservado de gás pulmonar³⁰³. Isso sugere hipóxia, que resulta da dificuldade em realizar trocas gasosas ao nível dos alvéolos pulmonares. Foi observado que essa disfunção pulmonar também pode comprometer o metabolismo do ferro. Desequilíbrios nos níveis de hemoglobina e ferritina foram relatados em pacientes com doença grave ou mortes causadas por COVID-19. Numa revisão sistemática com meta-análise conduzida por Taneri (2020), que incluiu 189 estudos (n = 57.563 pacientes), pacientes de alto risco com doença grave apresentaram níveis significativamente elevados de ferritina (diferença média ponderada (DMP), 473,25 ng/mL (IC 95% 382,52; 563,98)) e baixos níveis de hemoglobina (ADM, 4,08 g/L (IC 95% 5,12; - 3,05)) quando comparados com pacientes com doença de risco moderado ou baixo³⁰⁴. Em nosso estudo, tanto os grupos de pacientes com a doença (modelo

diagnóstico) quanto os pacientes com doença grave (modelo de gravidade da doença) apresentaram níveis baixos de hemoglobina e níveis elevados de ferritina. Todos os modelos baseados em ML desenvolvidos (ANN, PLS-DA, KNN e DT) destacaram que as diferenças nos níveis desses dois biomarcadores eram críticas, tanto para a previsão do diagnóstico de COVID-19 quanto para a gravidade da doença. Estudos recentes indicaram que a anemia e a hiperferritinemia são fortes biomarcadores para o prognóstico de mortalidade por SARS-CoV-2, além de outras doenças respiratórias graves ³⁰⁵⁻³⁰⁹.

A ferritina é uma proteína encontrada principalmente no fígado, medula óssea e baço e é a principal fonte de armazenamento de ferro do corpo. É considerado um biomarcador chave da desregulação imunitária, principalmente numa situação de hiperferritinemia, através da via direta de efeitos pró-inflamatórios e imunossupressores que contribuem para uma tempestade de citocinas ³¹⁰. Alguns estudos mostram que os desfechos fatais causados pela COVID-19 são acompanhados pela síndrome da tempestade de citocinas envolvendo altos níveis de marcadores inflamatórios, como a ferritina ^{311,312}. No presente estudo, a ferritina foi associada à predição da gravidade da COVID-19 e foi o biomarcador mais importante na predição do diagnóstico da doença por todos os modelos baseados em ML desenvolvidos, corroborando os dados da literatura.

Os distúrbios ácidos e básicos são indicadores importantes na patogênese e gravidade de diversas doenças, especialmente doenças respiratórias de origem infecciosa, como a pneumonia ³¹³⁻³¹⁵. A acidose pode ocorrer como resultado de um aumento significativo na pressão arterial de dióxido de carbono (acidose respiratória) ou de uma variedade de compostos inorgânicos ou orgânicos (acidose metabólica), como bicarbonato, ácido láctico arterial, cetonas, ou como resultado de insuficiência renal ou acidose hiperclorêmica; todos esses fatores atuam simultaneamente no aumento de prótons de hidrogênio e, conseqüentemente, na redução dos níveis de pH sanguíneo e respiratório ³¹⁶⁻³¹⁸. Os investigadores sugerem que a acidose metabólica causada pelo ácido láctico na COVID-19 se deve provavelmente à glicólise anaeróbica, que é favorecida em consequência da hipoxemia. Nessa condição, o piruvato, produto da via glicolítica, não é translocado para as mitocôndrias para acompanhar o processo oxidativo ^{319,320}; em vez disso, é convertido em lactato no citosol pela enzima LDH. Como a hipoxemia prejudica a oxigenação tecidual e a fosforilação oxidativa, as células obtêm ATP por meio da glicólise anaeróbica. Esse

fluxo depende da conversão do piruvato em lactato, o que resulta em altos níveis desse metabólito que sai das células. O consumo excessivo de lactato durante o processo de gliconeogênese culmina em acidose láctica³²¹. Foi relatado que no 18º dia da doença COVID-19, os níveis de ácido láctico começam a aumentar significativamente, desencadeando acidose metabólica, embora a pressão de dióxido de carbono seja aceitável³²².

Em nosso estudo, pacientes com diagnóstico positivo de COVID-19 (modelo de diagnóstico) e com doença grave (modelo de gravidade da doença) apresentaram níveis elevados de pressão de dióxido de carbono (análise de gases arteriais e venosos), dióxido de carbono total (análise de gases arteriais e venosos), ácido láctico e bicarbonato arterial (análise de gases arteriais) e pH venoso excepcionalmente baixo (pH = 7,3), o que também sugere acidose respiratória e metabólica. Estes resultados são semelhantes aos de outros países, indicando que estes desequilíbrios metabólicos são prevalentes em pacientes com COVID-19^{323–325}. Em geral, todos os dados disponíveis mostram que a maioria dos pacientes com doença grave apresenta comorbidades, como diabetes. Estudos recentes indicam que a acidose metabólica é influenciada pelo uso da metformina no tratamento do diabetes mellitus^{326,327}.

A presente análise detectou níveis de cinco biomarcadores de função, incluindo bilirrubina, bilirrubina direta, bilirrubina indireta, alanina transaminase e aspartato transaminase, aumentados em pacientes com doença positiva (modelo de diagnóstico) e grave (modelo de gravidade) em comparação com pacientes com doença não grave. Uma revisão sistemática e meta-análise realizada por Parohan (2020) revelou resultados semelhantes, onde todos os 1.455 pacientes com doença grave apresentavam níveis extremamente elevados de bilirrubina total (ADM 2,30 mmol/l; IC 95%, 1,24; 3,36; $p < 0,001$), alanina aminotransferase (ADM 7,35 U/L; IC 95%, 4,77; 9,93; $p < 0,001$) e aspartato aminotransferase (ADM 8,84 U/L; IC 95% 5,97; 11,71; $p < 0,001$), em comparação com 1.973 pacientes com não -doença grave³²⁸. Danos hepáticos também foram relatados em outras pneumonias virais (por exemplo, MERS e SARS) e estão diretamente associados à gravidade e mortalidade da doença [72–75]. Contudo, os mecanismos bioquímicos/fisiopatológicos que explicam a disfunção hepática causada pela COVID-19 ainda são desconhecidos. Não está claro se a disfunção hepática é devida ao SARS-Cov-2 ou é uma consequência da falência de múltiplos órgãos causados pelo vírus³²⁸.

Além disso, foram encontrados níveis elevados de proteína C reativa (PCR) nas amostras. Um aumento neste importante biomarcador em pacientes com COVID-19 foi relatado anteriormente por Chen (2020) (até 86%)³²⁹. Recentemente, uma revisão sistemática com meta-análise concluiu que níveis extremamente elevados de PCR estavam estatisticamente associados à gravidade da COVID-19 [77,78]. A PCR é uma proteína inflamatória na fase aguda de processos inflamatórios e infecciosos que é sintetizada principalmente nas células do fígado, mas também nas células musculares lisas, macrófagos, linfócitos e adipócitos. Altos níveis de PCR (aumentando até 100 vezes) são comumente encontrados durante infecções (os níveis plasmáticos de PCR aumentam cerca de 1–500 µg/mL em 24–72 horas) [76]. No entanto, o papel das isoformas da PCR e o seu envolvimento na progressão de doenças infecciosas ainda é desconhecido^{330,331}.

Por fim, além dos parâmetros bioquímicos já mencionados, também foram detectados baixos níveis de cálcio em nossa análise como preditor do diagnóstico de COVID-19 e da gravidade da doença. O cálcio é essencial para uma ampla variedade de processos no corpo, desde a contração muscular normal até atividades enzimáticas. Sabe-se que alterações na homeostase do Ca²⁺ podem contribuir para a morte celular por necrose e apoptose. As evidências mostram que os distúrbios do metabolismo do cálcio estão associados a doenças cardiovasculares e morte celular precoce^{332,333}. Níveis baixos de cálcio já foram associados ao aumento da mortalidade hospitalar em pacientes com doença arterial coronariana grave³³⁴, pacientes sépticos³³⁵, pneumonia bacteriana e pacientes com dengue³³⁶. Estudos mais recentes também associaram a hipocalcemia como um importante preditor de hospitalização e risco de mortalidade por COVID-19^{337–339}. Embora o presente estudo tenha mostrado resultados consistentes, ele apresenta algumas limitações. Estudos transversais, sem análise de acompanhamento dos dados dos pacientes, são propensos a viés de seleção, viés de informação e viés de confusão. Além disso, os níveis dos biomarcadores podem mudar durante a doença.

4.6.2 Estudo II: Análise de dados dos exames bioquímicos e hematológicos de pacientes COVID-19, HIV, Tuberculose e co-infectados HIV/TB atendidos em um Hospital Regional de referência do Norte de Moçambique (Hospital Geral de Marrere, Província de Nampula) visando predição do diagnóstico e investigação de biomarcadores associados a essas doenças.

No estudo II, vários modelos de *machine learning* (PLS-DA, LREG, KNN, XGBoost, SIMCA, ANN e SVM) foram testados para prever o diagnóstico de COVID-19, HIV/AIDS, TB e coinfeção HIV/TB, com base em dados bioquímicos. Considerando que os dados foram coletados entre abril-novembro de 2021 e considerando as recentes novas mutações registradas em amostras de pacientes com SARS-CoV-2 (por exemplo, lambda, gama, alfa e beta)^{285,340,341}, HIV (por exemplo, mutações 67N, 70R, 184V, 219Q, M184L e M184T),^{23,24} e *M. tuberculosis* (por exemplo, mutações V91W e delta 438A)³⁴², os modelos de ML deste estudo são complexos. Os resultados deste estudo mostraram alta acurácia diagnóstica na detecção de SARS-Cov-2 (AUC=0,94), HIV (AUC=0,95), TB (AUC=0,97) e coinfeção HIV/TB (AUC=0,96). Esses valores são semelhantes a outros modelos de ML de COVID-19, HIV e *M. tuberculosis* disponíveis na literatura (AUC ROC 70-99%)^{261,343}. É importante destacar que, em dados da vida real envolvendo pacientes com COVID-19 (ou HIV), os dados são quase sempre desequilibrados devido à prevalência da doença, pois há muitos pacientes com resultado de teste negativo do que positivo, como pode ser visto nos estudos recentes de Zunin (2022) e Alves (2021)^{344,345}. Portanto, esse problema de desequilíbrio também foi observado em nosso estudo.

Em nosso estudo anterior¹², utilizando amostras de plasma e soro de pacientes com COVID-19 analisadas por LC-MS (dados metabolômicos), o modelo PLS-DA foi capaz de prever o diagnóstico e a gravidade da COVID-19 com uma precisão em torno de 93%. Além disso, foram identificados novos potenciais biomarcadores de diagnóstico e gravidade (ribotimidina, N-acetil-glucosamina-1-fosfato, L-ornitina e 5,6-di-hidro-5-metiluracil)¹². Por outro lado, no presente estudo, propomos um novo método para diagnóstico de COVID-19 utilizando dados de exames de rotina de pacientes (exames bioquímicos e hematológicos). Embora ambos os estudos tenham utilizado o modelo PLS-DA para prever o diagnóstico de COVID-19, o potencial do nosso presente estudo (quando comparado com o estudo anterior) se destaca por utilizar dados de exames de rotina dos pacientes, mais acessíveis e baratos, facilitando a aplicação do modelo PLS-DA desenvolvido na prática clínica. Ao contrário do estudo anterior, onde é de difícil implementação, dado o alto custo dos equipamentos de cromatografia líquida de alta eficiência (HPLC) e espectrofotômetro de massa, que são equipamentos muito sofisticados e muito caros^{346,347}. Além da diferença no técnicas utilizadas, outro ponto a ser destacado é que neste estudo também analisamos amostras de pacientes com HIV, TB pulmonar e pacientes co-

infectados HIV/TB, aumentando a complexidade do modelo PLS-DA desenvolvido. Finalmente, identificamos alguns biomarcadores potenciais associados ao diagnóstico de COVID-19 [(cálcio, LDH, hemácias, leucócitos, neutrófilos, basófilos, eosinófilos, hemoglobina e hematócrito) e alguns associados ao diagnóstico de HIV e TB (linfócitos, hemácias, hematócrito, hemoglobina, AST, ALT e glicemia); esses biomarcadores são diferentes do nosso estudo anterior.

Atualmente, a aprendizagem automática e a inteligência artificial têm sido frequentemente utilizadas para auxiliar no diagnóstico precoce de doenças (incluindo COVID-19, VIH e TB) ³⁴⁸⁻³⁵⁰. No entanto, a maioria destes estudos compromete o seu modelo devido ao pequeno tamanho da amostra. Em nosso estudo, os modelos de ML foram treinados utilizando uma amostra relativamente grande (n= 6.498 pacientes), obtendo resultados robustos e precisos. Além disso, até onde sabemos, este é o primeiro estudo em que modelos de *machine learning* foram treinados para a detecção simultânea de COVID-19, HIV, TB e coinfeção HIV/TB.

Alterações nos níveis de cálcio em doenças virais já foram demonstradas em estudos anteriores ^{351,352}. Baixo cálcio está presente na COVID-19, de acordo com a revisão sistemática e meta-análise de Alemzadeh (2021) ³⁵³. Estudos recentes mostraram uma forte correlação entre a positividade da infecção por SARS-Cov-2 e o baixo nível de cálcio no organismo, conforme relatado por Yang (2021) e Cappellini (2020) ^{354,355}. Resultados semelhantes também foram abordados em nosso estudo, onde o cálcio foi considerado o principal biomarcador indicativo do diagnóstico de COVID-19, apresentando valores reduzidos quando comparados aos controles. Estudos *in vitro* e *in vivo* envolvendo animais infectados pelo SARS-Cov-2 mostraram que o gene SARS-Cov-E, gene localizado no aparelho do aparelho de Golgi, é altamente sintetizado durante a infecção viral e é responsável por codificar a proteína transmembrana do canal iônico de cálcio, permitindo a permeabilização do cálcio. Assim, há um desequilíbrio homeostático do cálcio intracelular que pode ativar as vias inflamatórias mediadas por TNF, IL-1b e IL-6, causando danos celulares e edema ^{356,357}. Dada a grande semelhança genômica entre SARS-CoV e SARS- Cov-2, os mecanismos bioquímicos entre eles podem ser semelhantes ³⁵⁸

Em nosso estudo, o modelo PLS-DA identificou a LDH como um potencial biomarcador diagnóstico de COVID-19, o que está de acordo com o que foi relatado em algumas revisões sistemáticas ^{48,49} e em nosso estudo anterior.⁹ Como a LDH está presente nas células pulmonares , possivelmente pacientes infectados pelo

SARS-Cov-2 liberam maior quantidade de LDH no sangue em decorrência das lesões sofridas pelas células pulmonares causadas pelo vírus. 50 Nosso estudo também demonstrou níveis elevados de AST e ALT em HIV/TB pacientes co-infectados, possivelmente porque essas enzimas apresentam níveis elevados durante o processo infeccioso viral ^{359,360}.

A neutropenia em pacientes infectados por SARS-Cov-2, HIV e coinfeção HIV/TB está de acordo com estudos anteriores da literatura ^{361–365}. Uma razão potencial é que os pacientes infectados por COVID-19 apresentaram agranulocitose transitória na doença inicial e pela demanda excessiva de neutrófilos no sangue periférico diante da infecção por SARS-Cov-2 ³⁶². Na infecção por HIV (ou coinfeção HIV/TB), a neutropenia é devida à progressão avançada da doença, que inclui baixos níveis de células CD4+, e altos níveis do vírus HIV-1 RNA que causa citotoxicidade ³⁶⁶. Embora não haja evidências que demonstrem que o HIV infecte e mate diretamente neutrófilos maduros, o HIV demonstrou ter a capacidade de infectar e matar células-tronco hematopoiéticas multipotentes pela Fas- processo dependente de apoptose, e proteínas virais são responsáveis pela supressão da proliferação de células progenitoras granulomonocíticas ^{367–369}.

A anemia é comum em doenças infecciosas graves, como COVID-19, infecção por HIV e TB, e representa uma das principais complicações hematológicas causadas por esses vírus, contribuindo para a redução da taxa de sobrevivência dos pacientes, baixa qualidade de vida e comprometimento do sucesso do tratamento ^{370–372}. Na infecção por SARS-Cov-2, a anemia é atribuída à presença de coagulação intravascular disseminada e microangiopatia trombótica pulmonar, que resulta em hemólise intravascular ⁶² enquanto no HIV causas multifatoriais podem estar relacionadas, como a presença de infecções oportunistas, deficiências nutricionais, alterações no sistema imunológico adaptativo, doenças crônicas pré-existentes e infecção de células estromais pelo HIV ^{373–376}. 65–70 No caso da infecção por M. tuberculosis, o fator nutricional é a principal causa ³⁷⁷. Apesar de nossos resultados consistentes, o desenho transversal adotado neste estudo constituiu a principal limitação do trabalho, pois o desenho deste estudo não permite o acompanhamento do paciente e os biomarcadores sanguíneos podem variar durante a doença.

A principal limitação do estudo é o desequilíbrio entre os pacientes (COVID-19, HIV, TB e grupo controle), o que pode ser justificado pela prevalência e incidência

dessas doenças. Dois outros fatores importantes contribuíram para o desequilíbrio dos dados: (i) os dados do estudo foram recolhidos num hospital da região norte de Moçambique (África) com falta de recursos financeiros e humanos, e esta é uma triste realidade em quase todos os hospitais em Moçambique. Coletamos esses dados manualmente diretamente dos prontuários dos pacientes e não existiam formulários eletrônicos dos pacientes que pudessem facilitar o processo de coleta de dados. Todo esse processo foi realizado pelos pesquisadores do estudo e o estudo não teve financiamento. (ii) O segundo e último ponto que contribuiu para o elevado desequilíbrio nos dados é a elevada estigmatização e discriminação dos pacientes que vivem com VIH/SIDA e tuberculose, e isso impactou na subcontagem de casos e no aumento das taxas de abandono do tratamento, reduzindo drasticamente a quantidade de dados disponíveis para coleta e, conseqüentemente, o elevado desequilíbrio dos grupos.

4.6.3 Estudo III: Predição de diagnóstico de COVID-19 usando dados clínicos de pacientes COVID-19 atendidos na rede de Farmácia Drugstore distribuída em todo território Nacional.

No estudo III, o foco foi desenvolver diversos modelos de inteligência artificial e *machine learning* utilizando mais de 3,3 milhões de pacientes testados para covid-19 (antígeno, IgG e IgM) em diversas farmácias comunitárias privadas distribuídas por todo o Brasil. O preço desses testes variou entre 90-220R\$³⁷⁸. Considerando o tamanho e abrangência da amostra, este é um dos maiores estudos sobre o perfil da doença COVID-19 realizados no mundo. Esses modelos de aprendizagem automática poderiam ajudar os profissionais de saúde a identificar e gerir casos de COVID-19 de forma mais eficiente. Os resultados também podem ajudar a compreender a história natural da doença na perspectiva dos cuidados de saúde primários e dos ambientes farmacêuticos.

É importante destacar que durante o período em que foram coletados os dados deste estudo (2021), quatro variantes importantes da COVID-19 circulavam no Brasil, sendo elas Alfa (B.1.1.7), Beta (B.1.351), Gama (P.1) e Delta (B.1.617.2)^{379,380}. A existência dessas variantes comprova a complexidade dos dados dos pacientes coletados no presente estudo e, conseqüentemente, reflete a robustez do

desempenho preditivo dos modelos de *machine learning* desenvolvidos no presente estudo na previsão do diagnóstico de COVID-19 nos três conjuntos de dados utilizados (antígeno, conjunto de dados IgG e IgM) ^{379,380}.

No Brasil, a principal fonte de atendimento para 75,00% da população é o sistema de saúde do país, o Sistema Único de Saúde (SUS), portanto, durante a pandemia de COVID-19, as farmácias privadas ajudaram o SUS a garantir que os pacientes recebam serviços de saúde adequados ³⁸¹. O conceito de resiliência é utilizado em estudos de sistemas de saúde para avaliar a capacidade de um país para responder a uma pandemia e oferece lições importantes para fortalecer os sistemas de saúde ³⁸¹. As farmácias têm desempenhado um importante papel colaborativo no sentido de fornecer cuidados mais acessíveis e convenientes para milhões de cidadãos no Brasil. Em janeiro de 2021, existiam no Brasil 114.352 farmácias cadastradas, sendo 77,00% farmácias comunitárias privadas. Além disso, existem mais de 114 mil estabelecimentos e mais de 100 mil farmacêuticos atuando em todas as regiões geográficas do Brasil ³⁸². Portanto, o presente estudo corrobora estudos anteriores nos quais os autores destacaram que os farmacêuticos têm sido prestadores de cuidados de saúde essenciais e acessíveis, dispostos a assumir papéis criticamente importantes durante a pandemia de COVID-19 em todo o mundo ³⁸³⁻³⁸⁵. É importante ressaltar que, além de oferecer testes, os farmacêuticos têm prestado outros serviços: verificação de possíveis interações medicamentosas, fornecimento de informações sobre tratamento farmacológico, comunicação da falta de disponibilidade de um medicamento, problemas com receitas (por exemplo, ausência do nome do paciente, falta de legibilidade da prescrição médica) e rejeição de receitas ³⁸⁶. Os farmacêuticos também têm participado na luta contra a desinformação, como a controvérsia sobre o uso de medicamentos sem benefício clinicamente comprovado (por exemplo, hidroxicloroquina e ivermectina) e a negação e hesitação da vacina ^{383,386}.

Em nosso estudo, os modelos de melhor desempenho para previsão de diagnóstico de COVID-19 incluem *Light Gradient Boosting Machine*, *Extreme Gradient Boosting*, *Gradient Boosting Classifier*, *Ada Boost Classifier* e *Logistic Regression*. Esses modelos alcançaram valores de acurácia diagnóstica variando de 0,71 a 0,85. Esses valores de precisão indicam quão bem os modelos podem distinguir entre casos positivos e negativos de COVID-19 usando os dados fornecidos ³⁸⁷. O conjunto de modelos de alto desempenho inclui métodos de conjunto complexos (por exemplo,

Light Gradient Boosting Machine, *Extreme Gradient Boosting* e *Gradient Boosting Classifier*) e modelos lineares mais simples (por exemplo, Regressão Logística). Esta combinação de tipos de modelos levanta questões sobre o compromisso entre a interpretabilidade do modelo e a precisão preditiva. Os resultados sugerem que modelos complexos e mais simples podem alcançar precisão competitiva no diagnóstico de COVID-19, o que é importante para aplicações médicas onde a interpretabilidade e a transparência são frequentemente desejadas ^{388,389}.

Também é relatado o tempo de treinamento dos modelos de *machine learning* para diagnóstico de COVID-19, variando entre 12 e 57 segundos. Esta informação é crucial, uma vez que são preferidos tempos de formação mais curtos para implementação prática, especialmente em ambientes médicos em tempo real ³⁹⁰. A capacidade de treinar esses modelos dentro de um prazo razoável pode impactar sua viabilidade para aplicações clínicas ³⁸⁷.

O diagnóstico preciso da COVID-19 é crucial para o manejo dos pacientes e para a saúde pública ³⁹¹. Os modelos de *machine learning*, quando integrados aos sistemas de saúde, poderiam ajudar os profissionais de saúde a tomar decisões mais informadas em relação ao atendimento, isolamento e tratamento do paciente ³⁴⁸. Os valores de precisão diagnóstica relatados estão na faixa de 71,0% a 85%, o que pode ser considerado moderadamente bom para uma triagem inicial ou como uma ferramenta adicional para o diagnóstico de COVID-19 ³⁹². No entanto, é importante considerar o contexto clínico e os potenciais consequências de falsos positivos e falsos negativos ³⁹².

O estudo também emprega validação cruzada K-Fold (k=10) para avaliar a robustez do desempenho dos modelos no diagnóstico de COVID-19. Os resultados indicam que os valores de precisão dos modelos permanecem consistentes em diferentes dobras dos dados, com variância muito baixa (valores DPR <0,1%). Isto sugere que os modelos não se ajustam excessivamente aos dados de formação e são susceptíveis de generalizar bem para dados novos e não vistos para o rastreamento da COVID-19 ^{393,394}.

O estudo sugere que certas variáveis desempenham um papel crucial na previsão se uma pessoa tem COVID-19. Essas variáveis incluem idade, sintomas como diarreia, tosse, coriza, pressão persistente no peito e febre. O fato de sintomas como tosse, coriza, diarreia e pressão torácica estarem entre as variáveis mais importantes para prever o diagnóstico de COVID-19 está de acordo com o que se

sabe sobre a apresentação clínica da doença ³⁹⁵⁻³⁹⁷. Esses sintomas são consistentes com os sintomas típicos de infecções respiratórias, incluindo COVID-19 ^{398,399}. Esta descoberta ressalta a importância da avaliação clínica e do relato de sintomas no diagnóstico e monitoramento de casos de COVID-19 ⁴⁰⁰.

A inclusão da idade como um preditor significativo é consistente com descobertas anteriores de que os indivíduos mais velhos correm um risco maior de doenças graves e complicações devido à COVID-19 ⁴⁰¹. A idade pode servir como um importante fator de risco na determinação da gravidade da doença e dos resultados potenciais para os pacientes. Os resultados da investigação indicam também que o gênero é a variável menos importante entre as consideradas ⁴⁰². Esta descoberta sugere que, no contexto dos modelos de aprendizagem automática estudados, o gênero pode não ser um forte preditor do diagnóstico de COVID-19. No entanto, é importante notar que isto não diminui o reconhecimento mais amplo de que o gênero pode desempenhar um papel na suscetibilidade a várias doenças, incluindo a COVID-19 ^{403,404}.

Existem muitos fatores independentes que podem impactar significativamente as flutuações no número de pacientes infectados, incluindo o comportamento dos pacientes, medidas de saúde pública e a disseminação de opiniões públicas através das redes sociais ⁴⁰⁵. Por estas razões, uma estrutura ágil e flexível como a fornecida pelas farmácias privadas contribuiu para absorver o impacto causado pela pandemia da COVID-19, ajudando a limitar as necessidades de saúde não satisfeitas por diferentes razões. No Brasil, a principal fonte de atendimento para 75,00% da população é o sistema de saúde do país, o Sistema Único de Saúde (SUS), portanto, durante a pandemia de COVID-19, as farmácias privadas ajudaram o SUS a garantir que os pacientes recebam serviços de saúde adequados ⁴⁰⁶. O conceito de resiliência é utilizado em estudos de sistemas de saúde para avaliar a capacidade de um país de responder a uma pandemia e oferece lições importantes para fortalecer os sistemas de saúde ⁴⁰⁷.

O estudo examinou três métodos de teste diferentes (antígeno, IgG, IgM) para prever o diagnóstico de COVID-19. Ao identificar variáveis importantes comuns entre estes métodos, a investigação implica que certos sintomas e fatores demográficos contribuem consistentemente para um diagnóstico preciso, independentemente da técnica de teste utilizada. Isto poderia fornecer informações sobre a eficácia e confiabilidade de diferentes métodos de teste. Os resultados da pesquisa têm

implicações para a prática clínica e estratégias de saúde pública. Os profissionais de saúde podem considerar os sintomas identificados e os fatores demográficos como pontos de referência importantes ao avaliar pacientes com suspeita de COVID-19. As autoridades de saúde pública também podem utilizar esta informação para refinar estratégias de rastreamento e testes para diagnósticos mais eficazes e precisos.

Como limitação do estudo, é importante observar que esses resultados são baseados no conjunto de dados específico e nos modelos de *machine learning* utilizados no estudo. A eficácia das variáveis identificadas pode variar com base em diferentes populações, regiões e sistemas de saúde. Mais pesquisas, validações e testes são necessários para confirmar a robustez dessas descobertas em diversos conjuntos de dados e ambientes. Outra limitação do estudo é que, embora tenhamos desenvolvido modelos de *machine learning* com grande volume de dados e bom desempenho na previsão do diagnóstico de COVID-19, é importante ressaltar que esses modelos são ferramentas e não devem substituir o julgamento clínico. Por fim, os resultados referem-se apenas ao primeiro semestre de 2021, as variantes do vírus podem ter mudado esse cenário e os dados foram coletados apenas em farmácias privadas – mas até 190 milhões de pessoas dependem exclusivamente do SUS. O efeito destas medidas ainda é muito cedo para avaliar, mas futuros inquéritos serológicos permitirão monitorizar a progressão da pandemia de COVID-19 e ajudar a avaliar a eficácia das mudanças políticas.

4.7 CONCLUSÃO

No quarto capítulo desta tese, modelos de *machine learning* foram rigorosamente treinados e validados com o propósito de prever o diagnóstico do COVID-19, além de investigar potenciais biomarcadores associados ao diagnóstico e prognóstico da doença. Utilizamos dados provenientes de exames bioquímicos, hematológicos e urinários de pacientes com COVID-19 tratados no Hospital Israelita Albert Einstein, no Brasil.

Todos os modelos baseados em *machine learning* (Redes Neurais Artificiais, Árvore de Decisão, PLS-DA e KNN) demonstraram eficácia na predição do diagnóstico de COVID-19 e na avaliação da gravidade da doença, alcançando uma precisão superior a 84%. Esses resultados se assemelham aos obtidos pela técnica

de RT-PCR, estabelecendo assim um patamar mínimo recomendado para testes diagnósticos. O modelo de Redes Neurais Artificiais obteve o melhor desempenho preditivo, atingindo uma precisão entre 94% e 98%, sugerindo sua aplicabilidade como ferramenta de apoio à tomada de decisão por profissionais de saúde.

Adicionalmente, identificamos associações entre o diagnóstico de COVID-19 e a gravidade da doença com biomarcadores como hiperserotoninemia, hipocalcemia, hipoxemia, hipóxia pulmonar, acidose respiratória, acidose metabólica, baixo pH urinário e níveis elevados de lactato desidrogenase. Esses biomarcadores representam potenciais alvos prognósticos que demandam investigação mais aprofundada em futuros ensaios clínicos.

Em outra vertente deste estudo, empregamos modelos de *machine learning* (PLS-DA, ANN, XGBoosted, KNN, LREG, SIMCA e SVM) para analisar um banco de dados contendo pacientes com COVID-19, HIV, TB e co-infectados HIV/TB, provenientes de um hospital de referência na região norte de Moçambique. O modelo PLS-DA demonstrou excelente desempenho no diagnóstico dessas doenças, alcançando taxas de acurácia variando entre 94% e 97%. Além disso, o modelo PLS-DA identificou diversos biomarcadores potenciais associados a cada uma das enfermidades, oferecendo informações valiosas sobre indicadores diagnósticos específicos.

Para a COVID-19, biomarcadores como cálcio, lactato desidrogenase, hemácias, leucócitos, neutrófilos, basófilos, eosinófilos, hemoglobina e hematócrito foram associados. Já a infecção pelo HIV apresentou associações com volume corpuscular médio, plaquetas, neutrófilos e volume médio de plaquetas. No caso da infecção por *Mycobacterium tuberculosis*, a largura de distribuição dos glóbulos vermelhos e a ureia foram identificadas como biomarcadores relevantes. Por fim, a coinfeção HIV/TB foi associada a biomarcadores como linfócitos, glóbulos vermelhos, hematócrito, hemoglobina, aspartato transaminase, alanina transaminase e glicemia. Essas descobertas destacam o potencial do modelo PLS-DA para otimizar o diagnóstico dessas doenças, enfatizando a importância dos biomarcadores específicos para rastreamento e detecção precoce.

Em uma abordagem mais abrangente, treinamos e validamos um total de quinze modelos de *machine learning* para prever o diagnóstico de COVID-19 utilizando dados clínico-epidemiológicos de quatro milhões de pacientes testados em

diversas farmácias comunitárias privadas em todo o Brasil. Dos quinze modelos desenvolvidos para cada conjunto de dados (antígeno, IgG e IgM), os cinco algoritmos mais eficazes na previsão de COVID-19 foram: *Light Gradient Boosting Machine*, *Extreme Gradient Boosting*, *Gradient Boosting Classifier*, *Ada Boost Classifier* e Regressão Logística. Notavelmente, os modelos baseados no conjunto de dados de pacientes testados pelo método IgM demonstraram a mais alta acurácia no diagnóstico de COVID-19 (85%), seguidos pelos modelos baseados nos métodos IgG (76%) e antígeno (71%).

Dentre as variáveis analisadas nos modelos de *machine learning*, aquelas mais relevantes para o diagnóstico da COVID-19 foram: idade, presença de diarreia, tosse, coriza, persistência de pressão no peito, febre e sexo.

Esses resultados não apenas fornecem *insights* cruciais para aprimorar a precisão diagnóstica (COVID-19, HIV e TB) e identificar potenciais alvos prognósticos, mas também delineiam um caminho promissor para o desenvolvimento de abordagens mais eficazes na detecção precoce e no manejo clínico de doenças infecciosas, destacando o potencial transformador da aplicação de modelos de *machine learning* na área da saúde.

5 CAPÍTULO V - EXPLORANDO A METABOLÔMICA INTEGRATIVA POR LC-MS, GC-MS E RMN COM INTELIGÊNCIA ARTIFICIAL NA DESCOBERTA DE NOVOS BIOMARCADORES ASSOCIADOS AO DIAGNOSTICO E PROGNÓSTICO DA COVID-19

Publicado em:

1.Cobre AF, Surek M, Stremel DP, Fachi MM, Lobo Borba HH, Tonin FS, Pontarolo R. Diagnosis and prognosis of COVID-19 employing analysis of patients' plasma and serum via LC-MS and machine learning. *Comput Biol Med.* 2022 Jul;146:105659. doi: 10.1016/j.compbimed.2022.105659.

2.Cobre AF, Alves AC, Gotine ARM, Domingues KZA, Lazo REL, Ferreira LM, Tonin FS, Pontarolo R. Novel COVID-19 biomarkers identified through multi-omics data analysis: N-acetyl-4-O-acetylneuraminic acid, N-acetyl-L-alanine, N-acetyltryptophan, palmitoylcarnitine, and glycerol 1-myristate. *Intern Emerg Med.* 2024 Feb 28. doi: 10.1007/s11739-024-03547-1.

5.1 RESUMO

Neste capítulo V, o objetivo é implementar e avaliar algoritmos de *machine learning* para prever diagnóstico, gravidade e letalidade da COVID-19, investigando biomarcadores associados. Amostras de soro e plasma de pacientes de diversos países foram analisadas por LC-MS, GC-MS e RMN. Amostras de voluntários saudáveis e de pacientes com pneumonia viral não relacionada ao COVID-19 foram utilizadas como controles. Os espectros obtidos dessas análises foram empregados no treinamento de sete modelos de ML: PLS-DA, ANNDA, XGBoostDA, SIMCA, SVM, LREG e KNN. Esses modelos foram utilizados para prever o diagnóstico, gravidade e mortalidade da COVID-19, bem como identificar biomarcadores associados a esses desfechos. O desempenho dos modelos foi avaliado considerando acurácia, sensibilidade e especificidade. Foram analisadas 1602 amostras. O modelo PLS-DA demonstrou o melhor desempenho na predição do diagnóstico, gravidade e letalidade da COVID-19, alcançando acurácias entre 93-97%. Foram identificados 23 biomarcadores associados ao diagnóstico e prognóstico da COVID-19, incluindo espermidina, taurina, L-aspártico, L-glutâmico, L-fenilalanina e xantina, entre outros. Níveis baixos de ribotimidina, 4-hidroxifenilacetoilcarnitina e uridina foram associados à positividade para COVID-19, enquanto níveis elevados de N-acetil-glucosamina-1-fosfato, cisteinilglicina, isobutirato de metila, L-ornitina e 5,6-di-hidro-5-metiluracil foram relacionados a maior gravidade e letalidade. Além disso, foram identificados cinco novos biomarcadores associados à gravidade e diagnóstico da COVID-19, incluindo ácido N-acetil-4-O-acetilneuramínico, N-acetil-L-alanina, N-acetiltriptofano, palmitoilcarnitina e 1-miristato de glicerol, elevados em pacientes graves em comparação com pacientes leves ou voluntários saudáveis. Os resultados deste estudo destacam a eficácia dos modelos de ML na previsão de desfechos relacionados à COVID-19 e identificam uma série de biomarcadores associados ao diagnóstico, gravidade e letalidade da doença. Esses achados contribuem para uma melhor compreensão da fisiopatologia da COVID-19 e podem auxiliar no desenvolvimento de estratégias de diagnóstico e tratamento mais precisas e eficazes.

Palavras-chave: COVID-19, *machine learning*, biomarcadores, diagnóstico, gravidade, letalidade, LC-MS, GC-MS, RMN, *machine learning*.

5.2 INTRODUÇÃO

A pandemia global da COVID-19 desencadeou uma corrida sem precedentes na busca por métodos inovadores e avançados que possam aprimorar o diagnóstico e prognóstico da doença ⁴⁰⁸. Nesse contexto, a Metabolômica Integrativa emerge como uma abordagem promissora, unindo técnicas analíticas de ponta, como LC-MS (Cromatografia Líquida acoplada à Espectrometria de Massas), GC-MS (Cromatografia Gasosa acoplada à Espectrometria de Massas) e RMN (Ressonância Magnética Nuclear), com a inteligência computacional proporcionada pela Inteligência Artificial (IA) ⁴⁰⁹. Esta sinergia entre tecnologias de ponta visa explorar o intrincado mundo dos metabólitos, desvendando nuances específicas da resposta metabólica do organismo à infecção por SARS-CoV-2 ⁴¹⁰.

Este capítulo V da tese de doutorado tem como foco a identificação de novos biomarcadores associados ao diagnóstico e prognóstico da COVID-19, utilizando uma abordagem integrativa que transcende as limitações das técnicas isoladas. A Metabolômica, ao analisar o perfil metabólico global, proporciona uma visão abrangente das mudanças bioquímicas associadas à presença do vírus ⁴¹¹. A combinação com técnicas de IA permite não apenas processar volumes massivos de dados, mas também discernir padrões complexos, essenciais para a descoberta de marcadores discriminatórios ⁴¹².

A utilização de LC-MS, GC-MS e RMN permite uma abordagem complementar, capturando uma gama diversificada de metabólitos que podem ser cruciais na caracterização dos estágios iniciais da infecção, bem como na diferenciação de fenótipos clínicos ^{412,413}. A inteligência artificial, por sua vez, desempenha um papel fundamental na interpretação desses dados complexos, identificando correlações e relações que escapariam à análise manual ⁴¹⁴.

Ao alcançar esses objetivos, este capítulo visa não apenas expandir o repertório de biomarcadores já conhecidos, mas também proporcionar *insights* valiosos para o desenvolvimento de estratégias de diagnóstico precoce e prognóstico mais precisas. A interdisciplinaridade entre a metabolômica e a Inteligência Artificial representa um passo significativo em direção a uma compreensão mais profunda da complexidade metabólica associada à COVID-19, destacando o potencial impacto positivo dessas abordagens integrativas na luta contra a pandemia.

5.3 OBJETIVO

5.3.1 *Objetivo geral:*

- Investigar e explorar a Metabolômica Integrativa por meio das técnicas LC-MS, GC-MS e RMN, aliadas à aplicação de Inteligência Artificial, com o intuito de identificar novos biomarcadores associados ao diagnóstico e prognóstico da COVID-19, contribuindo para uma compreensão mais profunda dos mecanismos metabólicos envolvidos na resposta do organismo à doença

5.3.2 *Objetivos específicos:*

- Coletar dados de metabolômica utilizando LC-MS, GC-MS e RMN para mapear de maneira abrangente os perfis metabólicos de pacientes com COVID-19 em diferentes estágios da doença;
- Aplicar técnicas avançadas de processamento de dados e Inteligência Artificial para analisar e interpretar os resultados obtidos, visando identificar padrões e correlações relevantes para a identificação de biomarcadores;
- Investigar a presença de biomarcadores específicos que possam ser utilizados como indicadores de diagnóstico precoce da COVID-19, utilizando abordagens integrativas de Metabolômica e Inteligência Artificial;
- Avaliar a associação entre os biomarcadores identificados e a progressão da doença, buscando estabelecer relações que possam contribuir para o prognóstico da COVID-19;
- Contribuir para o avanço do conhecimento científico relacionado aos processos metabólicos durante a infecção por COVID-19, fornecendo *insights* que possam orientar futuras intervenções terapêuticas e estratégias de manejo da doença.

5.4 MATERIAL E MÉTODOS

Neste capítulo V da tese, está subdividido em duas subsecções. A primeira secção foca no desenvolvimento de métodos de *machine learning* para predição de diagnóstico e investigação de biomarcadores prognósticos para COVID-19 usando banco de dados públicos de experimentos de metabólica envolvendo pacientes COVID-19.

5.4.1. *Estudo I: Diagnóstico e prognóstico de COVID-19 empregando análise de plasma e soro via LC-MS e Machine learning.*

5.4.1.1 *Fluxograma do estudo e pacientes:*

O fluxograma do estudo empregado para a realização do estudo I deste capítulo V é mostrado na Figura 5.1.

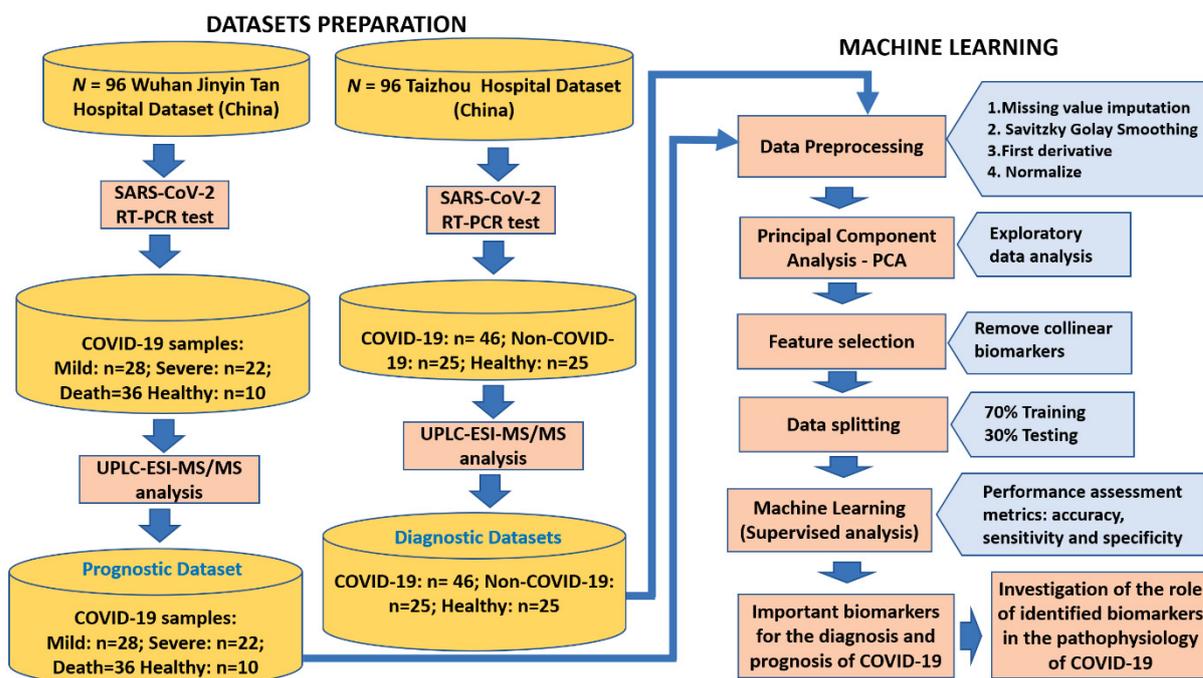


Figura 5.1. Fluxograma do estudo I: diagnóstico e prognóstico de COVID-19 empregando análise de plasma e soro via LC-MS e *Machine learning*. **Fonte:** O Autor (2024)

No estudo I, avaliamos conjuntos de dados públicos de duas coortes de pacientes diagnosticados com COVID-19 por RT-PCR na China (Wuhan), incluindo casos leves e graves e mortes associadas à doença (https://drive.google.com/drive/folders/1R_I_gu5D3SkD_9q_J93HOA9GuKxZiGNG)⁴¹⁵.

Amostras de sangue de todos os pacientes diagnosticados foram avaliadas por cromatografia líquida de ultra-eficiência (UPLC) acoplada à espectrometria de massa (LC-MS). A primeira coorte (Conjunto de dados I) refere-se a amostras de pacientes com COVID-19 diagnosticados no Hospital Wuhan Jinyin Tan (China) (nome do arquivo: C2_metaboanalyst_input_full.csv). Uma série de amostras foi registrada durante o curso da doença: amostras coletadas de 14 pacientes com sintomas leves em dois momentos do estudo (total de $14 \times 2 = 28$ amostras), amostras de 11 pacientes com sintomas graves (total de $11 \times 2 = 22$ amostras) e amostras coletadas em quatro momentos de 9 pacientes que faleceram durante o estudo (total de $9 \times 4 = 36$ amostras). Amostras de sangue de 10 voluntários saudáveis com testes de RT-PCR negativos foram utilizadas como controle. Amostras de plasma de todos esses pacientes também foram coletadas e analisadas por cromatografia líquida de ultra-eficiência (coluna C18) acoplada à espectrometria de massa quadrupolo e fonte de eletrospray (UPLC-ESI-MS/MS). Um total de 431 metabólitos (substâncias solúveis em gordura e solúveis em água) foram identificados e quantificados usando um banco de dados hospitalar interno e perfil de fragmentação de íons moleculares no modo MS/MS. Os fragmentos foram comparados com dados da literatura/banco de dados público internacional. Esta primeira base de dados incluiu, portanto, 96 amostras e 431 variáveis (metabólitos).

A primeira coorte (Conjunto de dados I) refere-se a amostras de pacientes com COVID-19 diagnosticados no Hospital Wuhan Jinyin Tan (China) (nome do arquivo: C2_metaboanalyst_input_full.csv)⁴¹⁵. Uma série de amostras foi registrada durante o curso da doença: amostras coletadas de 14 pacientes com sintomas leves em dois momentos do estudo (total de $14 \times 2 = 28$ amostras), amostras de 11 pacientes com sintomas graves (total de $11 \times 2 = 22$ amostras) e amostras coletadas em quatro momentos de 9 pacientes que faleceram durante o estudo (total de $9 \times 4 = 36$ amostras). Amostras de sangue de 10 voluntários saudáveis com testes de RT-PCR negativos foram utilizadas como controle. Amostras de plasma de todos esses pacientes também foram coletadas e analisadas por cromatografia líquida de ultra-

eficiência (coluna C18) acoplada à espectrometria de massa quadrupolo e fonte de eletrospray (UPLC-ESI-MS/MS). Um total de 431 metabólitos (substâncias solúveis em gordura e solúveis em água) foram identificados e quantificados usando um banco de dados hospitalar interno e perfil de fragmentação de íons moleculares no modo MS/MS. Os fragmentos foram comparados com dados da literatura/banco de dados público internacional. Esta primeira base de dados incluiu, portanto, 96 amostras e 431 variáveis (metabólitos).

A segunda coorte (Dataset II) refere-se a uma amostra de 46 pacientes com diagnóstico de COVID-19 no Hospital Taizhou (China) (nome do arquivo: C3_metaboanalyst_input_full.csv) [19]. Amostras de sangue de 25 voluntários saudáveis (RT-PCR negativo) e de 25 pacientes com síndrome de pneumonia, mas com RT-PCR negativo para SARS-CoV-2, foram utilizadas respectivamente como grupos de controle negativo e positivo. Amostras de soro de todos os pacientes foram analisadas por UPLC-ESI-MS/MS. Este segundo banco de dados foi responsável por 96 amostras e 941 metabólitos (identificados e quantificados usando ambos os modos de ionização; ESI- e ESI+).

5.4.1.2 Pré-processamento de dados para machine learning

O pré-processamento de dados é uma etapa importante para a análise de dados metabolômicos e refere-se à técnica de preparação (ou seja, limpeza e organização) dos dados brutos para torná-los adequados (ou seja, legíveis) para construção e treinamento de modelos baseados em ML ^{267,416}. Neste estudo, ambos os conjuntos de dados da COVID-19 (ou seja, diagnóstico e gravidade da doença) passaram por diferentes métodos de pré-processamento visando selecionar aquele que melhor se ajustava aos dados: (i) Imputação: os dados faltantes foram substituídos pelos valores medianos; (ii) Transformações: valor absoluto, Log10; (iii) Filtragem: linha de base (pontos especificados), linha de base (mínimos quadrados ponderados), derivada (Savitzky – Golay), suavização (Savitzky – Golay), detendência, ponderação de mínimos quadrados generalizados (GLSW), correção de sinal ortogonal (OSC) e externo ortogonalização de parâmetros (EPO); (iv) Normalização: normalização, variação normal padrão (SNV) e correção de dispersão multiplicativa (média MSC); (v) Escalonamento e centralização: escala automática, escala de grupo, escala de decaimento logarítmico, centro médio, centro mediano,

centro multiway, escala multiway e escala média sqrt. Todas as análises foram realizadas no software SOLO (Eigenvector Research).

5.4.1.3 *Desenvolvimento de modelos de machine learning*

Neste estudo, um modelo não supervisionado baseado em ML (análise de componentes principais - PCA) foi inicialmente desenvolvido com ambos os conjuntos de dados com o objetivo de identificar a estrutura dos dados e detectar possíveis amostras anômalas (ou seja, análises exploratórias) ^{269,273}. Para a previsão do diagnóstico de COVID-19 (Banco de Dados II) e da gravidade e letalidade da doença (Banco de Dados I), foram utilizados vários modelos supervisionados baseados em ML: SVM, análise discriminante (LDA), KNN, análise discriminante de redes neurais artificiais (ANNDA), análise discriminante por mínimos quadrados parciais (PLS-DA), SIMCA, análise discriminante de árvore intensificada por gradiente (XGBoostDA) e análise discriminante de regressão logística (LREG). Para a implementação destes modelos de classificação, 70% dos dados foram utilizados para o conjunto de treinamento (calibração) e os 30% restantes para o conjunto de testes.

A seleção da amostra para os conjuntos de treinamento e teste foi realizada aleatoriamente usando o algoritmo Kennard Stone. Para a implementação dos modelos, as classes (grupos) de amostras dos conjuntos de dados foram divididas individualmente em dois novos subconjuntos (ou seja, amostras de treinamento e amostras de teste), sendo este último subconjunto (amostras de teste) utilizado para prever cada classe específica. As amostras do Banco de Dados I (Wuhan, Jinyin Tan Hospital, n = 96 amostras) foram agrupadas da seguinte forma: classe 1 – saudável (representando 10 amostras, das quais 5 foram utilizadas para treinamento do modelo e as 5 restantes para predição de dados); classe 2 – óbito (36 amostras das quais 25 foram utilizadas para treinamento; 11 para predição de dados); classe 3 – COVID-19 grave (22 amostras das quais 15 foram utilizadas para treinamento; 7 para predição de dados); e classe 4 – COVID-19 leve (28 amostras; 20 utilizadas para treinamento e 8 para predição de dados). As amostras da Base de Dados II (Hospital de Taizhou, n = 96) foram categorizadas nas seguintes classes: classe 1 – COVID-19 (representando 46 amostras, das quais 32 foram utilizadas para treinamento dos modelos e as 14 restantes para predição de dados); classe 2 – saudável (25 amostras das quais 15 foram utilizadas para treinamento e 10 para predição de dados); e classe

3 – não COVID-19 (25 amostras; 15 utilizadas para treinamento e 10 para predição de dados).

O método de validação cruzada cega veneziana foi usado para selecionar o número de variáveis latentes (LVs) dos modelos baseados em ML ⁴¹⁷. Foram selecionados LVs com menores valores de erro de validação cruzada, raiz quadrada do erro médio de validação cruzada (RMSECV) e raiz quadrada média do erro quadrático de calibração (RMSEC). A raiz do erro médio de predição (RMSEP) foi a métrica utilizada para avaliar a capacidade preditiva dos modelos baseados em ML; modelos com RMSEP menor tiveram melhor desempenho. O desempenho do modelo foi avaliado considerando as métricas de precisão, sensibilidade e especificidade. Essas métricas foram calculadas utilizando as seguintes figuras de mérito: falso positivo (FP), falso negativo (FN), verdadeiro positivo (VP) e verdadeiro negativo (VN), conforme equações (1)–(3).

$$\text{Sensibilidade (recall)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (1)$$

$$\text{Especificidade} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (2)$$

$$\text{Acurácia} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (3)$$

Onde, FP = falso positivo; FN = falso negativo; VP = verdadeiro positivo; VN = verdadeiro negativo. A precisão dos modelos também foi avaliada considerando a área sob a curva característica de operação do receptor (ROC) (AUC). Os valores de AUC ROC foram calculados considerando os conjuntos de dados de ambas as amostras (isto é, conjuntos de treinamento e teste).

5.4.1.4 Identificação de biomarcadores mais importantes associados ao diagnóstico e severidade de COVID-19

Um gráfico VIP (importância variável na projeção) foi construído a partir do modelo baseado em ML apresentando o melhor desempenho visando identificar os 'top 10' biomarcadores mais importantes para prever o diagnóstico de COVID-19 e os 'top 10' biomarcadores mais importantes para prever a gravidade e a letalidade da doença. Uma pontuação VIP de uma variável original é calculada como uma soma ponderada das correlações quadradas entre o LV do modelo PLS-DA e a variável

original (por exemplo, metabolito). O número de termos na soma depende do número de LV do modelo PLS-DA que foram considerados significativos para distinguir os grupos (classes) de amostras. Os pesos correspondem à variância percentual explicada pelo LV no modelo PLS-DA. Uma variável original com pontuação VIP superior a 1 é considerada estatisticamente significativa para classificação de grupos (por exemplo, grupo COVID-19 vs. indivíduos saudáveis). Veja abaixo a equação de cálculo da pontuação VIP (4) ^{418,419}.

$$VIP^2_j = \sum f w_{if}^2 SSY_f / (SSY_{total\ expl.} F)$$

Onde: w_j = valor do peso PLS; SSY = porcentagem da variância Y explicada por cada variável latente específica; F = número de variáveis latentes do modelo PLS-DA; J = número de variáveis X.

As análises foram realizadas utilizando o software SOLO (Eigenvector Research) e o servidor web Metaboanalyst 5.0 ⁴²⁰; os resultados obtidos com essas diferentes ferramentas foram comparados qualitativamente.

5.4.2 Estudo II: Novos biomarcadores COVID-19 identificados por meio de análise de dados multiômicos: ácido N-acetil-4-O-acetilneuramínico, N-acetil-L-alanina, N-acetiltryptofano, palmitoilcarnitina e 1-miristato de glicerol

5.4.2.1 Fluxograma do estudo:

O fluxograma do estudo empregado para a realização do estudo II deste capítulo V é mostrado na Figura 5.2.

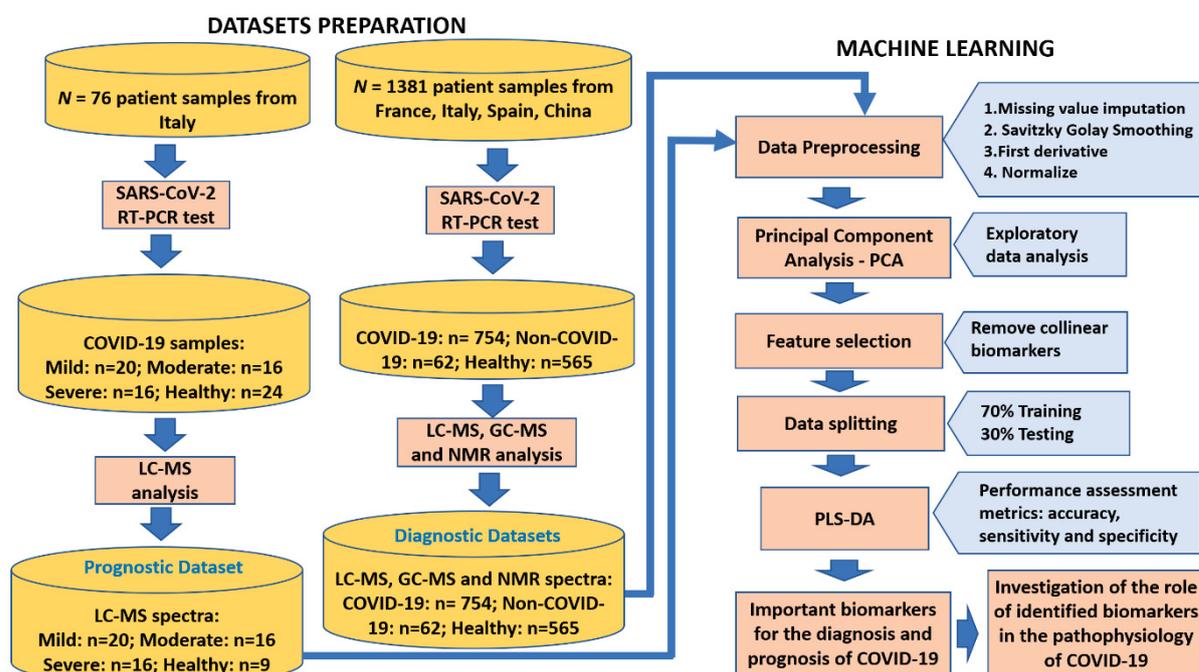


Figura 5.2. Fluxograma do estudo II: Análise de dados metabolômicos de pacientes com COVID-19 de vários países revela novos biomarcadores para diagnóstico precoce e prognóstico da infecção por SARS-CoV-2. **Fonte:** O Autor (2024).

Os dados utilizados para a realização deste estudo foram coletados do banco de dados da plataforma Metaboanalyst ⁴²⁰, que é um banco de dados público que contém dados experimentais de diversas doenças, incluindo a COVID-19.

Os seguintes critérios de inclusão foram adotados para a seleção dos conjuntos de dados: (i) amostras de pacientes com COVID-19 deveriam ser analisadas por ressonância magnética nuclear (RMN), cromatografia líquida acoplada à espectrometria de massa (LC-MS) ou cromatografia gasosa acoplada à espectrometria de massa (GC-MS); (ii) os conjuntos de dados deverão apresentar os metabólitos identificados (caracterizados); (iii) os pacientes devem ser devidamente categorizados de acordo com o resultado do teste COVID-19 (positivo ou negativo) ou de acordo com a gravidade da doença (leve, moderada, grave).

Assim, acabamos de coletar e avaliar sete diferentes conjuntos de dados públicos de metabolômica e lipidômica de uma coorte de pacientes com COVID-19 (diagnosticados por RT-PCR) de cinco países diferentes, Estados Unidos, França, Espanha, Itália e China. Esses dados (metabolômica e lipidômica) foram gerados pela análise do plasma de pacientes em diferentes estágios da doença utilizando três

técnicas analíticas diferentes, RMN, GC-MS e LC-MS. Os links de acesso rápido às bases de dados utilizadas neste estudo podem ser acessados abaixo. O resumo de todos os dados ômicos utilizados para o desenvolvimento deste estudo é apresentado na Tabela 5.1.

Tabela 5.1. Conjuntos de dados multi-ômicos de pacientes COVID-19 e seus respectivos países.

Estudo	Conjunto de dados	País	Método analítico	Tipo de amostra	Condição clínica	Nº Total
Bruzzone, 2020 ⁴²¹	1	Espanha	NMR	Plasma	Saudável	280
			NMR	Plasma	COVID-19	261
Shi, 2021 ⁴²²	2	China	GC-MS	Plasma	Saudável	57
			GC-MS	Plasma	COVID-19	60
			GC-MS	Plasma	Non-COVID-19	30
Albóniga, 2022[23]	3	Espanha	LC-MS	Plasma	Saudável	133
			LC-MS	Plasma	COVID-19	254
Barberis, 2022	4	Itália	GC-MS	Plasma	Saudável	26
			GC-MS	Plasma	Não-COVID	32
			GC-MS	Plasma	COVID-19	103
Blasco, 2020	5	França	LC-MS	Plasma	Saudável	45
			LC-MS	Plasma	COVID-19	55
	6	Itália	LC-MS	Plasma	Saudável	9
Caterino, 2021			LC-MS	Plasma	Leve	20
			LC-MS	Plasma	Moderado	16
			LC-MS	Plasma	Grave	16
Barberis, 2021	7	Itália	GC-MS	Soro	Saudável	24
			GC-MS	Soro	COVID-19	21

Nota: GC-MS: Gas Chromatography-Mass Spectrometry; LC-MS: Liquid Chromatography-Mass Spectrometry, NMR: Nuclear Magnetic Resonance Spectroscopy. **Fonte:** O Autor (2024)

5.4.2.2 Descrição das condições clínicas dos sete conjuntos de dados diferentes

5.4.2.2.1 Conjunto de dados I (conjunto de dados da Espanha): análise de RMN

O conjunto de dados I (conjunto de dados da Espanha) refere-se a amostras de plasma de pacientes com COVID-19 (n = 261) diagnosticados por RT-PCR e de voluntários saudáveis (n = 280) com resultados negativos no teste de RT-PCR para COVID-19 ⁴²¹. Amostras de voluntários saudáveis foram coletadas durante o período de exames de rotina (check-up) ocorridos antes do início da pandemia (2018-2019).

Os dois grupos de amostras foram analisados por ressonância magnética nuclear (RMN). A distribuição por sexo em ambos os grupos de pacientes foi estatisticamente a mesma (116 mulheres no grupo COVID-19 VS 146 mulheres no grupo saudável, $p = 1,00$), por outro lado, a média de idade em ambos os grupos de pacientes foi estatisticamente diferente (COVID - 19 = 65 anos vs saudável = 45 anos, $p < 0,001$). O tempo médio de internação dos pacientes com COVID-19 foi de 14 dias. Quase 9,34% ($n=24$) dos pacientes foram a óbito, e as comorbidades mais frequentes foram: hipertensão ($n = 116, 45,14\%$), eventos cardiovasculares ($n = 68, 26,46\%$) e diabetes ($n = 64, 24,90\%$). Os sintomas mais relatados foram: febre ($n = 177, 68,87\%$), fadiga ($n = 148, 57,59\%$) e tosse seca ($n = 135, 52,53\%$)⁴²¹. Os dados mostrados na tabela 5.1 foram divididos em treinamento (70%) e teste (30%), com vista a desenvolver modelos de ML.

5.4.2.3 Conjunto de dados II (China): análise GC-MS

O conjunto de dados II (conjunto de dados da China) consistiu em amostras de plasma de três grupos de pacientes: pacientes com COVID-19 diagnosticados por teste RT-PCR ($n=57$), grupo de voluntários saudáveis com resultado negativo no teste RT-PCR para COVID-19 ($n=60$) e pacientes não-COVID-19 que apresentaram sintomas semelhantes aos da COVID-19, mas que tiveram resultado negativo no teste RT-PCR para COVID-19⁴²². Os seguintes critérios de inclusão foram adotados para definir um paciente não-COVID-19: (i) redução da contagem de linfócitos ou leucócitos no início da infecção, (ii) febre ou sintomas respiratórios e (iii) manifestações imagiológicas de pneumonia. A idade mediana foi semelhante entre pacientes com COVID-19, não-COVID-19 e saudáveis, 51 anos (IQR, 38-59), 50,5 anos (IQR, 37,5-68,8) e 52 anos (44,3-59), respectivamente. Os homens foram mais prevalentes nos grupos COVID-19 [$n = 47 (59,5\%)$] e voluntários saudáveis [$n = 38 (55,9\%)$], mas foram menos prevalentes no grupo não-COVID-19 [$n = 11 (36,7\%)$]. As comorbidades mais importantes entre pacientes com COVID-19 e não-COVID-19 foram, respectivamente: hipertensão [$n = 19 (24,1\%)$ vs $n = 6 (20\%)$], diabetes [$n = 11 (13,9\%)$ vs $n = 2 (6,7\%)$]. Para o grupo COVID-19, 32 (40,5%) tiveram doença moderada e 47 (59,5%) desenvolveram doença grave⁴²². Os dados mostrados na tabela 5.1 foram divididos em treinamento (70%) e teste (30%), com vista a desenvolver modelos de ML.

5.4.2.4 Conjunto de dados III (conjunto de dados da Espanha): análise LC-MS

O conjunto de dados III (conjunto de dados da Espanha) consistiu em amostras de plasma de pacientes com COVID-19 (n = 254) diagnosticados por RT-PCR e do grupo de voluntários saudáveis (n = 133) com resultado negativo no teste de COVID-19 [23]. RT-PCR. Os pacientes com COVID-19 foram agrupados nas seguintes categorias: (i) COVID-19 grave: pacientes que desenvolveram síndrome respiratória grave; (ii) COVID-19 moderada: pacientes que têm oportunidades de radiografias de tórax e necessitam de oxigenoterapia, (iii) COVID-19 leve: pacientes que estavam assintomáticos uma semana após a infecção por SARS-CoV-2, e que também não necessitou de oxigenoterapia. Todas as amostras foram analisadas por eletroforese capilar acoplada por espectrômetro de massa de tempo de voo (CE-MS), com analisador de eletrospray [23]. A idade mediana do grupo saudável foi menor [42 anos (IQR, 36-51)] do que o grupo COVID-19 [71 anos (55-85)]; Mas dentro do grupo COVID-19, houve uma diferença significativa na idade: COVID-19 leve 57 anos (50-75), COVID-19 moderada 85 anos (75-89), COVID-19 grave 71 anos (59-92). O sexo feminino foi mais predominante entre os pacientes do grupo de voluntários saudáveis, mas para os pacientes com COVID-19 não houve diferença significativa (52%). A hipertensão (52%) foi a comorbidade mais prevalente no grupo COVID-19. Os dados mostrados na tabela 5.1 foram divididos em treinamento (70%) e teste (30%), com vista a desenvolver modelos de ML.

5.4.2.5 Conjunto de dados IV (conjunto de dados da Itália): análise GC-MS

O conjunto de dados VII (conjunto de dados da Itália) pertence à Itália e é formado por pacientes com COVID-19 (n = 103) diagnosticados por RT-PCR, pacientes não-COVID-19 (n = 32) que tiveram outras pneumonias, mas com sintomas semelhantes aos COVID-19, mas resultado negativo de RT-PCR para COVID-19, e grupo de voluntários saudáveis (n=26) ⁴²³. Amostras de plasma dos três grupos de pacientes foram coletadas e analisadas por GC-MS, que identificou 1.108 metabólitos, dos quais 556 pertenciam à classe lipídica (467 identificados pelo modo de ionização positiva e 89 pelo modo de ionização negativa). No grupo COVID-19, 84 dos pacientes apresentavam doença moderada e 19 pacientes apresentavam doença grave. Quanto

ao grupo não-COVID-19, 20 pacientes apresentavam doença moderada e 12 pacientes apresentavam doença grave. As idades médias (em anos) para pacientes com COVID-19, não-COVI-19 e saudáveis foram, respectivamente: $67,3 \pm 18,0$, $69,6 \pm 8,9$ e $50,1 \pm 5,3$. O percentual de mulheres nos grupos COVID-19, não-COVID-19 e saudável foi, respectivamente, 40,7% (n = 42); 59,3% (n = 19) e 57,7% (n = 15). O tempo (em dias) desde a admissão até o diagnóstico foi de $5,8 \pm 7,2$ no grupo COVID-19 e $7,7 \pm 6,5$ no grupo não-COVID-19. O tempo (em dias) desde o diagnóstico até a gravidade foi de $6,5 \pm 7,3$ para pacientes com COVID-19 grave e $1,8 \pm 4,9$ para pacientes sem COVID-19. No grupo COVID-19, as comorbidades mais frequentes foram hipertensão (n = 38), doenças cardiovasculares (n = 38), diabetes (n = 17) e aparelho digestivo (n = 16). Os sintomas mais relatados no grupo COVID-19 foram: febre (n = 52), tosse (n = 34) e dispneia (n = 27) ⁴²³.

5.4.2.6 Conjunto de dados V (França): análise LC-MS

O conjunto de dados V (conjunto de dados da França) é formado por uma coorte de pacientes com COVID-19 (n = 55) diagnosticados por RT-PCR em um hospital na França (Hospital Universitário de Tours) e amostras de voluntários saudáveis (n = 45) ⁴²⁴. Amostras de plasma de ambos os grupos de pacientes foram coletadas e analisadas por LC-MS. A média de idade (em anos) entre o grupo COVID-19 e o grupo saudável foi estatisticamente semelhante ($77,5 \pm 16,0$ vs $75,9 \pm 17,5$, p = 0,83). A distribuição por sexo entre os grupos COVID-19 e saudáveis também foi semelhante (mulheres 51% vs mulheres 49%, p = 1,00). Hipertensão (56,8%), Evento cardiovascular (55,6%), Insuficiência renal (46,7%), Tabagismo (42,3%) foram as comorbidades mais prevalentes no grupo COVID-19. O tempo (em dias) desde a realização do teste RT-PCR até a coleta da amostra de plasma no grupo COVID-19 e saudável foi estatisticamente semelhante ($3,6 \pm 2,6$ vs $2,6 \pm 1$). Os sintomas mais frequentes foram dispneia (64,8%), febre (66,7%), tosse (61,1%) e diarreia (14,8%), que são os sintomas mais frequentes relatados pelo grupo COVID-19. Um total de 66,7% do grupo COVID-19 foram hospitalizados e 11,1% morreram ⁴²⁴. Os dados mostrados na tabela 5.1 foram divididos em treinamento (70%) e teste (30%), com vista a desenvolver modelos de ML.

5.4.2.7 Conjunto de dados VI (conjunto de dados da Itália): análise LC-MS

O conjunto de dados VI (conjunto de dados da Itália) compreende uma coorte de pacientes com COVID-19 ($n = 56$) distribuídos de acordo com a gravidade da doença: COVID-19 leve ($n = 20$), COVID-19 moderado ($n = 16$) e COVID-19 grave ($n = 20$) de dois hospitais na Itália (Hospital Universitário Federico II e Hospital Cotugno, Nápoles-Itália) ⁴²⁵. O banco de dados também continha amostras de voluntários saudáveis que constituíam o grupo controle ($n = 9$). Amostras de plasma de todos os grupos de pacientes foram coletadas e analisadas por LC-MS e foram detectados um total de 630 metabólitos, 483 dos quais pertenciam à classe lipídica. A idade mediana do grupo COVID-19 foi de 58 anos, sendo 70% ($n = 36$) homens e 30% ($n = 16$) mulheres. A idade média do grupo controle era de 46 anos, 40% eram homens e 60% eram mulheres. Dados sobre comorbidades dos pacientes não estavam disponíveis ⁴²⁵. A Tabela 5.1 mostra a divisão do conjunto de dados VI em dois subconjuntos (treinamento e teste) para análise de *machine learning*.

5.4.2.8 Conjunto de dados VII (conjunto de dados da Itália): análise GC-MS

A base de dados IV (conjunto de dados da Itália) consistiu em uma coorte prospectiva formada por dois grupos: (i) pacientes com COVID-19 ($n = 24$) que eram profissionais de saúde em um hospital no norte da Itália (Hospital Universitário de Novara); (ii) grupo de voluntários saudáveis ($n = 21$) [27]⁴²⁶. Amostras de soro desses pacientes foram coletadas e analisadas por GC-MS nas quais foram detectados e quantificados 322 metabólitos, visando investigar quais desses metabólitos são importantes para o diagnóstico. As médias de idade (em anos) entre os grupos COVID-19 e saudáveis foram, respectivamente, $38,9 \pm 10,9$ anos e $36,5 \pm 10,1$. O número de mulheres entre os grupos COVID-19 e saudável foi igual ($n = 16$). O tempo médio entre a coleta das amostras de soro e o diagnóstico foi de $13,3 \pm 5,1$ dias. O tabagismo foi a única comorbidade relatada entre os grupos COVID-19 ($n = 2$) e saudável ($n = 3$). Febre ($n=13$), Anosmia ($n=12$) foram os sintomas mais importantes entre os pacientes com COVID-19. Os dados mostrados na tabela 5.1 foram divididos em treinamento (70%) e teste (30%), com vista a desenvolver modelos de ML⁴²⁶.

5.4.2.9 Pré-processamento de dados para machine learning

O pré-processamento é uma etapa crítica na metabolômica. Este visa remover o ruído experimental contido nos espectros, corrigir os problemas de linha de base dos espectros adquiridos, e outros problemas interferentes que por algum motivo não foi possível eliminá-los durante o processo de obtenção dos espectros. Isto permite converter dados espectrais brutos em espectros limpos e mais ajustados para a realização de análises quimiométricas. O pré-processamento dos espectros LC-MS, GC-MS e RMN foram pré-processados usando a mesma metodologia descrita no estudo I do presente capítulo.

5.4.2.10 *Análise de Componentes Principais – PCA*

Todos os modelos de *machine learning* foram desenvolvidos no software SOLO (Eigenvector Research, Copenhagen). A análise de componentes principais (PCA) foi aplicada para detectar amostras periféricas e avaliar o grau de separação no espaço PC entre casos e controles ⁴²⁷. Para os dados de gravidade, avaliamos o grau de separação entre COVID-19 leve, COVID-19 moderado e COVID-19 grave e indivíduos saudáveis. Para dados diagnósticos, o PCA foi empregado para verificar possível discriminação entre pacientes com e sem COVID-19 e voluntários saudáveis. A detecção de amostras discrepantes foi investigada através da construção do gráfico de alavancagem versus resíduos estudentizados, onde amostras com altos valores de alavancagem e com resíduos estudentizados simultaneamente foram identificadas como potenciais discrepantes e removidas do conjunto de dados ^{6,12}.

5.4.2.11 *Análise Discriminante de Mínimos Quadrados Parciais -PLS-DA*

O primeiro passo no desenvolvimento do modelo PLS DA foi a investigação de amostras discrepantes. A detecção de amostras discrepantes foi investigada através da construção do gráfico de alavancagem versus resíduos estudentizados, onde amostras com altos valores de alavancagem e com resíduos estudentizados simultaneamente foram identificadas como potenciais discrepantes e removidas do conjunto de dados. Os conjuntos de dados de cada um dos quatro países foram divididos aleatoriamente em dois subconjuntos: subconjunto de treinamento (70% das amostras) e subconjunto de teste (30% das amostras). A calibração e validação foram

realizadas automaticamente usando o algoritmo Kennard-Stone. A estimativa do erro de validação cruzada (RMSECV) e da raiz do erro quadrático médio de calibração (RMSEC) foi realizada usando o método de validação cruzada Venetian Blind ⁴¹⁷. Assim, o número de variáveis latentes do modelo PLS-DA foi escolhido considerando valores mais baixos de RMSEC e RMSECV. O desempenho do modelo PLS-DA foi avaliado utilizando sensibilidade (**Equação 1**), especificidade (**Equação 2**) e precisão (**Equação 3**) ²⁷⁸.

5.4.2.12 Identificação de biomarcadores associados ao diagnóstico e prognóstico da COVID-19

VIP foi aplicado aos modelos PLSDA para identificar biomarcadores de doenças relevantes. VIP quantifica a importância das variáveis individuais (biomarcadores) na previsão do modelo. A inspeção do gráfico VIP foi empregada para identificar visualmente variáveis importantes. No modelo PLS-DA, uma pontuação VIP é calculada com base na soma ponderada das correlações quadráticas entre a variável de dados original e as variáveis latentes do modelo PLS-DA. O peso corresponde ao percentual de variâncias explicadas por cada variável latente específica. Variáveis originais com pesos superiores a 1 no gráfico VIP são consideradas estatisticamente significativas por serem importantes para a diferenciação entre as diferentes classes das amostras em estudo ⁴¹⁹

5.5 RESULTADOS

5.5.1. Estudo I: Diagnóstico e prognóstico de COVID-19 empregando análise de plasma e soro via LC-MS e Machine learning

5.5.1.1 Análise exploratória

A Figura 5.3 mostra o modelo PCA para ambos os conjuntos de dados (Dataset I e II). Os métodos de pré-processamento que melhor se adequaram ao modelo foram uma combinação de imputação utilizando valores medianos, auto-escalonamento e GLSW. Em ambos os casos, o modelo PCA foi capaz de discriminar todas as classes de amostras. Nenhuma amostra discrepante foi identificada.

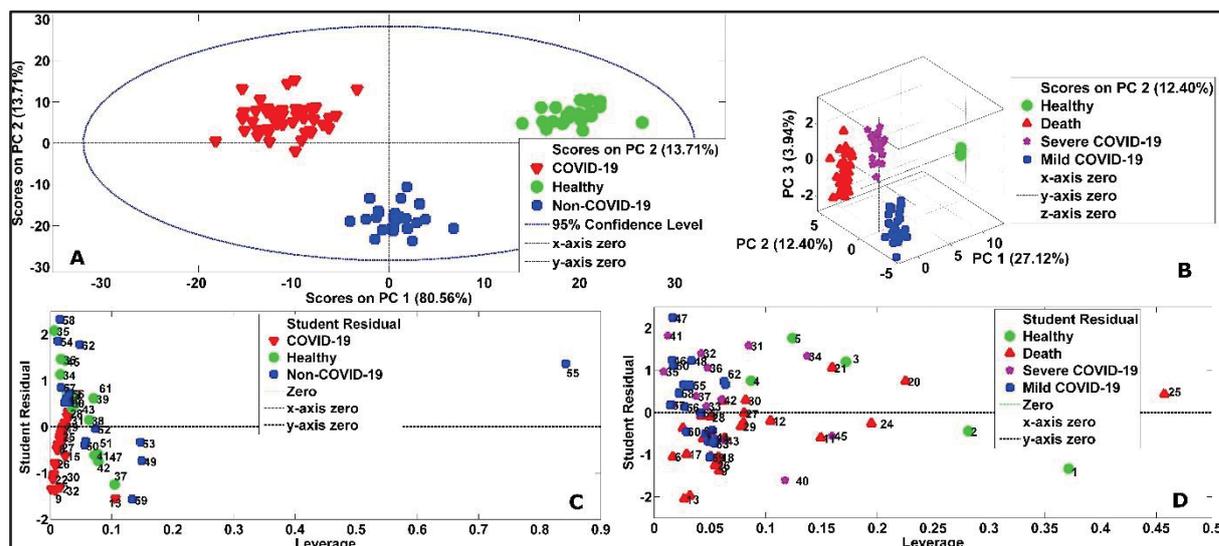


Figura 5.3. Análise exploratória de dados de pacientes do estudo I. (A) Modelo PCA do conjunto de dados de pacientes do hospital de Taizhou (amostras de sangue de 46 pacientes com COVID-19 diagnosticados por RT-PCR são representadas pelos triângulos vermelhos; amostras de sangue de 25 pacientes com síndrome de pneumonia, mas com RT-PCR negativo para SARS-CoV-2 são representados como quadrados azuis; amostras de sangue de 25 voluntários saudáveis com RT-PCR negativo são representadas por círculos verdes). (B) Modelo PCA do conjunto de dados de pacientes do hospital de Wuhan (amostras de sangue de 28 pacientes com COVID-19 leve são representadas pelos quadrados azuis; amostras de sangue de 36 pacientes com COVID-19 são representadas por estrelas rosa; amostras de sangue de 36 mortes por COVID-19 são representadas como triângulos vermelhos; amostras de sangue de 10 voluntários saudáveis negativas por RT-PCR são representadas como círculos verdes). (C) Gráfico de alavancagem versus resíduos de estudantes para a detecção de amostras discrepantes no conjunto de dados de pacientes do hospital de Taizhou (a amostra n. 55 apresentou valores de alavancagem altos, mas

não foi considerada uma discrepância, pois está dentro de $\pm 2,5$ desvios padrão dos resíduos de estudantes). (D) Gráfico de alavancagem versus resíduos de estudantes para a detecção de amostras discrepantes no conjunto de dados de pacientes do hospital de Wuhan (a amostra n. 25 tinha valores de alavancagem altos, mas não foi considerada uma discrepância, pois está dentro de $\pm 2,5$ desvios padrão dos resíduos de estudantes). **Fonte:** O Autor (2024).

5.5.1.2 Machine learning: modelos de classificação

A Tabela 5.2 mostra o desempenho dos modelos baseados em ML construídos com o software SOLO. O modelo PLS-DA apresentou os resultados mais promissores (alto desempenho) para prever o diagnóstico, gravidade e letalidade da COVID-19 com valores mais elevados de precisão, sensibilidade e especificidade de 93%, 94% e 97%, respectivamente. As curvas ROC do modelo PLS-DA usando amostras dos conjuntos de treinamento e teste na **Figura 5.4**. Os demais modelos baseados em ML (ANN, ANNDA, XGBoostDA, SIMCA, SVM, LREG e KNN) apresentaram baixo desempenho preditivo (Tabela 5.2).

Tabela 5.2. Desempenho preditivo dos modelos de *machine learning*

Conjunto de dados do Hospital Wuhan, China (Banco de Dados I)								
Classe	Modelo	VP	FN	VN	FP	Sensibilidade	Especificidade	Acurácia
COVID-19 leve	ANN	15	13	49	19	0,54	0,72	0,67
	KNN	20	8	53	15	0,71	0,78	0,76
	SVM	23	5	45	23	0,82	0,66	0,71
	PLS-DA	28	0	63	5	1,00	0,93	0,95
	SIMCA	17	11	39	29	0,61	0,57	0,58
	XGboost	25	3	60	8	0,89	0,88	0,89
	LREG	18	10	56	12	0,64	0,82	0,77
COVID-19 grave	ANN	16	6	65	9	0,73	0,88	0,84
	KNN	15	7	57	17	0,68	0,77	0,75
	SVM	18	4	60	14	0,82	0,81	0,81
	PLS-DA	19	3	71	3	0,86	0,96	0,94
	SIMCA	15	7	41	33	0,68	0,55	0,58
	XGboost	17	5	59	15	0,77	0,80	0,79
	LREG	16	6	66	8	0,73	0,89	0,85
Óbito por COVID-19	ANN	29	7	50	10	0,81	0,83	0,82
	KNN	24	12	53	7	0,67	0,88	0,80
	SVM	31	5	55	5	0,86	0,92	0,90
	PLS-DA	35	1	58	2	0,97	0,97	0,97
	SIMCA	28	8	43	17	0,78	0,72	0,74
	XGboost	33	3	51	9	0,92	0,85	0,88
	LREG	25	9	46	14	0,69	0,77	0,74
Saudável (controle)	ANN	10	0	74	12	1,00	0,86	0,88
	KNN	7	3	67	19	0,70	0,78	0,77
	SVM	9	1	70	16	0,90	0,81	0,82
	PLS-DA	10	0	82	4	1,00	0,95	0,96
	SIMCA	10	0	73	13	1,00	0,85	0,86

	XGboost	10	0	76	10	1,00	0,88	0,90
	LREG	10	0	61	19	1,00	0,71	0,74
Conjunto de dados do Hospital Taizhou, China (banco de dados II)								
Classe	Modelo	VP	FN	VN	FP	Sensibilidade	Especificidade	Acurácia
COVID-19 positivo	ANN	40	6	39	11	0,87	0,78	0,82
	KNN	33	13	40	10	0,72	0,80	0,76
	SVM	41	5	37	13	0,89	0,74	0,81
	PLS-DA	45	1	44	6	0,98	0,88	0,93
	SIMCA	34	12	42	8	0,74	0,84	0,79
	XGboost	39	7	41	9	0,85	0,82	0,83
	LREG	41	5	33	17	0,89	0,66	0,77
Não- COVID-19	ANN	20	5	59	12	0,80	0,83	0,82
	KNN	14	11	55	16	0,56	0,77	0,72
	SVM	22	3	44	17	0,88	0,62	0,69
	PLS-DA	23	2	64	7	0,92	0,90	0,91
	SIMCA	19	6	42	19	0,76	0,59	0,64
	XGboost	15	10	62	9	0,60	0,87	0,80
	LREG	17	8	49	12	0,68	0,69	0,69
Saudável (controle)	ANN	16	9	56	15	0,64	0,79	0,75
	KNN	18	7	63	8	0,72	0,89	0,84
	SVM	16	9	51	20	0,64	0,72	0,70
	PLS-DA	23	2	67	4	0,92	0,94	0,94
	SIMCA	19	6	60	11	0,76	0,85	0,82
	XGboost	21	4	58	13	0,84	0,82	0,82
	LREG	17	8	55	16	0,68	0,77	0,75

Nota: PLS-DA: partial least squares discriminant analysis (PLS-DA); ANN: artificial neural network; XGBoosted: eXtreme Gradient Boosting, KNN: K-Nearest Neighbors, LREG: Regressão logística; SIMCA: Soft independent modeling by class analogy, SVM: Support vector machine; TP: verdadeiro positivo, TN: verdadeiro negativo; FP: falso positivo e FN: falso negativo. **Fonte:** O Autor (2024).

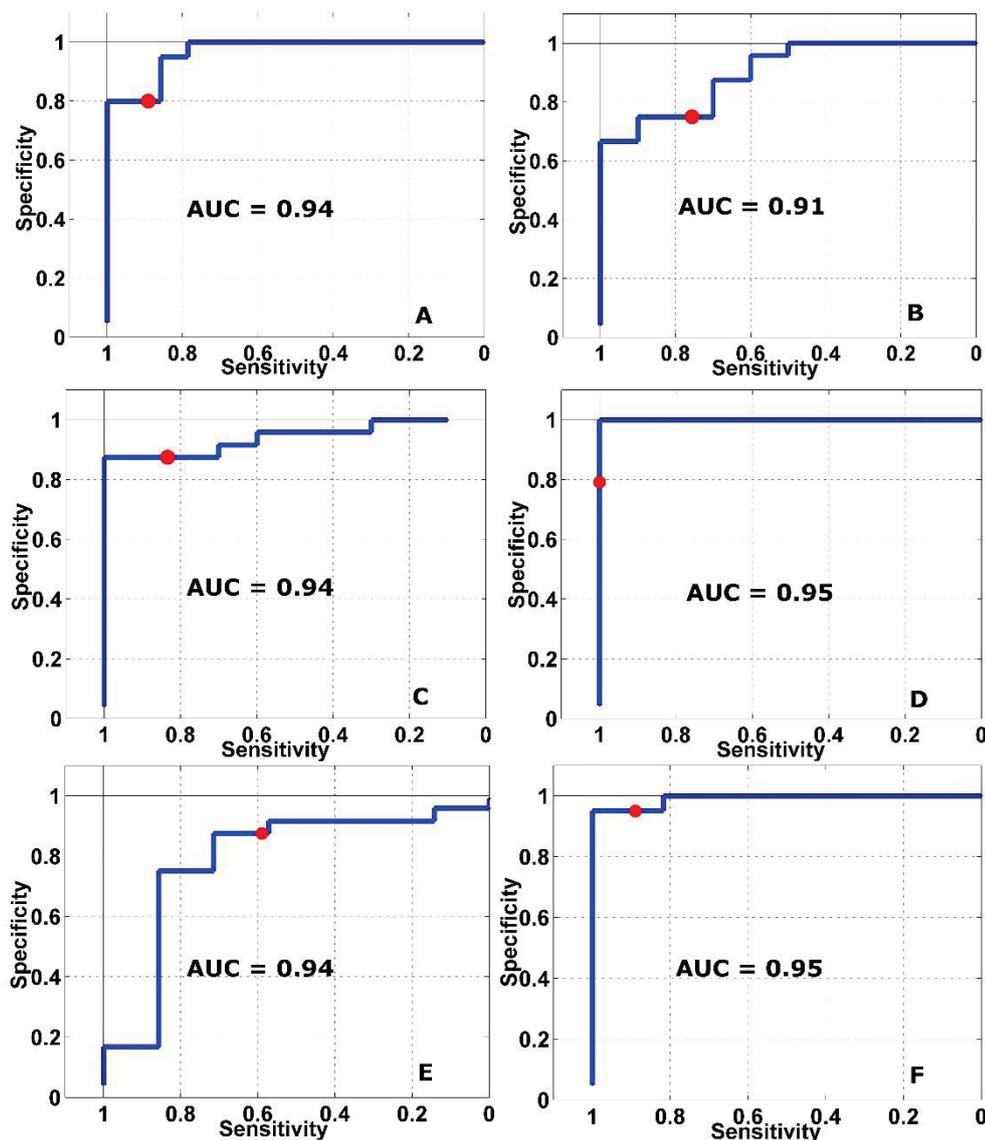


Figura 5.4. Área sob a curva ROC do desempenho do modelo PLS-DA para a predição do diagnóstico e severidade de COVID-19 usando dados de pacientes do estudo I. A área sob as curvas (AUC) reflete a precisão dos modelos PLS-DA na previsão de pacientes de diferentes classes de COVID-19 e voluntários saudáveis. As curvas incluem ambos os conjuntos de amostras (amostras de treinamento e de teste). (A) Conjunto de dados do hospital Thaizhou (conjunto de dados II): os resultados representam a precisão na previsão da classe de pacientes com COVID-19 diagnosticados por RT-PCR (AUC=0,93). (B) Conjunto de dados hospitalares de Thaizhou (conjunto de dados II): os resultados representam a precisão na previsão da classe de pacientes com síndrome de pneumonia, mas com RT-PCR negativo para SARS-CoV-2 (AUC=0,91). (C) Conjunto de dados do hospital Thaizhou (conjunto de dados II): os resultados representam a precisão na previsão da classe de voluntários saudáveis com RT-PCR negativo (AUC=0,94). (D) Conjunto de dados do hospital de Wuhan (conjunto de dados I): os resultados representam a precisão na previsão da classe de pacientes com COVID-19 agudo (AUC=0,95). (E) Conjunto de dados do hospital de Wuhan (conjunto de dados I): os resultados representam a precisão na

previsão de pacientes com COVID-19 grave (AUC=0,94). (F) Conjunto de dados do hospital de Wuhan (conjunto de dados I): os resultados representam a precisão na classificação dos óbitos por COVID-19 (AUC=0,97). **Fonte:** O Autor (2024).

5.5.1.3 Identificação dos biomarcadores importantes para a predição do diagnóstico e prognóstico da COVID-19

As Figura 5.5 e Figura 5.6 (gráficos VIP dos modelos PLS-DA) retratam os biomarcadores mais promissores para prever o diagnóstico e a gravidade/letalidade da COVID-19, respectivamente. O cálculo dos escores VIP desses biomarcadores (ver equação 4 - seção de materiais e métodos) incluiu dois parâmetros: (i) quatro LVs selecionados para o modelo PLS-DA por apresentarem menor erro de calibração (RMSEC) e validação cruzada (RMSECV), e (ii) e a variância total explicada por esses quatro LVs selecionados, que foi de 42,84% para o bloco X e 83,53% para o bloco Y (variável dependente, as classes das amostras).

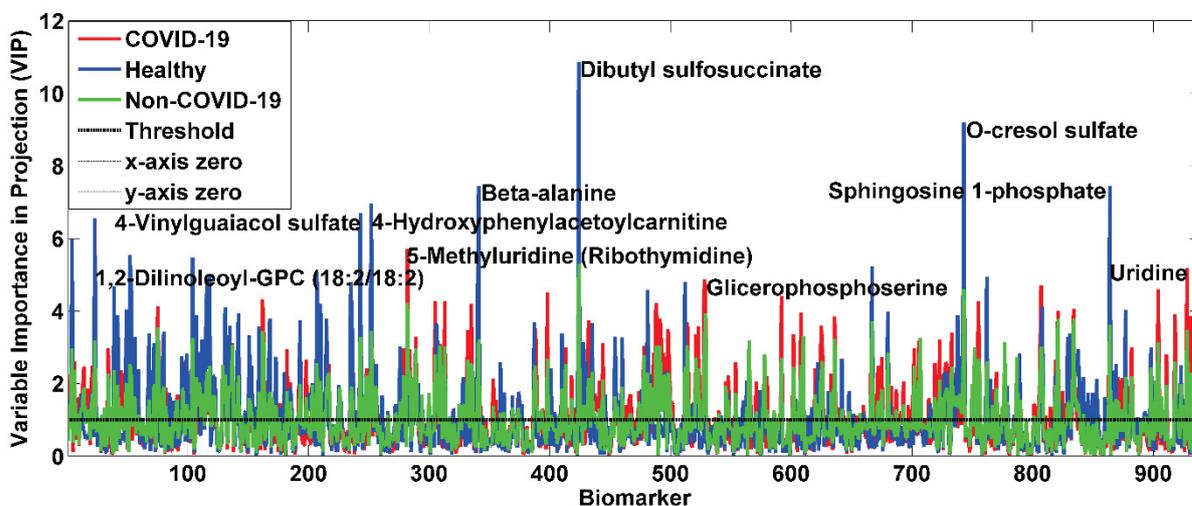


Figura 5.5. Gráfico de importância variável na projeção dos biomarcadores mais importantes para diagnóstico de COVID-19 (10 principais) do estudo I. O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na predição das diferentes classes das amostras (COVID-19 representado pela cor vermelha, não-COVID-19 pela cor azul e voluntários saudáveis pela cor verde). A linha preta tracejada paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP =1). Os metabólitos que contribuem significativamente para a predição das diferentes classes das amostras estão acima do limite (pontuação VIP > 1); os 10 principais biomarcadores foram destacados na figura. **Fonte:** O Autor (2024).

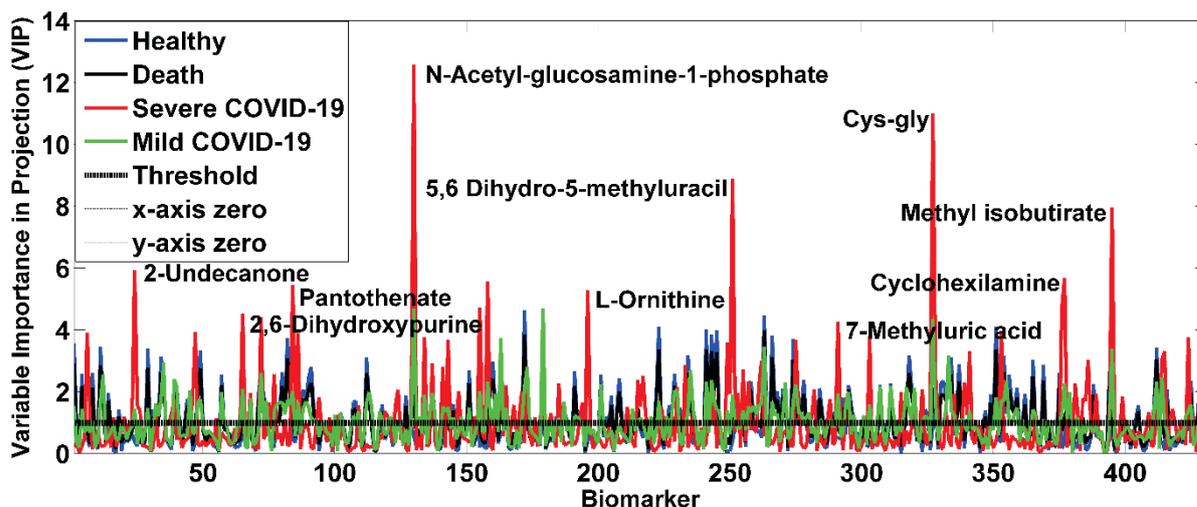


Figura 5.6. Gráfico de importância variável na projeção dos biomarcadores mais importantes para gravidade e letalidade da COVID-19 do estudo I. O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na previsão das diferentes classes das amostras (indivíduos saudáveis são representados em azul, COVID-19 leve é verde, COVID-19 grave está em vermelho e a morte é colorida em preto). A linha preta tracejada paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP = 1). Os metabólitos que contribuem significativamente para a previsão das diferentes classes das amostras estão acima do limite (pontuação VIP > 1); os 10 principais biomarcadores foram destacados na figura. **Fonte:** O Autor (2024).

O gráfico VIP do modelo PLS-DA revelou que os biomarcadores mais importantes para prever o diagnóstico de COVID-19 foram dibutil sulfosuccinato, sulfato de ortocresol, beta alanina, sulfato de 4-vinilguaiacol, 4-hidroxifenilacetoilcarnitina, ribotimidina, glicerofosfoserina e uridina (Figura 5.7). Esses três últimos biomarcadores foram encontrados em concentrações extremamente baixas (diminuídas por fatores de dois, três e quatro) em pacientes com diagnóstico de COVID-19 quando comparados com amostras de controle negativo (Figura 5.7).

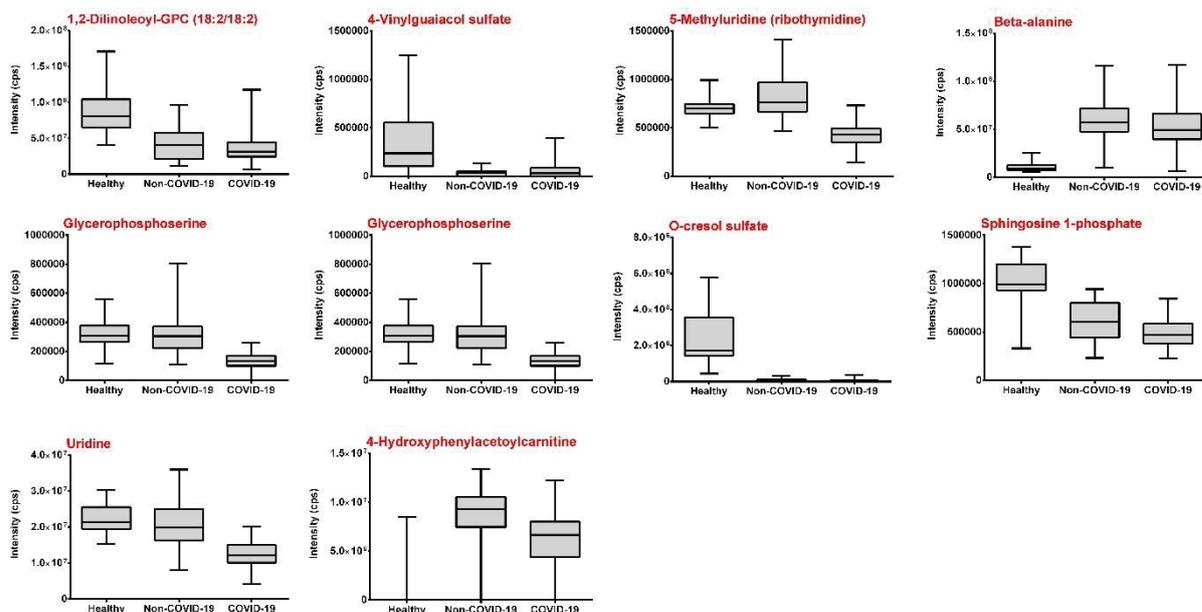


Figura 5.7. Perfil dos 10 principais biomarcadores sanguíneos associados ao diagnóstico de COVID-19. Os resultados estão agrupados de acordo com as classes: saudável (n=25), não-COVID-19 (n=25) e COVID-19 (n=46). As caixas indicam os intervalos interquartis (mediana); linhas horizontais indicam valores mínimos e máximos. **Fonte:** O Autor (2024).

Quanto à previsão da gravidade e letalidade da COVID-19, seis biomarcadores diferentes foram destacados no modelo como provavelmente associados a estes desfechos: ciclohexilamina, isobutirato de metila, 2-undecanona, cisteinilglicina, N-acetil-glucosamina-1-fosfato e 5 O ,6-di-hidro-5-metiluracil aumentou por fatores de três, quatro, cinco, seis, sete e sete, respectivamente, em pacientes com COVID-19 grave quando comparados com aqueles com doença aguda (Figura 5.8).

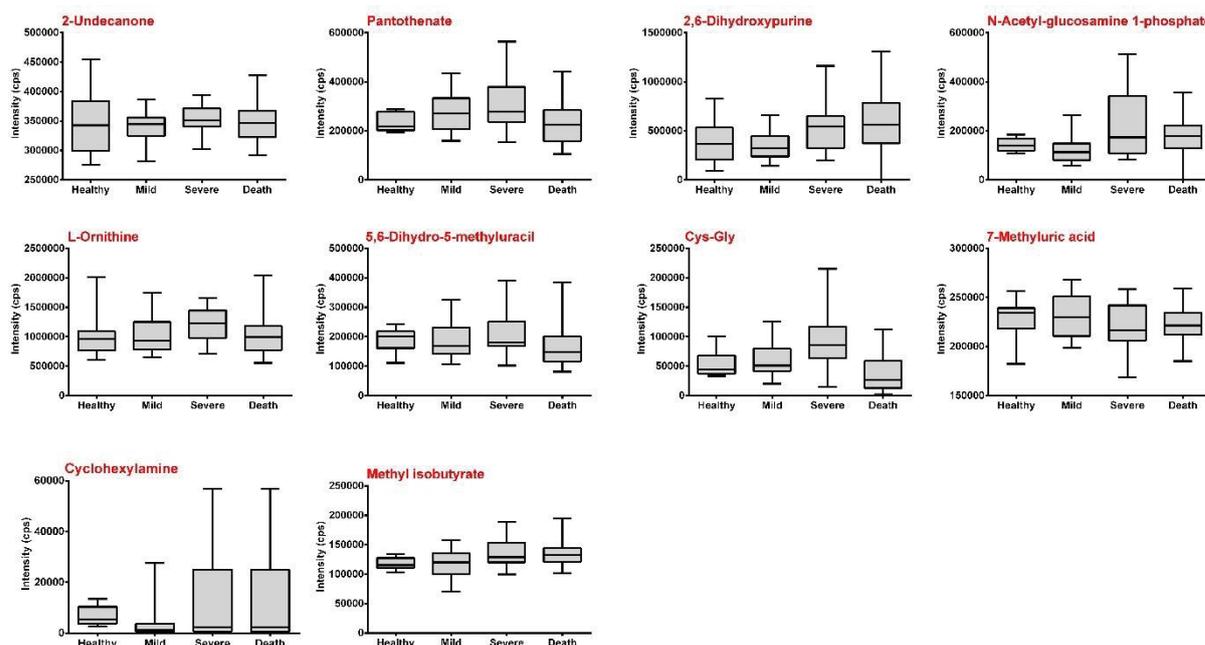


Figura 5.8. Perfil dos 10 principais biomarcadores sanguíneos associados à gravidade e letalidade da COVID-19. Os resultados estão agrupados de acordo com as classes: saudável (n=10), COVID-19 leve (n=28), COVID-19 grave (n=22) e óbito (n=36). As caixas indicam intervalos interquartil (mediana); linhas horizontais indicam valores mínimos e máximos. **Fonte:** O Autor (2024).

As análises acima mencionadas também foram realizadas (ou seja, reexecutadas) usando *Metaboanalyst 5.0* como pode ser observado nas figuras 5.9 e 5.10. Neste caso, o modelo PCA não foi capaz de distinguir as amostras das três classes (COVID-19, não-COVID-19, indivíduos saudáveis) utilizando os dados diagnósticos; a acurácia foi inferior a 80% (ou seja, inferior à obtida em nosso estudo [93-94%] usando o software *SOLO*). Da mesma forma, embora o modelo PCA dos dados de gravidade tenha sido capaz de distinguir pacientes saudáveis de mortes, não foi capaz de classificar os outros dois grupos (COVID-19 leve vs. COVID-19 grave). A precisão deste modelo ficou em torno de 80%, inferior à obtida com *SOLO* (94%-97%).

Também encontrou-se diferenças na identificação de biomarcadores dos modelos PLS-DA obtidos usando o software *SOLO* vs. *Metaboanalyst 5.0* (Tabela 5.3). As análises realizadas no *Metaboanalyst 5.0* mostraram os seguintes metabólitos em níveis extremamente baixos em pacientes com COVID-19 em comparação com aqueles sem a doença ou voluntários saudáveis: linoleato, palmitato, uréia, lactato, carnitina, prolina, glicerofosfoetanolamina, estearato, fenilalanina (ver Figura 5.9). Por outro lado, os metabólitos 6-metilmercaptopurina, ácido di-hidroxibenzenoacético, L-

fenilalanina, 6-metilmercaptapurina, ácido 4-di-hidroxibenzenoacético, L-fenilalanina, ácido formilantranílico, ácido tereftálico e ácido ftálico foram encontrados em altas concentrações em pacientes que morreram de COVID -19 (Figura 5.10).

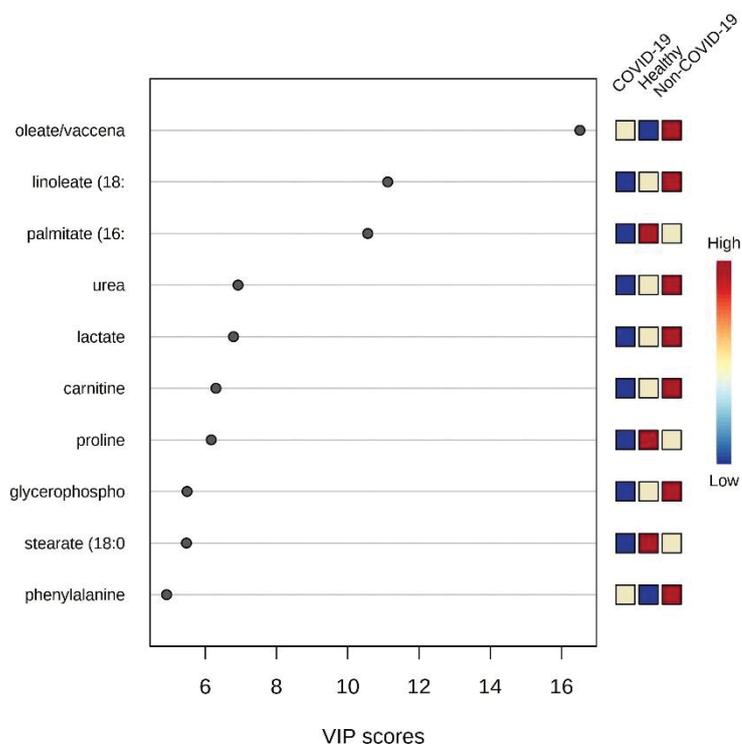


Figura 5.9. Gráfico de Importância Variável na Projeção (VIP) dos biomarcadores mais importantes para diagnóstico de COVID-19 (servidor web *Metaboanalyst 5.0*). O eixo Y representa os 10 metabólitos mais importantes na previsão do diagnóstico de COVID-19 e o eixo X representa a pontuação VIP que reflete a importância de cada metabólito na previsão das diferentes classes das amostras (COVID-19, não-COVID-19 e voluntários saudáveis). A mudança da cor azul para vermelha é proporcional ao aumento da intensidade do sinal do biomarcador. **Fonte:** O Autor (2024).

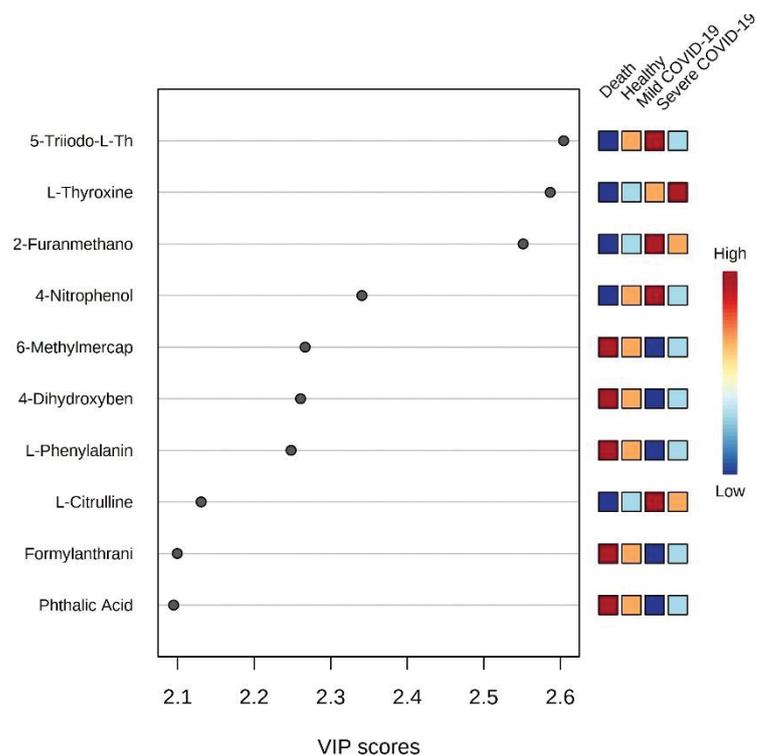


Figura 5.10. Gráfico de Importância Variável na Projeção (VIP) dos biomarcadores mais importantes para gravidade/letalidade da COVID-19 (servidor web *Metaboanalyst 5.0*). O eixo Y representa os 10 metabólitos mais importantes na previsão da gravidade do COVID-19 e o eixo X representa a pontuação VIP que reflete a importância de cada metabólito na previsão das diferentes classes das amostras (morte, COVID-19 grave, COVID-19 leve, indivíduos saudáveis). A mudança da cor azul para vermelha é proporcional ao aumento da intensidade do sinal do biomarcador. **Fonte:** O Autor (2024).

Tabela 5.3. Biomarcadores para predição do diagnóstico e gravidade/severidade da COVID-19 de acordo com os modelos PLS-DA do software SOLO vs. e Metaboanalyst 5.0

Rank	Diagnóstico de COVID-19 (Dataset II - Hospital de Thaizou)		Gravidade/mortalidade da COVID-19 (Dataset I - Hospital de Wuhan)	
	SOLO	Metaboanalyst	SOLO	Metaboanalyst
1	Sulfosuccinato de dibutila	Oleato	Fosfato de N-acetil-glucosamina-1	5-Triiodo-L-tironina
2	Sulfato de O-cresol	Linoleato	Cis-glicina	L-tiroxina
3	Beta-alanina	Palmitato	5,6Diidro-5-metiluracil	2-Furanometanol
4	Esfingosina 1-fosfato	Uréia	Isobutirato de metila	4-Nitrofenol
5	4-Vinylguaiacol sulfate	Lactato	2-Undecanona	6-Metilmercaptapurina
6	4-Hidroxfenilacetoilcarnitina	Carnitina	Pantotenato	Ácido 4-dihidroxibenzenoacético
7	1,2 Dilinoleoil-GPC (18:2/18:2)	Rolina	L-Ornitina	L-fenilalanina
8	5-Metiluridina	Glicerofosfoetanolamina	2,6 Dihidroxipurina	L-citrulina
9	Glicerofosfoserina	Estearato	Ciclohexilamina	Ácido formilantranílico
10	Uridina	fenilalanina	ácido 7-metilúrico	Ácido ftálico

Fonte: O Autor (2024)

5.5.2 Estudo II: Novos biomarcadores COVID-19 identificados por meio de análise de dados multi-ômicos: ácido N-acetil-4-O-acetilneuramínico, N-acetil-L-alanina, N-acetiltryptofano, palmitoilcarnitina e 1-miristato de glicerol

5.5.2.1 Análise exploratória

No estudo II, a visualização do gráfico de score do modelo PCA não identificou amostrou outliers. (Figuras 5.11 e 5.12) e o modelo PCA foi capaz de discriminar entre as amostras COVID19 e controles saudáveis, bem como entre amostras COVID-19, não-COVID-19 e saudáveis.

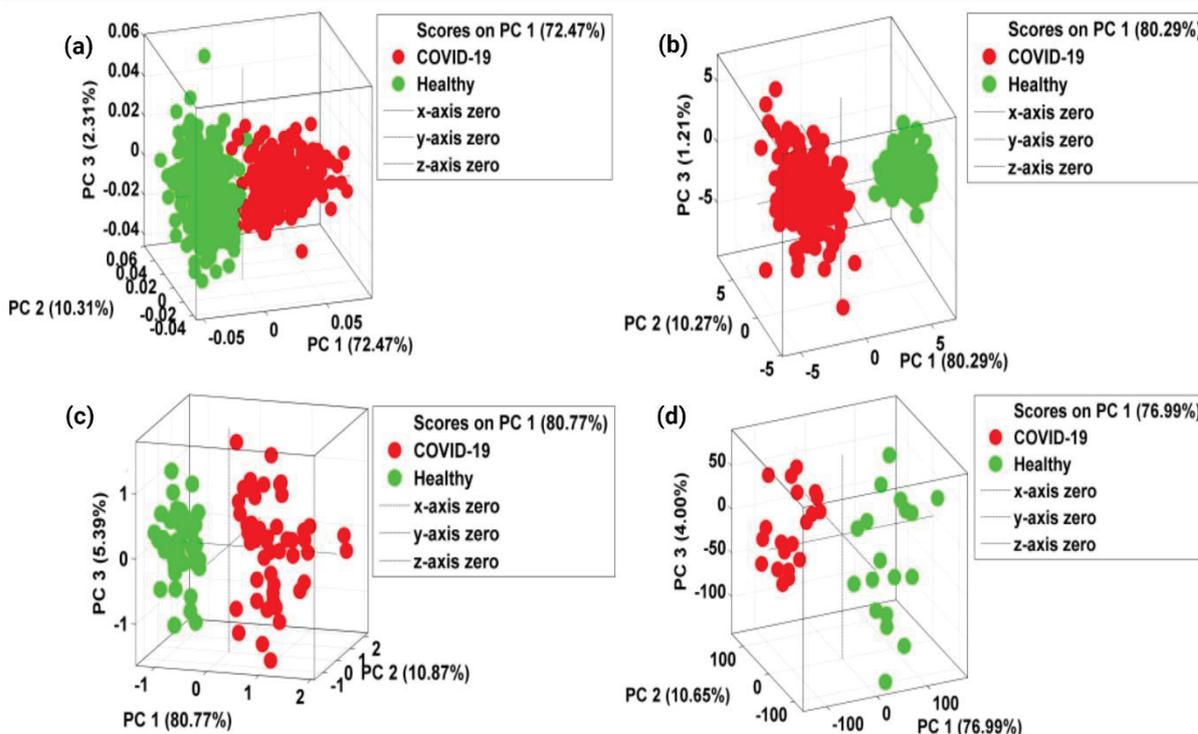


Figura 5.11. Modelo PCA para discriminação de amostras de pacientes com COVID-19 (círculos vermelhos) e voluntários saudáveis (círculos verdes) do estudo II. Em (a) é ilustrado um modelo PCA referente às amostras de pacientes com COVID-19 ($n = 261$) e voluntários saudáveis ($n = 280$) da Espanha analisadas por RMN (conjunto de dados 1). Em (b) é ilustrado um modelo PCA referente a amostras de pacientes com COVID-19 ($n = 254$) e voluntários saudáveis ($n = 133$) dos Estados Unidos analisadas por LC-MS (conjunto de dados 3). Em (c) é ilustrado um modelo PCA referindo-se a amostras de pacientes com COVID-19 ($n = 55$) e voluntários saudáveis ($n = 45$) da França analisadas por LC-MS (conjunto de dados 5). Em (d) é ilustrado um modelo PCA referente a amostras de pacientes com COVID-19 ($n = 21$) e voluntários saudáveis ($n = 24$) da Itália analisadas por GC-MS (conjunto de dados 7). Em todos os conjuntos de dados, o modelo PCA foi capaz de discriminar entre amostras de COVID-19 e amostras de voluntários saudáveis. **Fonte:** O Autor (2024).

Quanto aos dados de gravidade, os pacientes com COVID-19 leves, moderado e grave foram claramente separados (figura 5.12).

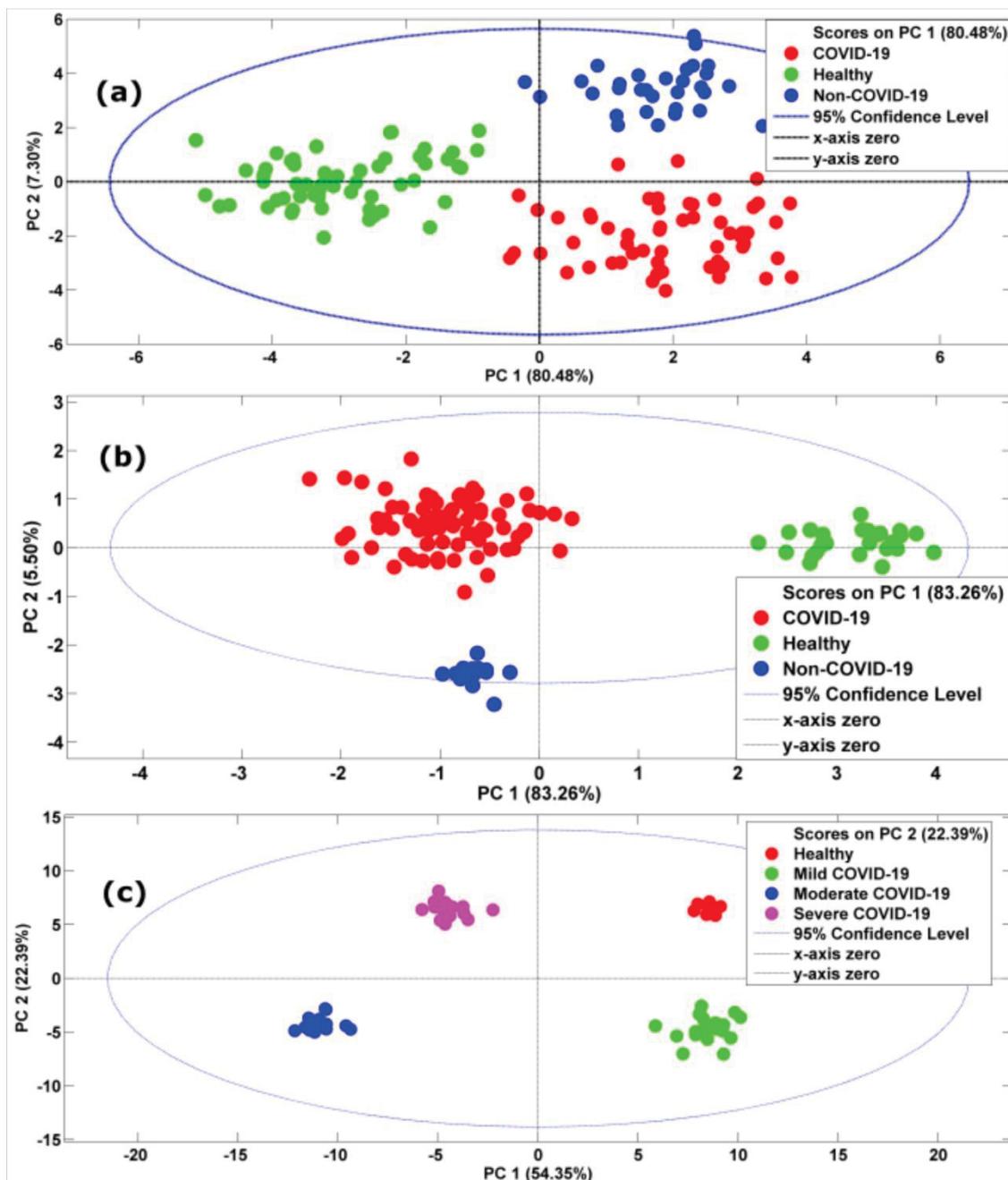


Figura 5.12. Modelo PCA para discriminação de amostras de dados de diagnóstico e gravidade da COVID-19, do estudo II. Para dados de diagnóstico, amostras de pacientes com COVID-19, voluntários saudáveis e não-COVID-19 são representadas por círculos antigos, verdes e azuis, respectivamente. Em (a), um modelo PCA é ilustrado referindo-se a amostras de pacientes com COVID-19 ($n = 60$), voluntários saudáveis ($n = 57$) e não-COVID-19 ($n = 30$) da Espanha analisadas por GC-MS (conjunto de dados 2). Em (b) é ilustrado um modelo PCA referente às amostras de pacientes com COVID-19 ($n = 103$), voluntários saudáveis ($n = 26$) e não-COVID-19 ($n = 40$) da Itália analisadas por GC-MS (conjunto de dados 4). Para os dados de gravidade, em (c) é ilustrado um modelo PCA referente a amostras de pacientes com COVID-19 grave ($n = 16$), COVID-19 moderada ($n = 16$), COVID-19 leve ($n = 20$) e

voluntários indivíduos saudáveis ($n = 9$) da Itália analisados por LC-MS (conjunto de dados 6). Amostras de pacientes saudáveis, COVID-19 grave, COVID-19 moderada, COVID-19 leve e voluntários saudáveis são representadas por círculos rosa, azul, verde e vermelho, respectivamente (conjunto de dados 6). Em todos os conjuntos de dados, o modelo PCA foi capaz de discriminar entre amostras de COVID-19 e amostras de voluntários saudáveis. **Fonte:** O Autor (2024).

5.5.2.2 Modelo PLS-DA: otimização de pré-processamento

A Figura 5.13 mostra os gráficos de *leverage* versus resíduos de *student* do modelo PLS-DA para as amostras dos quatro países estudados. Na Figura 5.13, embora algumas amostras tivessem valores de alavancagem elevados, não foram consideradas valores discrepantes porque estão dentro de ± 3 desvios padrão. Assim, nenhuma amostra foi excluída do conjunto de dados. Vários métodos de pré-processamento (isolados ou em combinação) e validação cruzada foram testados durante o treinamento dos modelos PLS-DA. A combinação de três métodos de pré-processamento (filtragem GLSW, normalização e *autoescale*) e a validação cruzada cega veneziana foram os que tiveram melhor desempenho preditivo (maiores valores de acurácia, sensibilidade e especificidade e menores valores *RMSECV*) em todos os diferentes conjuntos de dados estudados. A Figura 5.14 apresenta gráficos do número de variáveis latentes selecionadas para os modelos de treinamento em função dos baixos valores dos erros de treinamento *RMSECV*.

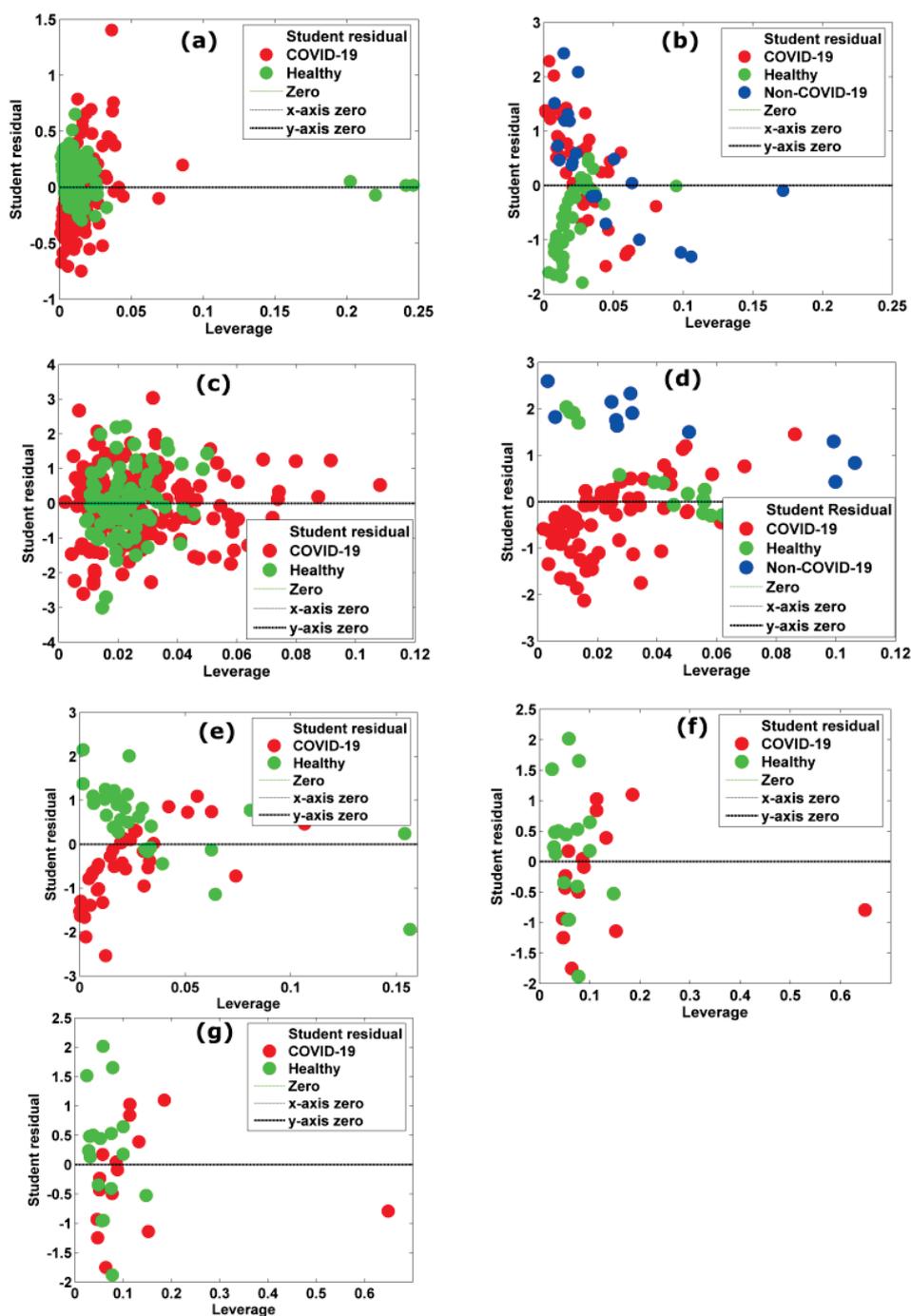


Figura 5.13. Gráfico de *leverage* versus resíduos de estudantes para detecção de valores discrepantes no conjunto de sete conjuntos de dados de amostras de pacientes da Espanha que foram analisadas por MNR (a), de amostras da China que foram analisadas por GC-MS (b), de amostras dos Estados Unidos que foram analisados por LC-MS (c), de amostras da Itália que foram analisadas por GC-MS (d), de amostras da França que foram analisadas por LC-MS (e), de amostras da Itália que foram analisadas por LC-MS (f) e as amostras da Itália que também foram analisadas por GC-MS (g). Neste gráfico, uma amostra é considerada outlier se apresentar simultaneamente valores elevados de Alavancagem e resíduos estudantis. Embora algumas amostras tivessem valores de alavancagem elevados, todas as amostras estavam dentro do intervalo de $\pm 2,5$ desvios padrão dos resíduos dos

alunos, pelo que não foram detectadas amostras discrepantes. Amostras voluntariamente saudáveis de COVID-19 e não-COVID-19 são representadas por círculos vermelhos, azuis e verdes. **Fonte:** O Autor (2024).

Para os dados de análise de plasma por RMN (dados da Espanha, conjunto de dados 1), o modelo PLS-DA foi treinado usando cinco variáveis latentes (Figura 5.14). Para as análises LC-MS da Espanha (conjunto de dados 3), França (conjunto de dados 5) e Itália (conjunto de dados 6), os modelos PLS-DA foram treinados usando sete, duas e duas variáveis latentes, respectivamente. Para amostras de plasma da Itália (conjunto de dados 2) analisadas por GC-MS, o modelo foi treinado usando três LV. Ainda nas análises de GC-MS, o conjunto de dados 4 e o conjunto de dados 7, ambos da Itália, foram treinados usando três LV, respectivamente. Todos os gráficos de *RMSECV* versus número de variáveis latentes são mostrados na Figura 5.14.

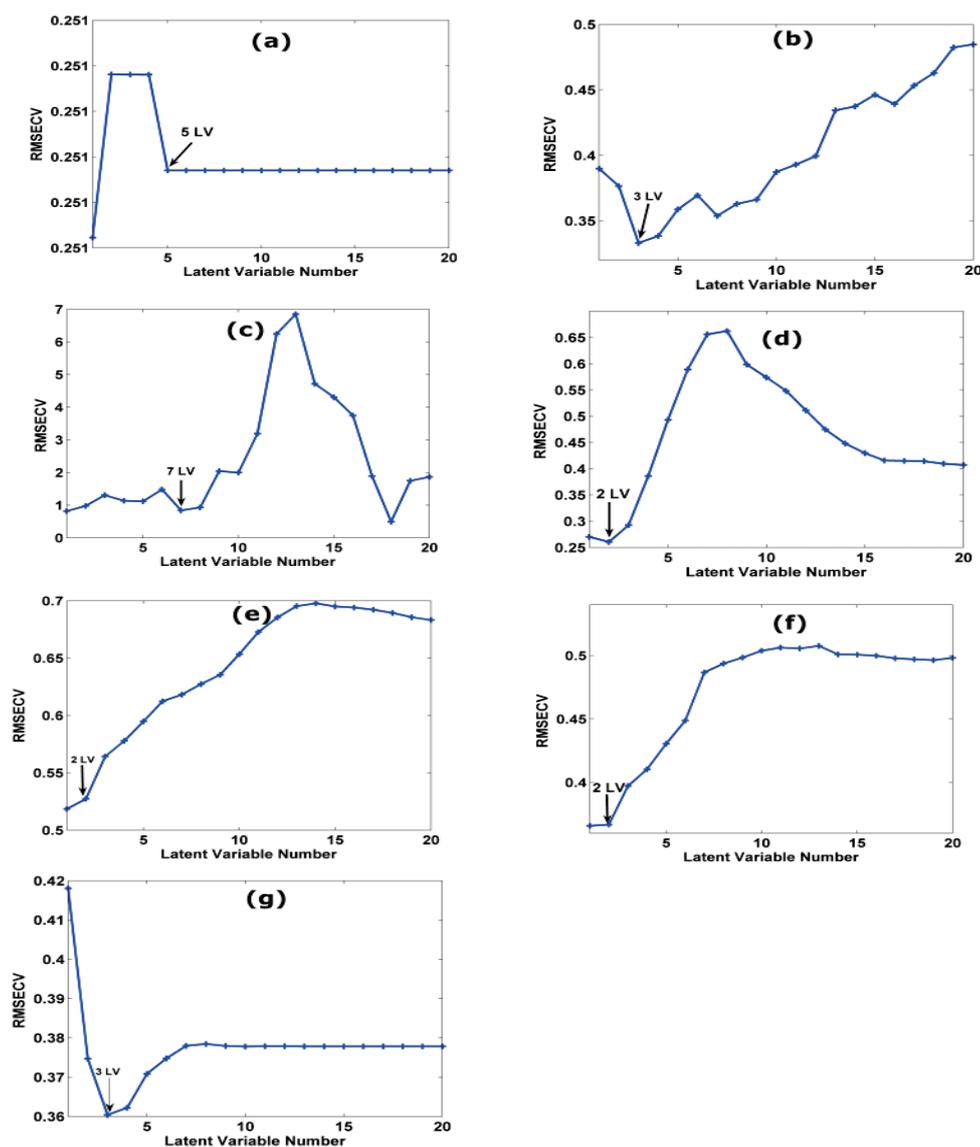


Figura 5.14. Gráficos da Raiz Quadrática Média do Erro da Validação Cruzada (RMSECV) versus Número de Variáveis Latentes (LV). Este gráfico tem como objetivo selecionar o número de variáveis latentes a serem utilizadas para treinamento do modelo PLS-DA. O número do VE é selecionado considerando valores menores de RMSCV. Assim, o número de LVs utilizados para treinar os modelos PLS-DA a partir de dados da Espanha (conjunto de dados 1, dados MNR), China (conjunto de dados 2, dados GC-MS), Espanha (conjunto de dados 3, dados LC-MS), Itália (conjunto de dados 4, dados GC-MS), França (conjunto de dados 5, dados LC-MS), Itália (conjunto de dados 6, dados LC-MS) e Itália (conjunto de dados 7, dados GC-MS) são mostrados nas figuras (a), (b), (c), (d), (e), (f) e (g). **Fonte:** O autor (2024).

O desempenho de todos os modelos PLS-DA na previsão do diagnóstico e gravidade da COVID-19 em todos os países estudados está resumido na **Tabela 5.4.**

Tabela 5.4. Resultados de desempenho dos modelos PLS-DA na previsão do diagnóstico e gravidade da COVID-19 em cada um dos sete conjuntos de dados avaliados.

Dataset	País	Método Bioanalítico	Grupo de pacientes	Total de amostras	VP	FN	VN	FP	Sensibilidade	Especificidade	Acurácia
1	Espanha	MNR	Healthy	280	273	7	251	10	0,98	0,96	0,97
		MNR	COVID-19	261	249	12	269	11	0,95	0,96	0,96
2		GC-MS	Healthy	57	55	2	81	9	0,96	0,90	0,93
	China	GC-MS	COVID-19	60	54	6	83	4	0,90	0,95	0,93
		GC-MS	Non-COVID-19	30	29	1	108	9	0,97	0,92	0,93
3	Espanha	LC-MS	Healthy	133	125	8	268	12	0,94	0,96	0,95
		LC-MS	COVID-19	254	243	11	149	10	0,96	0,94	0,95
4	Itália	GC-MS	Healthy	26	26	0	372	15	1,00	0,96	0,96
		GC-MS	COVID-19	103	96	7	24	2	0,93	0,92	0,93
5	França	LC-MS	Healthy	45	43	2	51	4	0,96	0,93	0,94
		LC-MS	COVID-19	55	52	3	40	5	0,95	0,89	0,92
6	Itália	LC-MS	Healthy	9	9	0	50	2	1,00	0,96	0,97
		LC-MS	Mild	20	20	0	38	3	1,00	0,93	0,95
		LC-MS	Moderate	16	16	0	42	3	1,00	0,93	0,95
		LC-MS	Severe	16	16	0	44	1	1,00	0,98	0,98
7	Itália	GC-MS	Healthy	24	23	1	21	0	0,96	1,00	0,98
		GC-MS	COVID-19	21	21	0	22	2	1,00	0,92	0,96

Nota: GC-MS: Gas Chromatography-Mass Spectrometry; LC-MS: Liquid Chromatography-Mass Spectrometry, NMR: Nuclear Magnetic Resonance Spectroscopy; VP: verdadeiro positivo, VN: verdadeiro negativo; FN: falso positivo e FP: falso negativo. Fonte: O Autor (2024). **Fonte:** O Autor (2024)

5.5.2.3 Identificação dos biomarcadores mais importantes para a predição do diagnóstico e prognóstico da COVID-19

Os gráficos VIP das variáveis mais importantes no diagnóstico da COVID-19 são mostrados nas Figuras 5.15-5.21, ao passo que o gráfico 7 mostra as variáveis mais importantes na previsão da severidade de COVID-19.

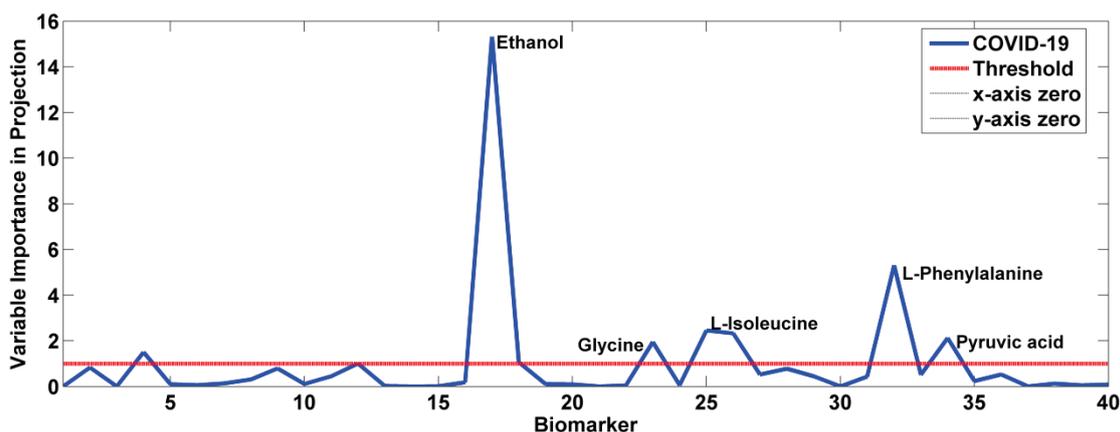


Figura 5.15. Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 utilizando dados da Espanha (conjunto de dados 1, dados MNR). O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na previsão do diagnóstico de COVID-19. A linha preta tracejada paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP = 1). Os metabólitos que contribuem significativamente para a predição das diferentes classes das amostras estão acima do limite (pontuação VIP > 1). **Fonte:** O Autor (2024).

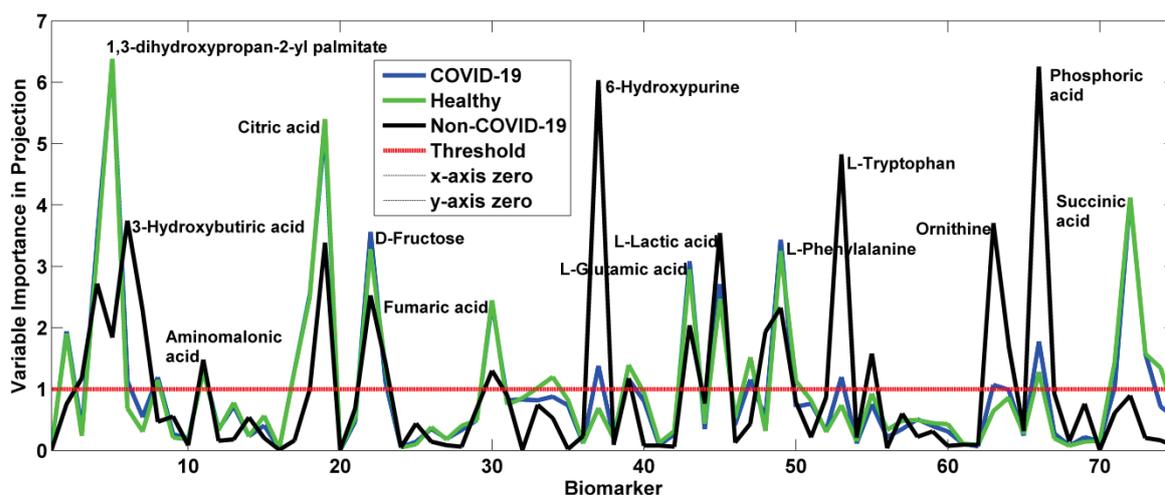


Figura 5.16. Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 utilizando dados da China (conjunto de dados 2, GC-MS Data). O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na previsão das diferentes classes de amostras (COVID-19 representada pela cor azul, não-COVID-19 pela cor preta e voluntários saudáveis pela cor verde). A linha preta tracejada paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP = 1). Os metabólitos que contribuem significativamente para a predição das diferentes classes de amostras estão acima do limite (pontuação VIP > 1); os 10 principais biomarcadores foram destacados na figura. **Fonte:** O Autor (2024).

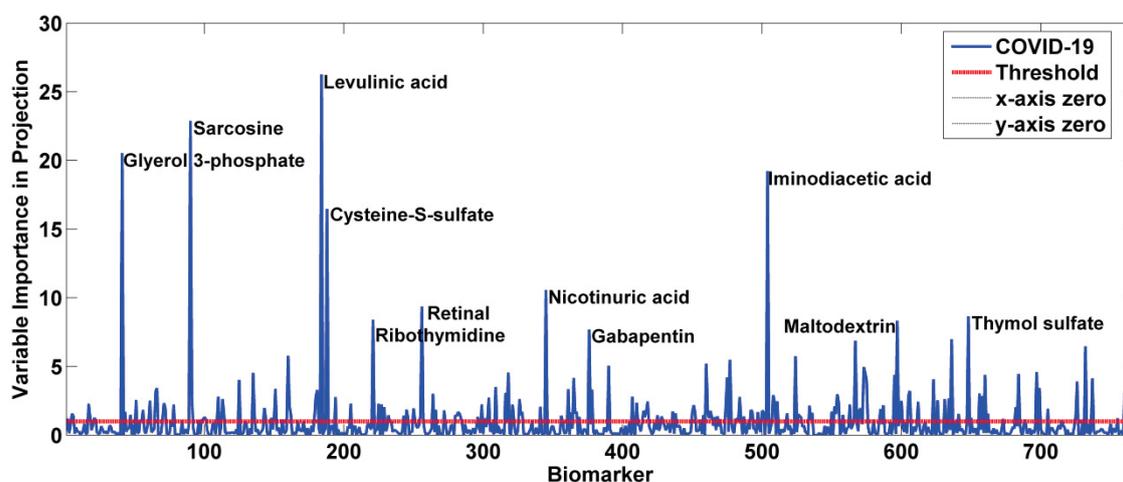


Figura 5.17. Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 usando conjunto de dados da Espanha (conjunto de dados 3, dados LC-MS). O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na previsão do diagnóstico de COVID-19. A linha preta tracejada paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP = 1). Os metabólitos que contribuem significativamente para a predição das diferentes classes das amostras estão acima do limite (pontuação VIP > 1). **Fonte:** O Autor (2024).

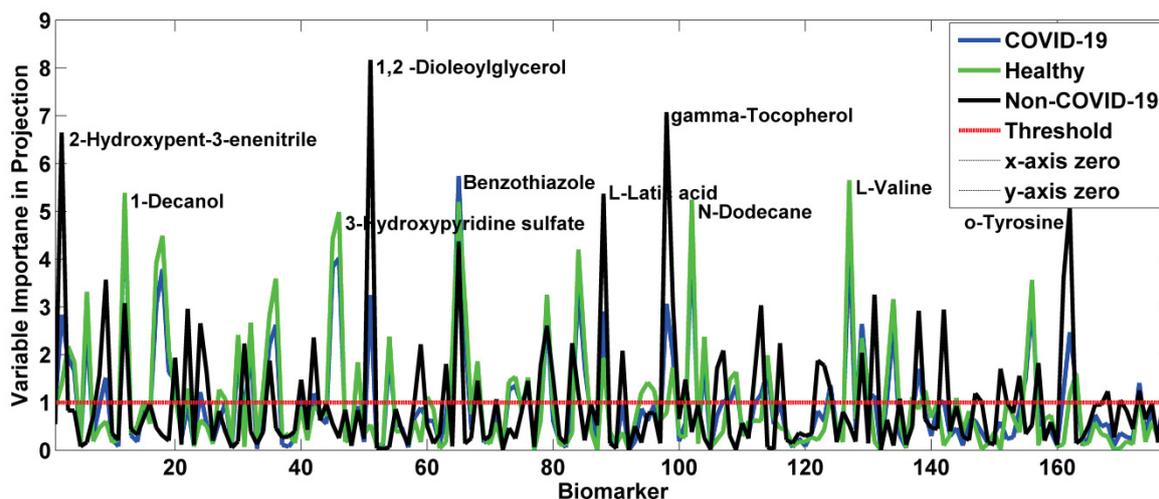


Figura 5.18. Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 utilizando dados da Itália (conjunto de dados 4, GC-MS Data). O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na previsão das diferentes classes de amostras (COVID-19 representada pela cor azul, não-COVID-19 pela cor preta e voluntários saudáveis pela cor verde). A linha preta tracejada paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP = 1). Os metabólitos que contribuem significativamente para a previsão das diferentes classes das amostras estão acima do limite (pontuação VIP > 1). **Fonte:** O Autor (2024)

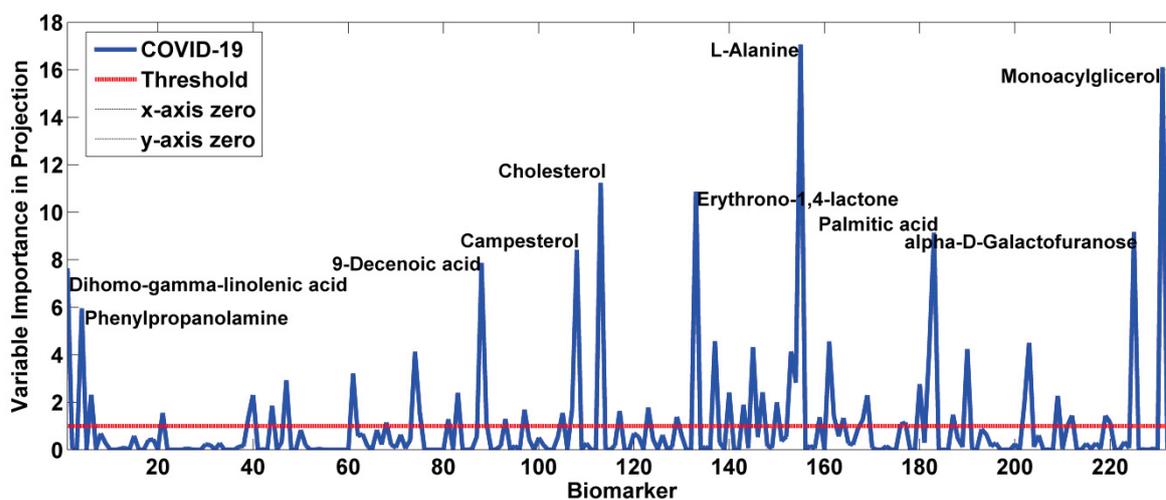


Figura 5.19. Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 utilizando dados da França (conjunto de dados 5, GC-MS Data). O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na previsão das diferentes classes de amostras (COVID-19 representada pela cor azul, não-COVID-19 pela cor preta e voluntários saudáveis pela cor verde). A linha preta tracejada paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP = 1). Os metabólitos que contribuem significativamente para a previsão das diferentes classes das amostras estão acima do limite (pontuação VIP > 1). **Fonte:** O Autor (2024).

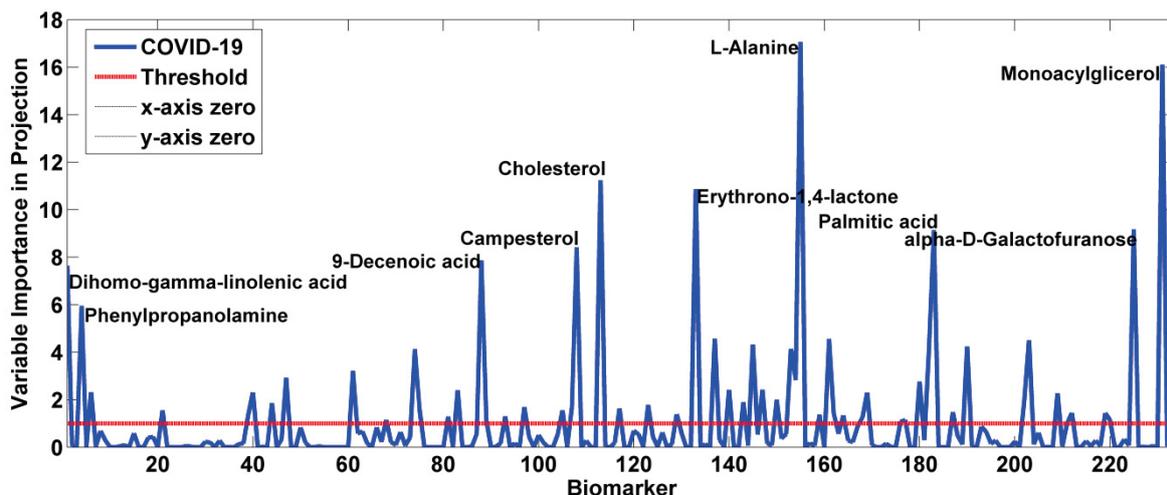


Figura 5.20. Importância da variável no gráfico de projeção dos biomarcadores mais importantes para o diagnóstico de COVID-19 utilizando dados da Itália (conjunto de dados 7, dados LC-MS). O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na previsão do diagnóstico de COVID-19. A linha preta tracejada paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP = 1). Os metabólitos que contribuem significativamente para a predição das diferentes classes das amostras estão acima do limite (pontuação VIP > 1). **Fonte:** O Autor (2024).

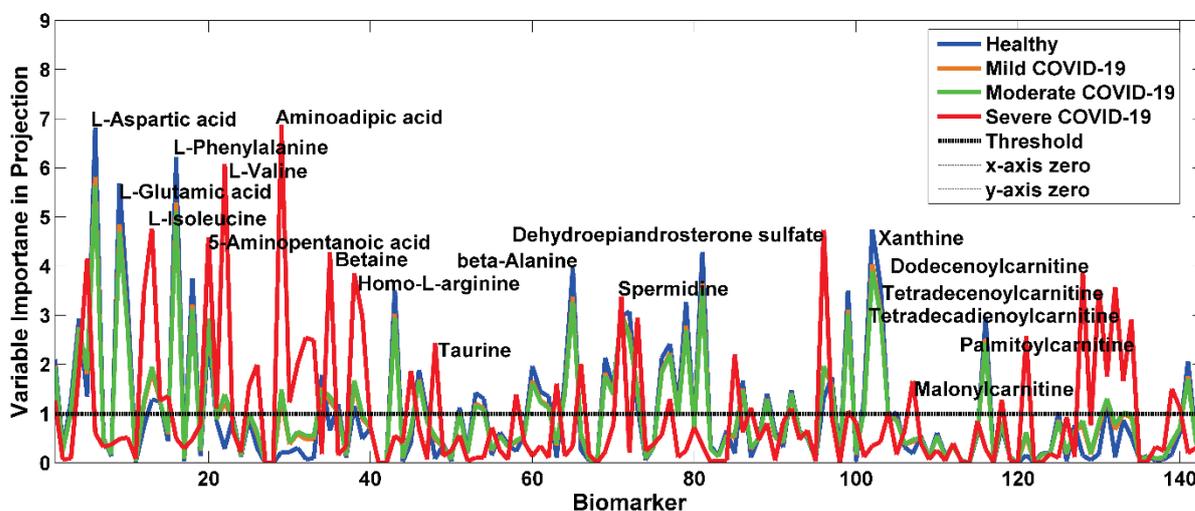


Figura 5.21. Importância variável no gráfico de projeção (VIP) dos biomarcadores mais importantes para a gravidade da COVID-19 usando dados da Itália (dados LC-MS, conjunto de dados 6). O eixo X representa todos os metabólitos analisados; O eixo Y representa a pontuação VIP que reflete a importância de cada metabólito na previsão das diferentes classes de amostras (COVID-19 grave representado pela cor rosa COVID-19 moderado representado pela cor azul, COVID-19 leve representado pela cor verde e voluntários saudáveis pela cor vermelha cor). A linha preta tracejada paralela ao eixo X representa o limite de pontuação VIP (limiar de pontuação VIP = 1). Os metabólitos que contribuem significativamente para a predição das diferentes

classes das amostras estão acima do limite (pontuação VIP > 1). **Fonte:** O Autor (2024).

Os gráficos *boxplot* que mostram a gama de biomarcadores associados à gravidade são mostrados na Figura 5.22 (dados de LC-MS da Itália, conjunto de dados 6). Por outro lado, os *boxplots* dos biomarcadores associados ao diagnóstico para todos os outros países são mostrados nas Figuras 5.23-5.25. Para os pacientes da Espanha (conjunto de dados 1, análises de RMN), os pacientes com COVID-19 apresentaram níveis mais elevados de L-fenilalanina e ácido pirúvico do que o grupo de voluntários saudáveis (Figura 9). Para as análises de GC-MS da China (conjunto de dados 2), os pacientes com COVID-19 apresentaram níveis mais elevados de ácido succínico, L-fenilalanina, ácido láctico, ácido glutâmico, ácido fumárico e D-frutose do que voluntários saudáveis ou pacientes com outras pneumonias (não- COVID-19); no entanto, os pacientes com COVID-19 apresentaram níveis mais baixos de palmitato de 1,3-di-hidroxiopropan-2-il e ácido cítrico do que indivíduos saudáveis ou não-covid-19.

Na análise de LC-MS de dados dos EUA (conjunto de dados 3), os pacientes com COVID-19 apresentaram níveis mais elevados de glicerol-3-fosfato, retinal e cisteína-S-sulfato do que voluntários saudáveis. Por outro lado, os biomarcadores sarcosina, ácido levulínico, ribotimidina e ácido iminodiacético estavam em níveis mais baixos em pacientes com COVID-19 do que em voluntários saudáveis. Para dados da França analisados por LC-MS (conjunto de dados 4), apenas o biomarcador 1,2 dioleoilglicerol apresentou níveis mais elevados em pacientes com COVID-19 do que em voluntários saudáveis. Os seguintes biomarcadores já estavam em níveis mais baixos em pacientes com COVID-19 do que em voluntários saudáveis ou não-COVID-19: 2-hidroxipent-3-enetrila, 1-decanol, sulfato de 3-hidroxipiridina, benzotiazol e ácido L-láctico. Para os dados de diagnóstico das amostras analisadas por LC-MS (conjunto de dados 5), os pacientes com COVID-19 apresentaram níveis mais elevados de ácido 2-hidroxi-butírico, citosina, asparagina, isoleucina e N-acetil glucosamina-1-fosfato do que os voluntários saudáveis. No entanto, os pacientes com COVID-19 apresentavam níveis baixos de indoxil sulfato, miristato de glicerol e N-acetil triptofano (figura 13). Finalmente, para os dados de diagnóstico da Itália analisados pelo GC-MS (conjunto de dados 7), os seguintes biomarcadores estavam em níveis mais elevados em

pacientes com COVID-19 do que em voluntários saudáveis: ácido dihomogama-linoléico, ácido 9-decenóico, campesterol e eritromo-1,4-lactona.

Para os dados de gravidade da Itália analisados por LC-MS (conjunto de dados 6), os seguintes biomarcadores estavam em níveis mais elevados em pacientes com COVID-19 grave do que em pacientes com COVID-19 leve: espermidina, taurina, L-aspartico, L-glutâmico, L-fenilalanina e xantina. Pacientes com COVID-19 grave apresentaram níveis mais baixos dos seguintes biomarcadores quando comparados aos pacientes com COVID-19 leve: ácido 5-aminopentanóico, sulfato de diidropiandrosterona, dodecenoilcarnitina e L-palmitoilcarnitina.

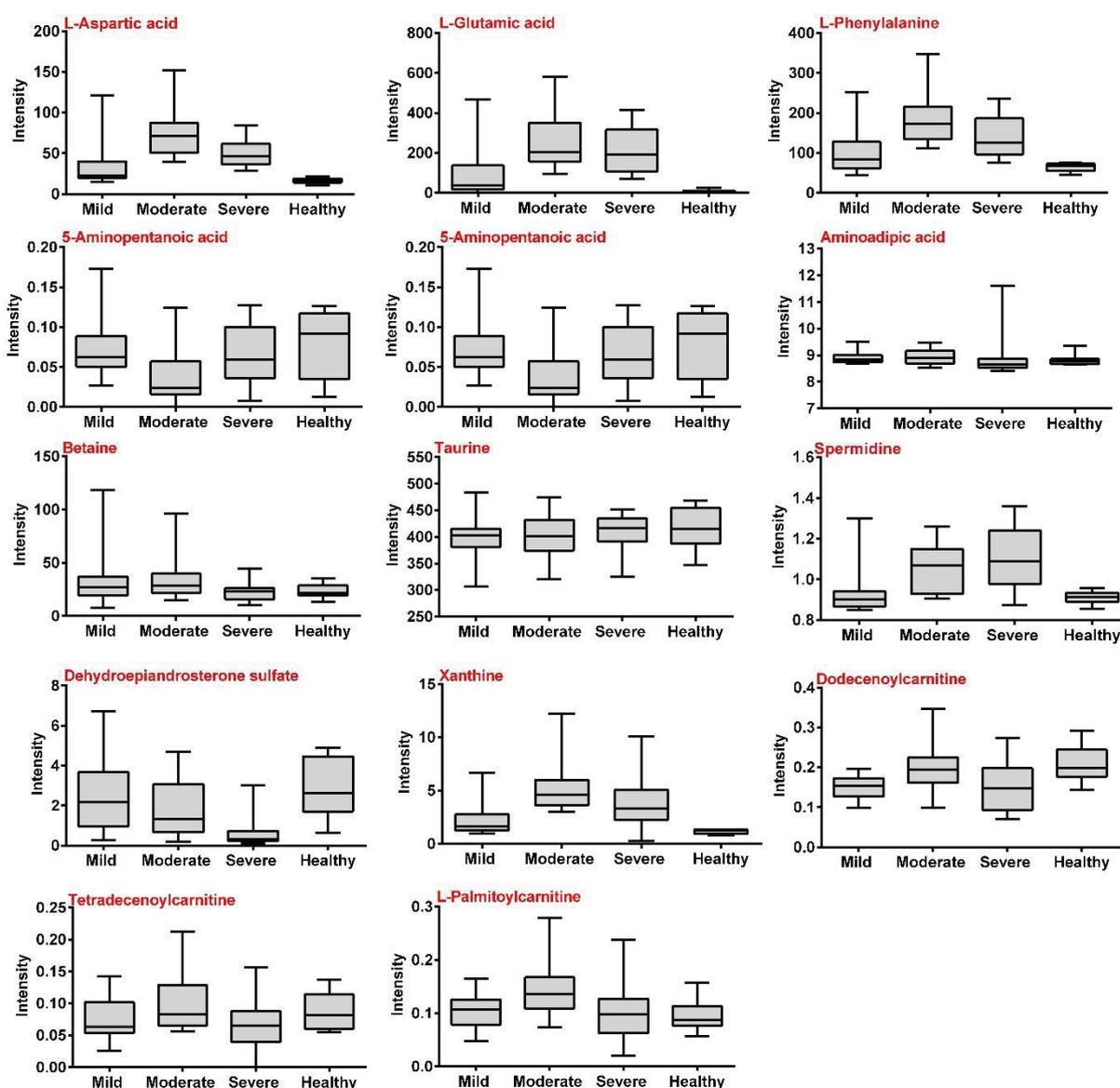
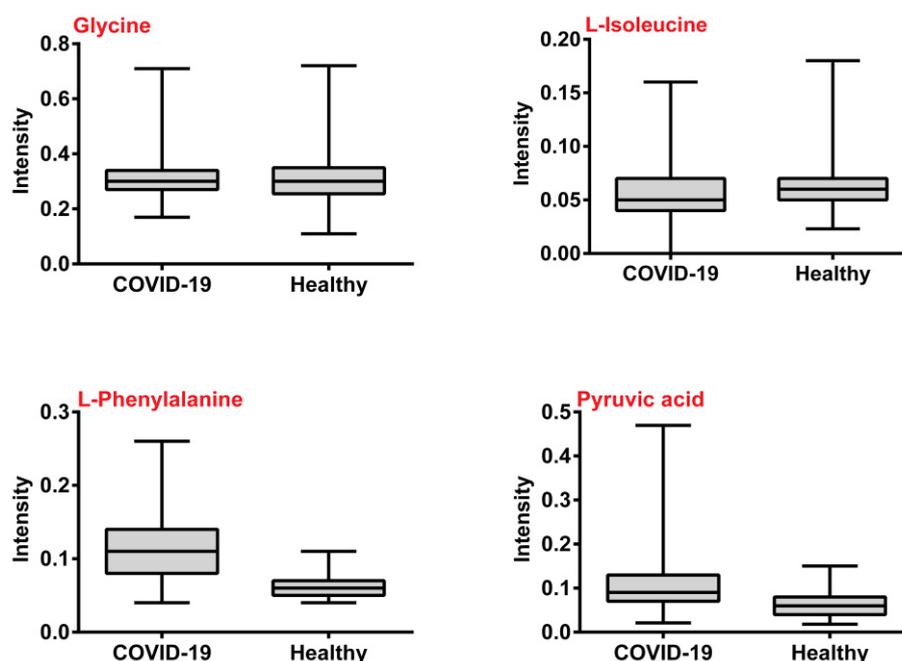


Figura 5.22. Perfil dos principais biomarcadores sanguíneos associados à gravidade da COVID-19 utilizando dados da Itália (conjunto de dados LC-MS, conjunto de dados

6). Os resultados são agrupados de acordo com classes: pacientes com COVID-19 grave (n = 16), COVID-19 moderado (n = 16), COVID-19 leve (n = 20) e voluntários saudáveis (n=9). As caixas indicam intervalos interquartis (mediana); linhas horizontais indicam valores mínimos e máximos. **Fonte:** O Autor (2024).

Figura 5.23. Perfil dos principais



biomarcadores sanguíneos associados ao diagnóstico de COVID-19 utilizando dados de Espanha (conjunto de dados MNR). Os resultados estão agrupados de acordo com as classes: saudável (n = 280) e COVID-19 (n = 261). As caixas indicam intervalos interquartis (mediana); linhas horizontais indicam valores mínimos e máximos. **Fonte:** O Autor (2024).

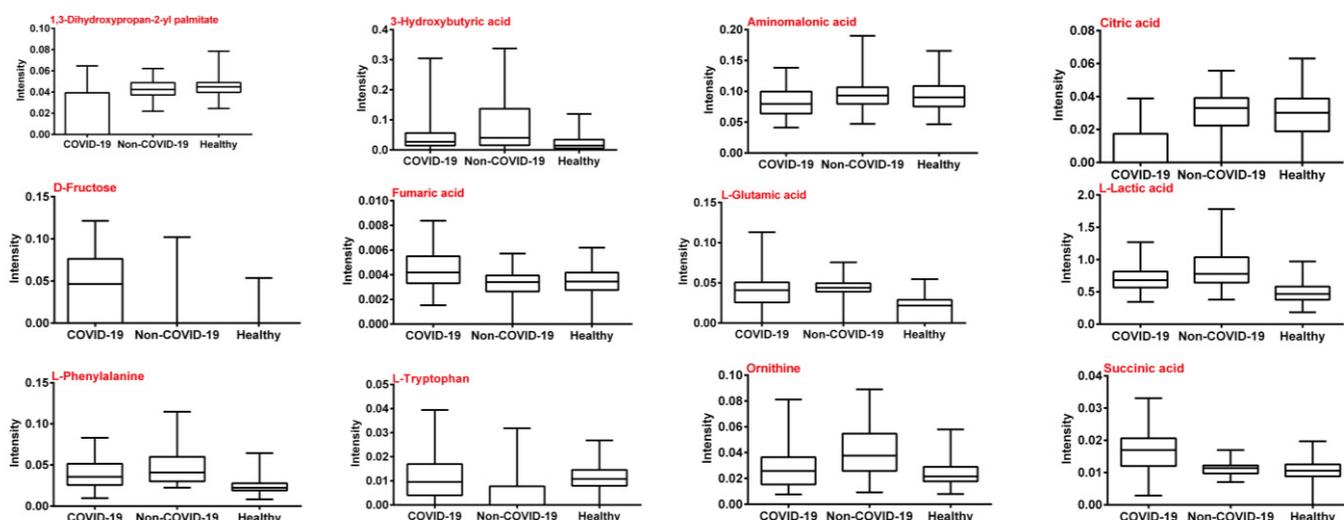


Figura 5.24. Perfil dos principais biomarcadores sanguíneos associados ao diagnóstico de COVID-19 utilizando dados da China (conjunto de dados GC-MS). Os resultados estão agrupados de acordo com as classes: saudável (n = 57), não-COVID-19 (n = 30) e COVID-19 (n = 60). As caixas indicam intervalos interquartis (mediana); linhas horizontais indicam valores mínimos e máximos.

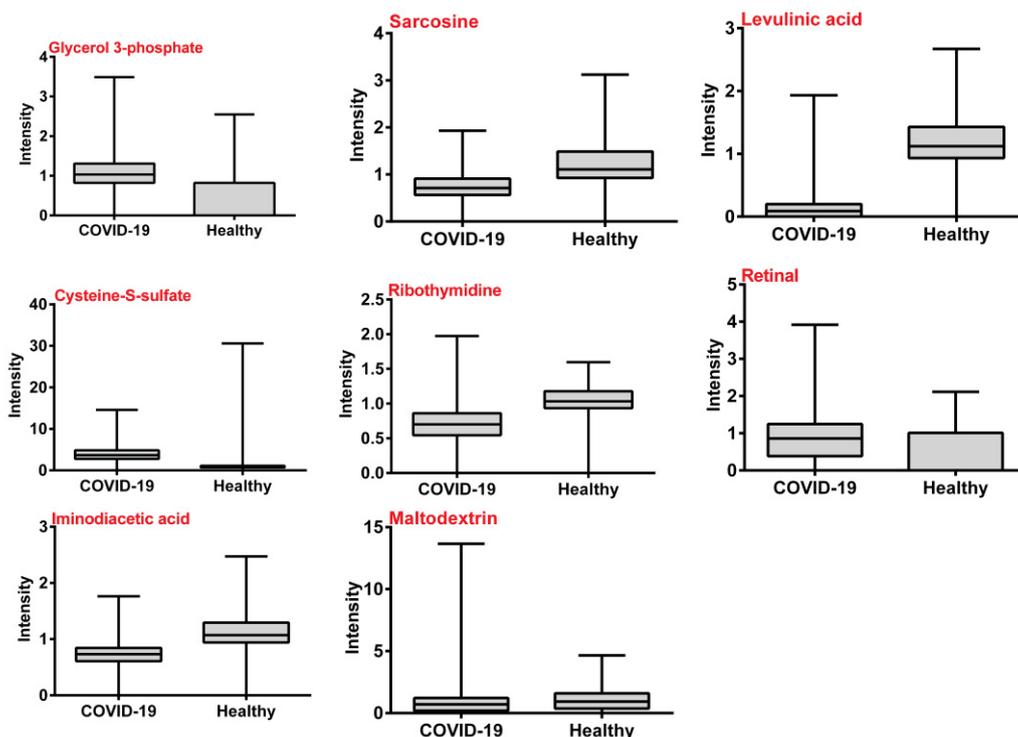


Figura 5.25. Perfil dos principais biomarcadores sanguíneos associados ao diagnóstico de COVID-19 utilizando dados dos da Itália (conjunto de dados LC-MS). Os resultados estão agrupados de acordo com as classes: saudável (n = 133) e COVID-19 (n = 254). As caixas indicam intervalos interquartis (mediana); linhas horizontais indicam valores mínimos e máximos. **Fonte:** O Autor (2024).

A Tabela 5.5 mostra os biomarcadores comuns encontrados nos diferentes conjuntos de dados que foram importantes na previsão do diagnóstico e da gravidade da COVID-19. É importante ressaltar que alguns desses biomarcadores já haviam sido previamente identificados pelo nosso grupo, em estudos anteriores envolvendo outros pacientes da COVID-19.

Tabela 5.5. Biomarcadores importantes na previsão do diagnóstico e gravidade da COVID-19 foram comuns em duas ou mais bases de dados diferentes investigadas no estudo

Biomarcador	Thaizou dataset*(China)	Wuhan dataset*(China)	Dataset 1 (Espanha)	Dataset 2 (China)	Dataset 3 (Espanha)	Dataset 4 (Itália)	Dataset 5 (França)	Dataset 6 (Itália)	Dataset 7 (Itália)	Desfecho
Alanina	+	-	-	-	-	-	-	+	+	Diagnóstico
Ribotimidina	+	-	-	-	+	-	-	-	-	Diagnóstico
L-Ornitina	+	+	-	+	-	-	-	+	-	Severe
L-Valina	-	-	-	-	-	+	-	+	+	Severidade
Ácido L-Glutâmico				+	-	-	-	+	+	Severidade
Isoleucina	+	-	-	-	-	-	-	+	+	Severidade
Fosfato de N-acetil-glucosamina-1	+	-	-	-	-	+	-	-	-	Severidade
2,6 dihidroxipurina	-	-	-	+	-	-	-	-	-	Diagnóstico
4-Hidroxiifenilacetol carnitina	-	+	-	-	-	-	-	+	-	Severidade

Nota: (+) presente; (-) ausente.; * Conjunto de dados do nosso estudo anterior. **Fonte:** O Autor (2024).

5.6 DISCUSSÃO

5.6.1. ESTUDO I: DIAGNÓSTICO E PROGNÓSTICO DE COVID-19 EMPREGANDO ANÁLISE DE PLASMA E SORO VIA LC-MS E MACHINE LEARNING

Conseguimos desenvolver e avaliar o desempenho de sete modelos diferentes baseados em ML (ANNDA, PLS-DA, XGBoostDA, SIMCA, SVM, LREG e KNN) para prever o diagnóstico de COVID-19, bem como a gravidade e mortalidade da doença usando plasma e amostras de soro de pacientes de dois hospitais de referência na China. Mais de 1300 metabólitos solúveis em água e solúveis em gordura foram avaliados. Como o curso da COVID-19 é extremamente variável entre os pacientes (especialmente devido às mutações emergentes do SARS-Cov-2 e à variabilidade biológica), o perfil metabolômico desses casos também é incerto, exigindo, portanto, modelos mais robustos baseados em ML ^{428–430}.

Em nosso estudo, o modelo PLS-DA apresentou o melhor desempenho (AUC ROC 87%–97%), com valores de acurácia semelhantes aos de outros modelos baseados em ML disponíveis na literatura nesta área (AUC ROC 70%–99%) ²⁶¹. O modelo PLS-DA é atualmente um dos algoritmos baseados em ML mais comumente usados para analisar dados de metabolômica e outras ciências ômicas (por exemplo, genômica, transcriptômica e proteômica), sendo recomendado por especialistas na área ^{431,432}. Na verdade, uma revisão sistemática que avaliou o número de citações de estudos publicados entre 1990 e 2018 e disponíveis na Web of Science mostrou um aumento nas publicações que citam o PLS-DA (n=2242), enquanto outros algoritmos (por exemplo, ANN, SVM, RF, regressão logística, aprendizagem profunda) foram menos comumente mencionados (n = 500) ⁴³³. As principais razões para isso incluem as características intrínsecas do PLS-DA, considerado um algoritmo versátil com melhores vantagens preditivas e descritivas em relação a outros modelos. Um estudo recente realizado por Mendes et al., (2019) ⁴³² comparou o desempenho preditivo de oito algoritmos de ML (PLS-DA, ANN, não-ANN, floresta aleatória (RF, máquinas de vetores de suporte de kernel de função de base radial (SVM), regressão logística e regressão de componentes principais (PCR)), usando 10 conjuntos de dados de metabolômica clínica disponíveis nos repositórios Metabolights e Metabolomics Workbench, e relatou o PLS-DA como o modelo com melhor desempenho ⁴³².

O PLS-DA pode prever dados altamente multivariados em um espaço de coordenadas menores chamado LV (ou componentes principais) que descrevem a variação entre os dados de entrada (por exemplo, metabólitos) e os dados de saída (por exemplo, classe de amostra) antes de regredir para uma variável dependente. Isso permite que conjuntos de dados com mais variáveis do que amostras sejam modelados sem recorrer a variáveis de pré-seleção (essencial para estudos geradores de hipóteses). Além disso, ao considerar o LV, problemas relativos à multicolinearidade entre os diferentes metabólitos em qualquer sistema biológico podem ser evitados (ou seja, os LV não se correlacionam entre si) ^{432,434}. Uma vez otimizado, o modelo PLS-DA pode ser reduzido a um modelo de regressão linear comum, permitindo prever o valor de cada metabólito/biomarcador no conjunto de dados ^{431,432}. Outros modelos baseados em ML, incluindo RNA multicamadas, geralmente requerem tamanhos de amostra maiores para alcançar um alto desempenho preditivo. Como consequência, o número de variáveis incluídas nesses modelos é menor que o número de amostras ^{292,293,435,436}, o que não é o cenário da maioria dos conjuntos de dados metabolômicos ²⁷¹. Em nosso estudo, conseguimos usar um conjunto de dados incluindo 180 amostras e 1.300 variáveis.

Considerando a grande complexidade dos dados metabolômicos e as propriedades intrínsecas dessas informações, valores faltantes, heterocedasticidade, parâmetros pouco informativos e variabilidade biológica são comuns. O pré-processamento de dados é, portanto, fundamental para melhorar a qualidade da informação, transformando a matriz de dados brutos em um conjunto 'mais limpo' ^{281,420}. Várias estratégias de pré-processamento estão disponíveis, incluindo imputação de dados ausentes, filtragem, transformações, normalização baseada em amostra, normalização baseada em metabólito, normalização baseada em amostra e metabólito e baseada em padrão interno ^{281,282}. Este processo pode ser realizado em ferramentas online gratuitas como MetaboAnalyst, NOREVA, ANPELA, NormalizeMets, MMEASE e Data Analysis ^{281,282,420,437-441}. Outro desafio na análise metabolômica é a integração de dados de diferentes experimentos e a remoção simultânea de variações biológicas e experimentais indesejadas ⁴⁴¹[55]. MMEASE é uma ferramenta online que permite mesclar esses dados e remover o efeito de variações indesejadas entre amostras, o que aumenta a eficiência das análises estatísticas e leva a resultados mais robustos e confiáveis ⁴⁴¹. Os dados são mesclados de acordo com o ID de alinhamento para tempo de retenção (TR) e massa

exata (m/z) de um determinado metabólito considerado como referência. Se tanto o TR quanto o m/z do metabólito de referência estiverem dentro da faixa tolerável, este procedimento é automaticamente aplicado aos metabólitos nos cromatogramas das amostras restantes de outros experimentos ⁴⁴¹. Uma alternativa para eliminar problemas do efeito lote é utilizar o método Z-score, que transforma os dados para média zero e desvio padrão de 1, normalizando a distribuição dos sinais analíticos [49]. Em nosso estudo, como apenas dados espectrais estavam disponíveis (o TR dos metabólitos estava ausente), o método MMEASE não foi empregado. No entanto, os conjuntos de dados foram padronizados usando o método de pré-processamento GLSW, que calcula uma matriz de filtros com base nas diferenças entre grupos de amostras que de alguma forma deveriam ser 'semelhantes' [28,56]. Segundo este método, no caso de problemas de classificação, amostras semelhantes seriam aquelas cujos dados das mesmas amostras foram analisados em instrumentos diferentes ou mesmo em períodos diferentes ⁴⁴². Em nosso estudo, utilizamos dois bancos de dados de pacientes com COVID-19 de dois experimentos diferentes, cujas amostras foram obtidas por dois modelos diferentes de equipamentos de LC-MS e em diferentes períodos ⁴⁴².

Os procedimentos metodológicos para otimizar o processamento de dados metabolômicos estão mais bem descritos nas diretrizes da NOREVA (Normalização e Avaliação de Dados Metabolômicos baseados em MS), NormalizeMets, MMEASE, MetaboAnalyst e ANPELA ^{281,282,420,437-441}.. O pré-processamento de dados geralmente é realizado em cinco etapas: filtragem de dados e imputação de valores faltantes (S1), correção de amostras de controle de qualidade (S2), transformação de dados (S3), normalização de dados (S4) e avaliação de desempenho (S5). Durante S1, a filtragem concentra-se na remoção de características não informativas consideradas propriedades intrínsecas dos dados metabolômicos, enquanto a imputação busca substituir valores ausentes ou inválidos decorrentes de razões técnicas/biológicas por valores específicos baseados em informações disponíveis, preservando assim a estrutura do conjunto de dados, e reduzindo a imprecisão ou limitação das análises. A correção de amostras de controle de qualidade (S2) visa reduzir a interferência de sinais prejudiciais ou incontroláveis nos dados metabolômicos para garantir a estabilidade e consistência dos dados baseados em amostras de controle de qualidade. Isto permite corrigir problemas relacionados à variação na intensidade do sinal, variabilidade intra e inter-amostra e desvios na

precisão da qualidade. Nas etapas S3 e S4, a transformação e normalização dos dados metabolômicos visa corrigir problemas de heterocedasticidade e variações indesejadas, transformando a distribuição de dados assimétricos em simétricos, preservando as variáveis existentes. Por fim, S5 consiste em avaliar o desempenho dos dados de pré-processamento com base em cinco critérios: (i) capacidade de reduzir a variação intragrupo entre amostras (métrica: desvio absoluto mediano agrupado); (ii) efeito na análise metabólica diferencial (métrica: pureza); (iii) consistência do método em marcadores descobertos em diferentes conjuntos de dados (métrica: consistência relativa ponderada); (iv) influência do método na precisão da classificação (métrica: área sob a curva); (v) nível de correspondência entre dados normalizados e de referência (métrica: alterações logarítmicas das concentrações) 281,282,420,437–441. Em nosso estudo, as etapas de pré-processamento de dados mencionadas acima foram seguidas (ou seja, foram empregadas imputação de valores faltantes e filtragem de dados por meio de GLSW ⁴⁴²; os dados foram normalizados usando escala automática ⁴⁴⁰. A imputação de mediana é um método amplamente utilizado em metabolômica, pois, diferentemente da média, não é afetada por valores extremos (outliers), o que preserva a estrutura dos dados e fornece um valor mais confiável do conjunto de dados ⁴⁴⁰. A autoescala é uma abordagem baseada na centralização da média seguida pela divisão de cada coluna ou variável (por exemplo, proteína ou qualquer outro metabólito) pelo desvio padrão da coluna, assumindo que todos os metabólitos são igualmente importantes ⁴⁴².

Estudos recentes utilizando amostras de sangue ou urina de pacientes diagnosticados com COVID-19 destacaram que alguns biomarcadores predizem a gravidade e a letalidade da doença. Yao et al. (2020), utilizando o modelo SVM, descobriram que níveis elevados de neutrófilos estavam associados a casos mais graves [58], enquanto Patterson et al. (2020), por meio do modelo random floresta, destacaram que um aumento de interleucina 6 (IL-6) e interferon-gama (IFN- γ) está relacionado a um pior prognóstico ⁴⁴³. Por outro lado, utilizando o software SOLO (Eigenvector Research), encontramos diferentes e novos biomarcadores potencialmente associados ao curso da doença. Altos níveis de ribotimidina, 4-hidroxifenilacetoilcarnitina e uridina foram associados à positividade para COVID-19, enquanto altos níveis de N-acetil-glucosamina-1-fosfato, cisteinilglicina, isobutirato de metila, ornitina e 5,6-di-hidro-5-metiluracil foram relacionados à gravidade e mortalidade da COVID-19. As diferenças entre os resultados do estudo podem ser

devidas às diferentes amostras (ou seja, tipo de amostra, origem), ao curso fisiopatológico multifatorial da doença que ainda não foi totalmente elucidado, bem como aos diferentes métodos analíticos/ modelos empregados pelos autores ^{444,445}. Em relação a este último, também descobrimos que as análises realizadas no software SOLO resultaram em modelos com maior desempenho preditivo em comparação com aqueles do Metaboanalyst 5.0 e identificaram diferentes biomarcadores para diagnóstico de COVID-19 e previsão de gravidade/letalidade. Isto pode ser devido às diferenças nos métodos de pré-processamento. Enquanto o SOLO permite a combinação de autoescalamento e GLSW, o Metaboanalyst 5.0 aplica apenas esta primeira abordagem, o que significa que o GLSW foi, neste caso, um fator determinante para a obtenção de modelos mais robustos. Embora o SOLO não seja um software livre, ele permite a seleção de diferentes estratégias de pré-processamento, proporcionando maior autonomia aos analistas, o que deve ser considerado no desenvolvimento de estudos baseados em ML.

Atualmente, a COVID-19 é amplamente considerada uma doença respiratória e vascular viral. No entanto, pode afetar outros órgãos importantes, como os do trato gastrointestinal e os sistemas hepato-biliar, cardiovascular, renal e nervoso central. Evidências recentes mostram que o SARS-CoV-2 pode causar disbiose na microbiota fecal e modificar o microbioma do trato oral e respiratório, levando a alterações nos níveis de vários metabólitos microbianos no sangue ou nas suas vias metabólicas ⁴⁴⁶⁻⁴⁴⁸. Embora as evidências sobre o assunto ainda sejam escassas, foi relatado que a microbiota é responsável por cerca de 50% de todos os metabólitos do sangue, o que levanta questões sobre o seu papel em doenças multifatoriais, como a COVID-19 ^{447,448}.

Li et al. (2019), ao avaliarem o perfil da microbiota nasofaríngea de pacientes com COVID-19, descobriram que as amostras positivas foram significativamente enriquecidas com a assinatura de dois táxons bacterianos (*Cutibacterium* e *Lentimonas*) e tiveram menor abundância de outros táxons bacterianos, incluindo *Prevotellaceae*. Esta última é uma família do filo *Bacteroidetes* comumente encontrada na microbiota oral e fecal, recentemente associada ao metabólito ribotimidina (nucleosídeo metilado), que foi aumentado nas amostras positivas para COVID-19 em nosso estudo. Quando superexpressas, essas proteínas contribuem ativamente para a gravidade da pneumonia e dos sintomas semelhantes à pneumonia e são, portanto, potenciais biomarcadores para o diagnóstico e gravidade da doença

^{449,450}. Da mesma forma, níveis elevados de 2-undecanona, um composto orgânico volátil de cadeia longa geralmente produzido durante infecções bacterianas adquiridas em hospitais causadas por *Pseudomonas aeruginosa*, podem estar associados a casos graves de infecções respiratórias, incluindo COVID-19^{451,452}. Esta substância pode ser encontrada em pacientes com fibrose cística ⁴⁵³. Na verdade, a fibrose pulmonar é uma complicação grave de algumas pneumonias virais, muitas vezes levando à dispneia e ao comprometimento da função pulmonar. Descobriu-se que pacientes com COVID-19 confirmado apresentavam diferentes graus de fibrose pulmonar durante e após a alta hospitalar ⁴⁵⁴. A esfingosina 1-fosfato (um produto do metabolismo dos esfingolípídios da membrana ou secretado pelas células), atua através de receptores acoplados à proteína G e regula o tráfego de células imunológicas, diversos processos imunológicos e fibrose ⁴⁵⁵. A via deste metabolito está implicada na função normal da vasculatura pulmonar; parece estar prejudicado na disfunção pulmonar aguda, enquanto é induzido durante a fibrose crônica. Mais estudos sobre a alteração dos níveis deste composto na COVID-19 são necessários para elucidar o seu papel na infecção.

Outro metabolito microbiano, agora associado a bactérias orais que causam cáries e periodontite (por exemplo, *Porphyromonas gingivalis*, *Prevotella* sp. e *Tannerella forsythia*), é o isobutirato de metila ⁴⁵⁶. Análises metagenômicas de pacientes infectados com SARS-CoV-2 demonstraram leituras elevadas de bactérias cariogênicas e periodontopáticas, endossando a noção de uma conexão entre o microbioma oral e as complicações do COVID-19 ⁴⁵⁶. Também encontramos altos níveis de ciclohexilamina (um composto potencialmente cancerígeno eliminado na urina) em pacientes com COVID-19 grave. Isto provavelmente ocorre devido a outra disbiose causada pelo SARS-CoV-2, que permite a hiperproliferação de bactérias intestinais que metabolizam o ciclamato (adoçante artificial ainda utilizado em algumas categorias de alimentos na China) ^{457,458}. Outros compostos comumente encontrados em alimentos e produtos manufaturados (por exemplo, fumaça de tabaco) são os cresóis (xenobióticos). O-cresol e 4-vinilguaiacol são convertidos em sulfatos através do metabolismo de fase II (ou seja, um processo conjunto entre o microbioma e o hospedeiro) e eliminados pela urina ⁴⁵⁹. Estudos anteriores demonstraram baixos níveis de sulfato de o-cresol e sulfato de 4-vinilguaiacol em pacientes com COVID-19, o que pode ser devido às altas taxas de eliminação urinária desses metabolitos (por exemplo, possível dano renal causado pela doença) ⁴⁶⁰.

COVID-19 também pode impactar negativamente o peso corporal e o estado nutricional ⁴⁶¹. Isso pode ocorrer devido à perda de apetite e redução da ingestão de nutrientes, medo e estresse dos pacientes em relação à doença e alterações metabólicas causadas pela infecção. Por exemplo, o metabólito 4-hidroxifenilacetilcarnitina, encontrado aumentado em pacientes com COVID-19 em nosso estudo, pertence ao metabolismo da tirosina e foi previamente associado ao excesso de peso em pacientes com síndrome metabólica. Outros estudos também relataram um aumento na inflamação e nos níveis séricos de leptina em pacientes com COVID-19, assim como em outras doenças infecciosas que podem contribuir para a anorexia ⁴⁶²⁻⁴⁶⁴. Esses metabólitos devem ser investigados como potenciais biomarcadores da gravidade da infecção viral.

Outro metabólito importante é a uridina, um nucleotídeo de pirimidina para a síntese de RNA que está associado à homeostase da glicose, ao metabolismo de lipídios e aminoácidos, à regulação da síntese de glicogênio e à deposição de lipídios (82). Durante o seu catabolismo, a uridina é convertida em β -alanina, seguida de secreção para o cérebro e tecidos musculares. A beta-alanina e a histidina são componentes da carnosina, uma molécula com comprovados efeitos antiinflamatórios, antioxidantes e antiglicantes ⁴⁶⁵. Em nosso primeiro modelo, os níveis de beta-alanina foram baixos em pacientes com COVID-19, enquanto os de uridina eram elevados. Isto pode indicar inibição do catabolismo da uridina durante o curso da infecção. Um estudo recente encontrou uma proporção significativamente baixa de arginina/ornitina entre adultos e crianças infectadas com SARS-Cov-2. Ornitina e citrulina são aminoácidos resultantes da degradação da arginina pela enzima arginase. A depleção dessas substâncias pode contribuir para a disfunção endotelial, desregulação das células T e coagulopatias que são comumente observadas na COVID-19 [84]. O alto nível de ornitina em pacientes com COVID-19 relatado em nosso estudo pode indicar aumento da atividade da enzima arginase.

A N-acetil-glucosamina-1-fosfato (GlcNAc-1-P) é um substrato da via biossintética das hexosaminas, convertida pela enzima UDP-GlcNAc pirofosforilase em UDP-GlcNAc (este metabólito pode usar a rota O-glicosilação). Esta conversão é um passo importante na produção de citocinas durante a infecção pelo vírus influenza, conforme demonstrado em modelos in vivo (modelos murinos) ⁴⁶⁶. Os pesquisadores acreditam que a inibição da via da hexosamina é um mecanismo usado por vírus respiratórios, incluindo SARS-Cov-2, para infectar células hospedeiras ^{467,468}. O nível

elevado de GlcNAc-1-P em pacientes com COVID-19 grave revela uma modificação potencial da via biossintética da hexosamina. Adicionalmente, como GlcNAc-1-P é um componente intracelular, a sua presença no plasma indica a existência de dano celular. A infecção por SARS-CoV-2 leva à piroptose, que geralmente é mais prevalente em casos graves. Mais da metade dos pacientes hospitalizados com COVID-19 apresentam níveis elevados de lactato desidrogenase, outro marcador de dano celular. Reguladores do estresse oxidativo, como a cisteinilglicina, um metabólito intermediário na via metabólica da glutatona, também têm sido associados a danos celulares em doenças virais. Altos níveis de cisteinilglicina oxidada foram relatados em indivíduos infectados pelo HIV e relacionados a um maior risco de danos pulmonares na COVID-19, provavelmente devido ao aumento do estresse oxidativo ^{469,470}. Outros metabólitos, como o 5,6-di-hidro-5-metiluracil (di-hidrotimina), um produto intermediário da degradação da timina, podem atuar como marcadores de danos ao DNA [92]. Descobrimos que os níveis desta substância eram elevados em pacientes com COVID-19 grave, mas estudos recentes demonstraram que a proteína spike do SARS-CoV-2 pode inibir a reparação de ADN danificado ⁴⁷¹. Outros metabólitos identificados em níveis extremamente baixos em pacientes com COVID-19 (por exemplo, linoleato, palmitato, ureia, lactato, carnitina) ao usar o software Metaboanalyst 5.0, ou aqueles encontrados em níveis elevados em pacientes que morreram da doença (por exemplo, 6-metilmercaptipurina, L-fenilalanina, ácido tereftálico) também devem ser avaliados mais detalhadamente.

Nosso estudo tem algumas limitações. Embora tenhamos utilizado aproximadamente 1.300 biomarcadores diferentes para treinamento e validação do modelo, estes podem não representar com precisão o universo de metabólitos disponíveis no sangue. Ainda assim, foi possível obter modelos com alto desempenho (acurácia >90%) para predição de diagnóstico, gravidade e letalidade da COVID-19 que podem ser utilizados na prática diária. Sete modelos diferentes baseados em ML baseados em dados de dois conjuntos diferentes da China foram construídos em nosso estudo; no entanto, outros conjuntos de dados e algoritmos podem levar a resultados diferentes.

5.6.2 Estudo II: Novos biomarcadores COVID-19 identificados por meio de análise de dados multi-ômicos: ácido N-acetil-4-O-acetilneuramínico, N-acetil-L-alanina, N-acetiltryptofano, palmitoilcarnitina e 1-miristato de glicerol

Neste estudo, modelos de *machine learning* (PCA e PLS-DA) foram desenvolvidos para analisar sete conjuntos de dados multi-ômicos de pacientes com COVID-19 em diferentes estágios da doença (aguda, leve, moderada e grave) provenientes de cinco países diferentes (Itália), França, Espanha e China) foram analisados para identificar potenciais novos biomarcadores diagnósticos e prognósticos. A partir dessas análises, foram identificados níveis elevados de um total de 23 biomarcadores diagnósticos e prognósticos da COVID-19. Dos 23 metabólitos, quatro deles (N-acetil glucosamina-1-fosfato, ornitina e ribotimidina) já haviam sido descritos na literatura pela primeira vez como novos biomarcadores associados à patogênese e gravidade da COVID-19, em estudo conduzido por nosso grupo de pesquisa envolvendo outros pacientes ¹². Portanto, o presente estudo consolida o papel desses quatro biomarcadores na compreensão dos mecanismos moleculares e fisiopatológicos da doença.

Como o curso da COVID-19 é muito variável (causado principalmente pela variabilidade biológica entre os pacientes por pertencerem a países diferentes, mutações do SARS-Cov-2) a proteômica e metabolômica destes pacientes é muito incerta e complexa, exigindo a aplicação de o modelo *PLS-DA*, considerado modelo de *machine learning* padrão ouro para análise de metabolômica, lipidômica e dados de outras ciências ômicas (por exemplo, proteômica, glicômica, genômica e transcriptômica), devido ao seu maior desempenho preditivo no diagnóstico e prognóstico de doenças ^{431,432}. A acurácia diagnóstica e prognóstica de todos os modelos PLS-DA desenvolvidos em nosso estudo variou de *AUC ROC* 80-96%, estando de acordo com outros estudos anteriores disponíveis na literatura (*AUC ROC* 70-99%) ^{12,261}. Outro fator adicional que aumenta a complexidade dos dados é o fato das amostras terem sido analisadas em três técnicas analíticas diferentes, *RMN*, *LC-MS* e *GC-MS*, permitindo a identificação de uma grande diversidade de biomarcadores diagnósticos e prognósticos, com características diferentes. diferenças físico-químicas, diferentes pesos moleculares e diferentes funções biológicas no organismo do paciente. Isto oferece uma oportunidade para compreender melhor, a nível

molecular, os mecanismos bioquímicos e fisiopatológicos da infecção e da progressão da doença ^{472,473}.

Em nosso estudo, a partir da análise de amostras de plasma por *RMN* (dados da Espanha) e *GC-MS* (dados da China), os pacientes com COVID-19 apresentaram níveis elevados de L-fenilalanina, citosina, asparagina, isoleucina L-aspártica, L - glutâmico do que o grupo de voluntários saudáveis ou pacientes com outras pneumonias. Resultados semelhantes foram encontrados em um estudo recente sobre alterações no perfil de aminoácidos em pacientes com infecção por SARS-CoV-2, que encontraram níveis elevados de fenilalanina, ácido glutâmico, glutamato, triptofano, alanina, glicina e histidina em pacientes adultos com COVID-19 ⁴⁷⁴. Na COVID-19, esses aminoácidos têm sido associados à gravidade da doença e estão diretamente envolvidos em processos catabólicos, pois as citocinas inflamatórias promovem a degradação do tecido muscular resultando na liberação de aminoácidos, que são posteriormente utilizados na gliconeogênese. via para suprir as demandas metabólicas causadas pela infecção por SARS-CoV-2 ^{12,475}. Como os pacientes com COVID-19 apresentam frequentemente hipoxemia devido a dificuldades respiratórias ^{476,477}, o piruvato produzido no citosol, em vez de penetrar na mitocôndria (cadeia respiratória) para produzir energia, é retido no citosol e convertido em lactato (respiração anaeróbica), com baixa produção de energia, processo catalisado pela enzima aldeído desidrogenase ^{478,479}. Como consequência do aumento dos níveis de lactato, é o surgimento da acidose metabólica que causa complicações no quadro clínico do paciente ⁴⁸⁰.

Esses resultados também foram confirmados em nosso estudo, onde pacientes com COVID-19 da China apresentaram níveis mais elevados de lactato do que voluntários saudáveis. Ainda nos dados da China do nosso estudo, foram observados níveis mais baixos de ácido cítrico em pacientes com COVID-19 do que em voluntários saudáveis. Sabe-se que o ácido cítrico é um dos compostos produzidos no ciclo de Krebs (ciclo do ácido cítrico), ao nível das mitocôndrias, mas como o processo infeccioso ocorre em estado anaeróbico, espera-se que os níveis de ácido cítrico sejam baixos, pois moléculas de piruvato são constantemente convertidas em lactato, em vez de produzir acetil coenzima A, que é o substrato inicial do ciclo de Krebs (57). Em nosso estudo, também foram observados níveis mais elevados de ácido glutâmico em pacientes com COVID-19 do que em pacientes saudáveis. Resultados semelhantes foram encontrados em estudos recentes de Paéz-Franco (2022) e

Reverté (2021), onde foram observados níveis mais elevados de ácido glutâmico em pacientes com COVID-19 moderadamente grave do que em pacientes com doença leve ou moderada [58,59]. Nas células do sistema imunológico, a glutamina é convertida em glutamato, aspartato e alanina por oxidação parcial a CO₂, em um processo denominado glutaminólise, e essa conversão desempenha um papel fundamental no funcionamento eficaz das células do sistema imunológico, além disso, através da via das pentoses. fosfato, uma via metabólica paralela à via da glicólise, as células podem produzir ribose-5-fosfato, que é um precursor dos açúcares pentoses observadas na estrutura do RNA e do DNA, bem como glicerol-3-fosfato para a síntese de fosfolipídios ^{481,482}. Outra via biossintética do ácido glutâmico é através do ciclo gama-glutamil, que é a via sintética do GSH, um tripéptido com potente atividade antioxidante. A infecção por COVID-19 leva a interrupções neste ciclo [24,58,62], promovendo um aumento descontrolado dos níveis de ácido glutâmico e piroglutâmico, causando acidose metabólica e insuficiência hepática [63,64]. Em nosso estudo, os níveis de ácido cítrico foram reduzidos em pacientes com COVID-19 do que em pacientes com outras pneumonias ou em voluntários não saudáveis. O citrato é um metabolito intermediário do ciclo de Krebs (que ocorre nas mitocôndrias) e, devido à baixa disponibilidade de oxigênio causada pela COVID-19, isto resulta em níveis reduzidos de todos os metabolitos do ciclo de Krebs, incluindo o citrato, resultando na diminuição da produção de ATP. Resultados semelhantes aos nossos foram encontrados no recente estudo de Shi (2021); e isso comprova a importância desse metabolito na patogênese e diagnóstico da doença ⁴⁸³.

Em nosso estudo, os níveis de glicerol 3-fosfato foram muito mais elevados em pacientes com COVID-19 do que em voluntários saudáveis. É importante destacar que o glicerol 3-fosfato é um metabolito importante no metabolismo glicídico e lipídico, e participa ativamente da regulação da imunidade adquirida pelo organismo em resposta a infecções virais [65]. No COVID-19, alterações nos níveis de glicerol 3-fosfato têm sido usadas como um indicador de dano imunológico causado pelo vírus e como um biomarcador associado à gravidade da doença ^{484,485}. Outra molécula que estava em níveis mais elevados em pacientes com COVID-19 do que em pacientes saudáveis era o sulfato de cisteína-S. Sabe-se que o cisteína-S-sulfato é um metabolito endógeno envolvido em processos inflamatórios ⁴⁸⁶. O papel do cisteína-S-sulfato no COVID-19 é pouco conhecido, porém há um estudo recente citando a participação da cisteína-S -sulfato nos processos inflamatórios do COVID-19, e os

autores propõem esta molécula como um possível biomarcador que deve ser considerado na patogênese e diagnóstico do COVID-19 ⁴⁸⁷.

Identificamos várias moléculas de acil-carnitina (dodecenoilcarnitina, tetradecenoilcarnitina, tetradecadienoilcarnitina, palmitoilcarnitina e malonilcarnitina) que estavam em níveis mais elevados em pacientes graves com COVID-19 do que em voluntários saudáveis ou levemente doentes. Resultados semelhantes foram encontrados no estudo recente de Lauro (2022) [70]. O aumento dos níveis de acil-carnitina pode indicar um aumento acentuado no transporte de lipídios para as mitocôndrias para o processo de beta-oxidação produzir energia para as células, sendo uma resposta à baixa produção de energia dos processos anaeróbios resultantes da disfunção respiratória causada pela COVID-19 [51]. Além dos biomarcadores acima mencionados, também identificamos pela primeira vez alguns biomarcadores que estão associados ao diagnóstico e gravidade da Covid-19 nomeadamente: Ácido N-Acetil-4-O-acetilneuramínico, N-Acetil-L-Alanina, N-Acetiltryptofano, palmitoilcarnitina e 1-miristato de glicerol.

Como limitações do nosso estudo, podemos citar que embora o estudo tenha sido realizado envolvendo amostras de plasma e soro de pacientes de diversos países (China, Spina, Itália e França), reconhecemos que o constante surgimento de novas variantes do vírus SARS -CoV-2 pode alterar o metaboloma, proteoma e lipoma de pacientes com COVID-19 e conseqüentemente o surgimento de novos biomarcadores associados a estas novas variantes. Estudos metabólicos de coorte longitudinais de acompanhamento de longo prazo monitorando esses pacientes são necessários.

5.7 CONCLUSÃO

Neste estudo, sete algoritmos diferentes baseados em *ML* (*PLS-DA*, *KNN*, *XGboost*, *SVM*, *ANN*, *SIMCA* e *LREG*) foram construídos para prever o diagnóstico, gravidade e letalidade do COVID-19 usando dois bancos de dados diferentes. O modelo *PLS-DA* apresentou o melhor desempenho, com precisão de aproximadamente 93%. Este modelo pode ajudar no diagnóstico precoce da COVID-19 e orientar a gestão da doença com intervenções adicionais adaptadas à prática diária. Finalmente, alguns dos biomarcadores associados ao diagnóstico e prognóstico de COVID-19 encontrados no conjunto de amostras do nosso estudo (ou seja, 5,6-di-hidro-5-metiluracil, cisteinilglicina, ribotimidina, esfingosina 1-fosfato, ciclohexilamina, uridina e ornitina) já foram citados anteriormente na literatura científica, o que reforça seu papel na infecção. Por outro lado, relatamos pela primeira vez biomarcadores adicionais (ou seja, N-acetil-glucosamina-1-fosfato e 4-hidroxifenilacetilcarnitina) que devem ser avaliados posteriormente como indicadores prognósticos de COVID-19.

Concluindo, neste estudo, modelos de *machine learning* (*PCA* e *PLS-DA*) foram usados para analisar vários conjuntos de dados multiômicos de pacientes com COVID-19 de diferentes países (China, Espanha, Itália e França) para identificar novos biomarcadores associados à rápida diagnóstico e prognóstico da COVID-19, visando prevenir sintomas a longo prazo. Todos os modelos *PLS-DA* apresentaram desempenho de predição diagnóstica e prognóstica em torno de 90%, assim como os resultados de RT-PCR. Um total de 23 biomarcadores foram identificados como associados ao diagnóstico e prognóstico de COVID-19 (por exemplo, espermidina, taurina, L-aspártico, L-glutâmico, L-fenilalanina e xantina, N-acetil glucosamina-1-fosfato, ornitina e ribotimidina). Além disso, também relatamos pela primeira vez novos biomarcadores (ácido N-acetil-4-O-acetilneuramínico, N-acetil-L-alanina, N-acetiltryptofano, palmitoilcarnitina e 1-miristato de glicerol) que devem ser estudados com vista a serem utilizados na prática clínica como indicadores de diagnóstico precoce e prognóstico da COVID-19, permitindo intervenção rápida e prevenindo o desenvolvimento de COVID longa.

6 CAPÍTULO VI - ABORDAGEM INTEGRATIVA NA BUSCA POR INIBIDORES DA PROTEÍNA SPIKE (RBD) DO SARS-COV-2: TRIAGEM VIRTUAL, ANÁLISE DE DRUG-LIKENESS, PREDIÇÕES ADMET, DINÂMICA MOLECULAR E INTELIGÊNCIA ARTIFICIAL PARA O TRATAMENTO DA COVID-19

Publicado em:

1. Cobre AF, Maia Neto M, de Melo EB, Fachi MM, Ferreira LM, Tonin FS, Pontarolo R. Naringenin-4' glucuronide as a new drug candidate against the COVID-19 Omicron variant: a study based on molecular docking, molecular dynamics, MM/PBSA and MM/GBSA. *J Biomol Struct Dyn.* 2023 Jul 2:1-14. doi: 10.1080/07391102.2023.2229446.

2. COBRE, Alexandre de F. et al. Machine learning-based virtual screening, molecular docking, drug-likeness, pharmacokinetics and toxicity analyses to identify new natural inhibitors of the glycoprotein spike (s1) of SARS-CoV-2. *Química Nova*, v. 46, p. 450-459, 2023. doi: 10.21577/0100-4042.20230038

6.1 RESUMO

Introdução: Este capítulo VI da tese teve como objetivo identificar compostos bioativos naturais (NBCs) como potenciais inibidores da região de domínio de ligação (RBD) da proteína *Spike* (S1) do vírus SARS-CoV-2 tipo selvagem e da mutação ômicron, por meio de simulações computacionais (*in silico*). Material e métodos: compostos naturais bioativos foram selecionados do banco de dados ZINC e avaliados através de triagem virtual e docking molecular para identificar aqueles com maior afinidade pela região de domínio de ligação (RBD) da proteína *Spike* do vírus SARS-CoV-2. Em seguida, foram analisados quanto à sua adequação farmacológica, avaliando parâmetros de *drug-likeness* e ADMET. Os selecionados foram submetidos a simulações de dinâmica molecular, incluindo o cálculo das energias livres de ligação, utilizando antivirais (remdesivir e molnupinavir) utilizados atualmente como referência. Além disso, análises de QSAR foram realizadas utilizando algoritmos de inteligência artificial como *Random Forest*, *XGBoosting* e *Histogram-based Gradient Boosting*, para prever a bioatividade anti-SARS-CoV-2 dos compostos selecionados.

Resultados: Um total de 170.906 compostos foi analisado por virtual screening e docking molecular. Para a RBD do tipo selvagem, apenas NBC5 (feselol), NBC14, NBC15 e NBC27 apresentaram simultaneamente afinidade de ligação pela RBD e resultados ADMET favoráveis. Feselol e três outros NBCs foram identificados como candidatos promissores para o tratamento da COVID-19. Entretanto, análises de QSAR (*machine learning*) revelaram que apenas feselol demonstrou atividade anti-SARS-CoV-2 significativa. Para a proteína *Spike* (RBD) da mutação ômicron, os principais NBCs com alta afinidade foram identificados. O ligante ZINC000045789238 (naringenina-4'-O glicuronídeo) apresentou bioatividade significativa e uma ligação favorável, com destaque para a formação de múltiplas ligações de hidrogênio. Os descritores moleculares relacionados ao aumento da bioatividade anti-SARS-CoV-2 foram identificados por meio da metodologia *SHAP values*. **Conclusão:** Os resultados destacam feselol como um potencial inibidor da proteína *Spike* para o vírus tipo selvagem, enquanto o naringenina-4'-O glicuronídeo emerge como uma promissora opção para a variante mutante ômicron. Esses achados destacam a importância da investigação de compostos naturais bioativos como potenciais agentes terapêuticos contra a COVID-19 e suas variantes.

Palavras-chave: SARS-CoV-2, compostos bioativos naturais, docking molecular, ADMET, dinâmica molecular, QSAR, *Machine learning*.

6.2 INTRODUÇÃO

A pandemia global de COVID-19, causada pelo coronavírus SARS-CoV-2, instaurou uma urgência sem precedentes na busca por terapias eficazes e específicas para combater essa doença viral altamente contagiosa ⁴⁸⁸. A proteína *Spike (RBD)* do SARS-CoV-2 desempenha um papel crucial na infecção, facilitando a ligação do vírus às células hospedeiras humanas por meio da interação com o receptor da enzima conversora de angiotensina 2 (ECA-2) ⁴⁸⁹.

Baseando-se nisso, o presente capítulo se dedica à exploração de estratégias inovadoras na busca por potenciais inibidores da proteína *Spike (RBD)* do SARS-CoV-2, abordando tanto a variante selvagem quanto a mutação Ômicron ⁴⁹⁰. A crescente complexidade das variantes do vírus representa um desafio crucial na busca por tratamentos eficazes contra a COVID-19 ⁴⁹¹. Neste contexto, este estudo adota uma abordagem integrativa, amalgamando técnicas avançadas como triagem virtual via docking molecular, análise de *drug-likeness*, previsões farmacocinéticas e de toxicidade, dinâmica molecular e inteligência artificial (QSAR) ⁴⁹². O objetivo primordial é identificar compostos com potencial terapêutico, considerando não apenas sua capacidade de interação com a proteína-alvo, mas também suas propriedades farmacológicas e cinéticas.

A interdisciplinaridade dessas técnicas permite uma análise abrangente e criteriosa, transcendendo a mera identificação de interações moleculares. Ao integrar dados provenientes de diferentes abordagens, busca-se compreender não apenas a eficácia dos candidatos a inibidores, mas também sua viabilidade para o desenvolvimento clínico. Este estudo busca, portanto, contribuir significativamente para o arsenal terapêutico contra a COVID-19, fornecendo informações valiosas para o desenho e a implementação de tratamentos eficazes e seguros.

6.3 OBJETIVOS

6.3.1 *Objetivo geral:*

- Investigar e identificar potenciais inibidores da proteína Spike (RBD) do SARS-CoV-2, tanto para a variante selvagem quanto para a mutação Ômicron, utilizando uma abordagem integrativa que combina triagem virtual via docking molecular, análise de *drug-likeness*, previsões farmacocinéticas e de toxicidade, dinâmica molecular e inteligência artificial (QSAR), visando contribuir para o desenvolvimento de tratamentos eficazes contra a COVID-19.

6.3.2 *Objetivos específicos:*

- Realizar triagem virtual utilizando técnicas de docking molecular para identificar compostos com potencial ligação à proteína Spike (RBD) do SARS-CoV-2, considerando tanto a variante selvagem quanto a mutação Ômicron;
- Avaliar a adequação dos compostos identificados por meio da análise de *drug-likeness*, garantindo propriedades farmacológicas desejáveis;
- Realizar previsões farmacocinéticas e de toxicidade dos compostos selecionados para avaliar sua viabilidade como candidatos a fármacos;
- Aplicar técnicas de dinâmica molecular para compreender as interações dinâmicas entre os compostos candidatos e a proteína alvo, a fim de validar e aprimorar as previsões iniciais;
- Desenvolver modelos de inteligência artificial (QSAR) para prever a atividade biológica dos candidatos a inibidores da proteína *Spike (RBD)*, melhorando a compreensão do seu potencial terapêutico;
- Integrar os resultados das diferentes abordagens utilizadas para identificar os principais compostos candidatos com potencial terapêutico contra a COVID-19, fornecendo informações valiosas para pesquisas futuras e desenvolvimento de fármacos.

6.4 MATERIAL E MÉTODOS

6.4.1 Fluxograma do estudo

O fluxograma do estudo empregado para a realização deste capítulo VI é mostrado na Figura 6.1.

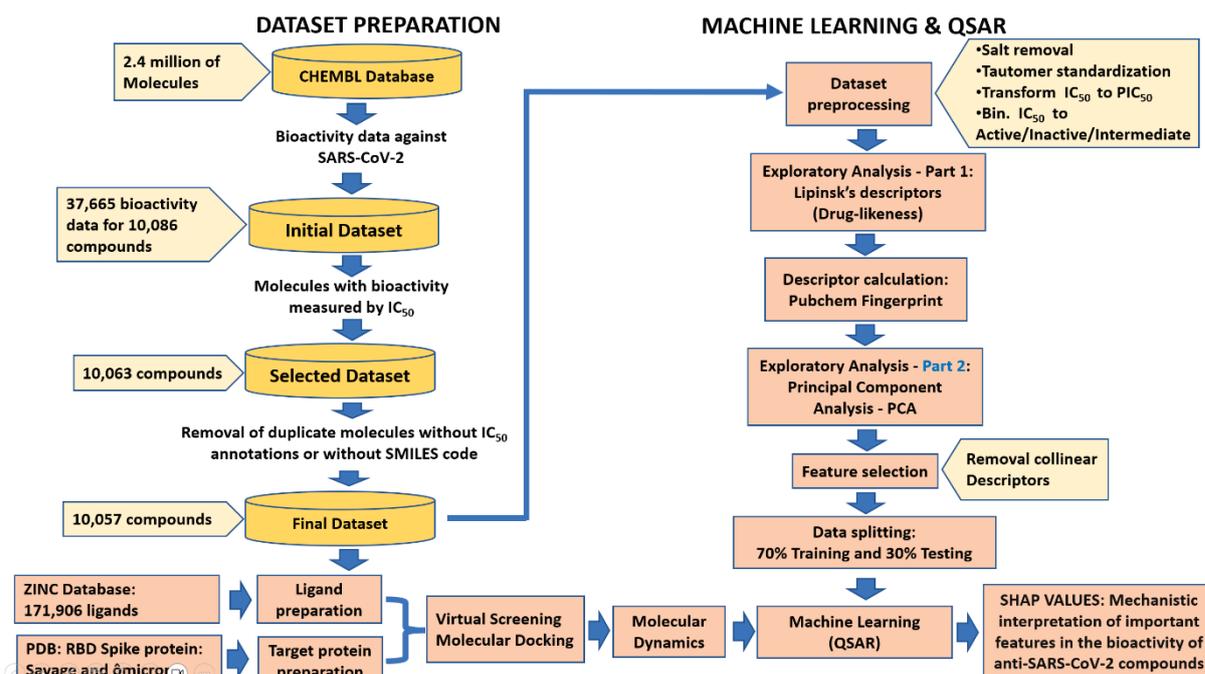


Figura 6.1. Fluxograma do estudo. Fonte: O Autor (2024)

6.4.2 Seleção de proteína Spike (RBD): variante selvagem e Ômicron

A estrutura da glicoproteína *Spike* (S1) selvagem (Wugan, China) determinada por microscopia eletrônica) e da mutação ômicron (África do Sul) foram obtidas do banco de dados público *Research Collaboratory for Structural Bioinformatics* (RCSB PDB), onde estão disponíveis estruturas tridimensionais de macromoléculas, incluindo todas as proteínas SARS-CoV-2⁴⁹³. O PDB auxilia na aquisição de estruturas 3D de alvos terapêuticos, o que permite a realização de estudos de triagem virtual baseadas na estrutura do ligante (SBVS)⁴⁹³. O alvo S1 foi selecionado devido ao seu fragmento RBD (região de domínio de ligação) que interage com o ECA-2 humana, responsável por o reconhecimento e a penetração do vírus nas células pulmonares⁴⁹⁴.

Para a identificação da proteína *Spike* (RBD) do SARS-CoV-2 tipos selvagem e da mutação ômicron, foi realizada uma busca sistemática de estruturas

cristalográficas no PDB (Figura 6.2). As estruturas cristalográficas da proteína *Spike* (*RBD*) selvagem e ômicron foram avaliadas considerando os seguintes critérios: (i) o resultado da classificação percentual de validação, (ii) disponibilidade de informações sobre o *RBD* com *ACE-2* humana, (iii) a presença de uma sequência completa de aminoácidos e (iv) disponibilidade do artigo científico publicado.

Inicialmente foram identificados um total de 7287 e 4630 registros da proteína *Spike* selvagem e ômicron, Respectivamente. Após a triagem dos registros foram selecionados os códigos: PDB ID 6M17⁴⁹⁴ para a proteína *Spike* tipo selvagem e PDB ID foi 7T9L⁴⁹⁵ para a proteína *Spike* da variante ômicron (Figura 6.2). Essas duas estruturas cristalográficas foram usadas como alvos nas triagens virtuais do presente capítulo.

6.4.3 Preparo da proteína alvo

A estrutura 3D da glicoproteína *Spike* (S1) tipo selvagem (PDB ID 6M17)⁴⁹⁴ e da variante Ômicron (PDB ID 7T9L)⁴⁹⁵ no formato PDB foi preparada no software *AutodockTools* (ADT), onde (i) todas as moléculas de água de cristalização foram removidas para evitar impedimentos estéricos no momento do acoplamento com os ligantes; (ii) foram adicionados os hidrogênios polares; e (iii) foram incluídas as cargas de Kollman⁴⁹⁶. Finalmente, investigamos o estado de protonação dos aminoácidos tituláveis da proteína *Spike* (S1) *RBD* em pH 7,4 usando o software *ProPka 3.0*^{497,498}. Após a preparação, o arquivo da molécula foi convertido para o formato PDBQT no ADT. O software *Discovery Studio Visualizer* (DSV) foi usado para visualizar a estrutura da proteína preparada⁴⁹⁹.

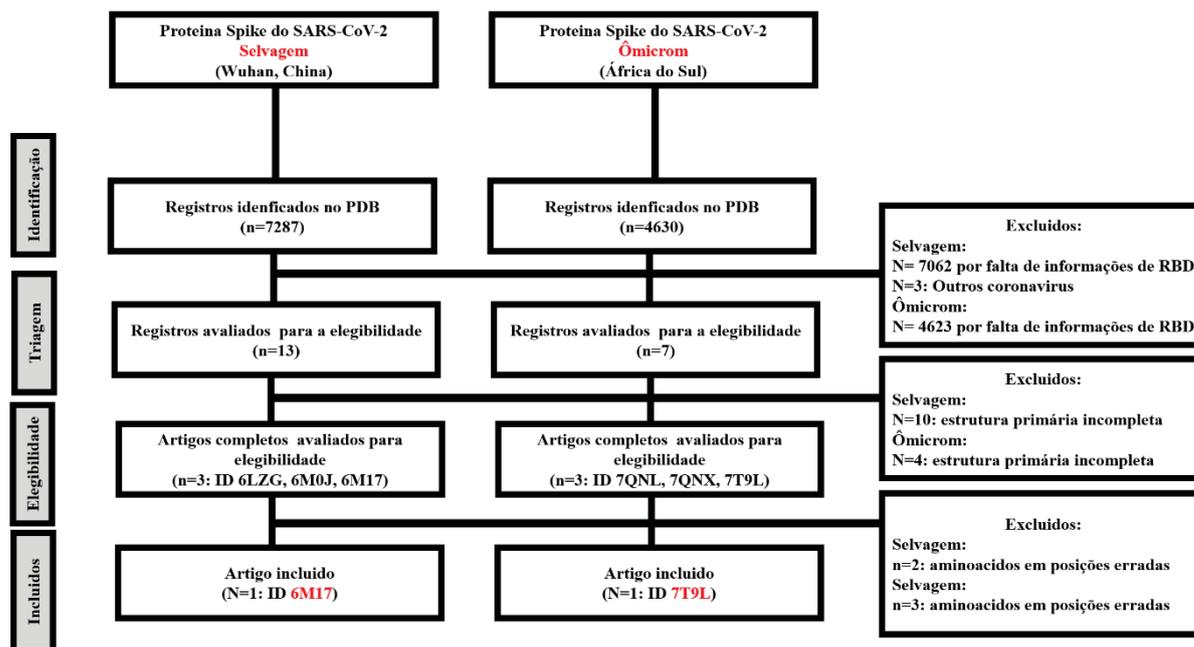


Figura 6.2. Identificação da proteína *Spike* do tipo selvagem (PDB ID 6M17) e da mutação ômicron (PDB ID 7T9L) do vírus SARS-CoV-2. Identificação dos aminoácidos da região de domínio de ligação da proteína *Spike* do SARS-CoV-2. **Fonte:** O Autor (2024).

Após a preparação dos alvos moleculares das variantes selvagem (PDB ID 6M17)⁴⁹⁹ e ômicron (PDB ID 7T9L)⁵⁰⁰, foram lidos atentamente os artigos científicos originais das estruturas do alvo molecular, visando identificar quais resíduos de aminoácidos do RBD da proteína *Spike* interagem com o receptor *ACE-2* humano. Na etapa seguinte, a sequência dos resíduos de aminoácidos foi salva em formato FASTA e identificados os resíduos mutantes.

6.4.4 Coleta e preparo dos ligantes

Um total de 170.906 moléculas bioativas naturais disponíveis comercialmente foram obtidas do banco de dados ZINC15⁵⁰¹. A base de dados ZINC foi escolhida porque é a única base de dados com filtros para selecionar apenas compostos disponíveis comercialmente e é uma base de dados muito utilizada na descoberta de medicamentos por grandes indústrias farmacêuticas⁵⁰¹. Para a obtenção desses ligantes foram usados os seguintes filtros: *substances/for sale/in vitro/biogenic*. Todos os 71 mil compostos selecionados possuem dados de biotividade *in vitro* testada contra outras doenças. Na sequência, foram obtidas as suas estruturas 3D em um

arquivo no formato mol2, que é o formato reconhecido pela maioria dos softwares de estudo in silico.

Após a sua obtenção, todos os ligantes foram preparados no software *PyRx virtual screening tools*, e o preparo consistiu em dois passos ⁵⁰². O primeiro passo foi a minimização da energia de cada ligante utilizando os seguintes parâmetros: (i) campo de força: UFF; (ii) algoritmo de otimização: gradientes conjugados; (iii) o número total de etapas: 200; (iv) número de etapas para atualização: 1; (v) pare se a diferença de energia for menor que 0,1. Dessa forma, suas conformações mais estáveis foram obtidas e puderam ser usadas para interagir com a proteína alvo. Também atribuímos cargas parciais de Gasteiger a todos os ligantes usando o software *AutoDock Raccoon* ⁵⁰³. Depois disso, foram corrigidos os estados de protonação de todos os ligantes para o pH 7,4 usando o software Open Babel ⁵⁰⁴.

O segundo passo foi a conversão do formato dos arquivos do ligante de 'mol2' para o formato PDBQT, utilizando o software *AutodockVina*, que é acoplado ao software de ferramentas de triagem virtual *PyRx* ⁵⁰².

6.4.5 Triagem virtual e docking molecular

Antes de realizar a triagem virtual no software *PyRx virtual screening tools*, inicialmente foram ajustadas as coordenadas do centro X, Y, Z e o tamanho (angstrom) de X, Y, Z da grid-box e a exaustividade no *Autodock tools* e *Autodock Vina*. Os ligantes mais promissores foram aqueles que apresentavam valores mais altos (maior afinidade) de energia de ligação com proteínas RBD tanto da variante selvagem quanto da mutação ômicron ⁵⁰⁵. Os ligantes com resultados mais favoráveis durante as análises de triagem virtual foram submetidos a análise de docking molecular utilizando o software *Autodock Tools 4.2.0* e *Autodock Vina* ^{496,506}. A razão para utilizar dois softwares diferentes na análise de docking (ferramentas *Autodock 4.2.0* e *AutodockVina*) foi para avaliar o desempenho e a precisão dos resultados de docking molecular ^{507,508}. Os complexos ligante-proteína com maior afinidade de ligação ($\Delta G > -7$ kcal/mol) foram considerados para análises posteriores. O valor de corte da energia de afinidade foi baseado na literatura ^{509,510}. Os medicamentos remdesivir e molnupinavir foram utilizados como referência para comparar os resultados do docking molecular e da triagem virtual entre nossos ligantes com o RBD do SARS-CoV-2 do tipo selvagem e da variante ômicron, respectivamente.

A triagem virtual e o docking molecular foram realizados utilizando os seguintes parâmetros: (i) exaustividade de 16; (ii) as coordenadas do centro da caixa de grade foram otimizadas em $x = 228,052$, $y = 172,488$ e $z = 253,630$; e (iii) o tamanho da caixa de grade foi otimizado em $x = 50 \text{ \AA}$, $y = 70 \text{ \AA}$, $z = 50 \text{ \AA}$. É essencial destacar que o tamanho da nossa *gridbox* é semelhante ao grid-box do recente estudo de docking publicado por Liu (2023) (Liu, 2023), que também utilizou a mesma estrutura cristalográfica da proteína Spike usada em nosso estudo (ID do PDB: 7T9L). Na modelagem de docking, a busca pelas conformações (poses) dos ligantes com maior estabilidade de ligação com a proteína *Spike (RBD)* foi realizada utilizando uma função de pontuação denominada soma de energias parametrizadas e uma função de busca aleatória denominada algoritmo genético. Nesta análise, os resultados da energia de ligação da proteína e do ligante foram comparados com os valores do desvio quadrático médio (RMSD) para validar os resultados obtidos nas análises de triagem virtual ⁴⁹⁶.

6.4.6 Avaliação da influência do estado estereoisomérico, tautomérico e estado de protonação em pH fisiológico dos ligantes na sua afinidade pela RBD

Considerando que os estados estereoisomérico e tautômero do fármaco e seu estado de protonação em pH fisiológico ($\text{pH} = 7,4$) são fundamentais para a afinidade de ligação às proteínas, foi realizado um estudo aprofundado nos ligantes que apresentavam maior afinidade pela proteína *Spike (RBD)* do tipo selvagem e da variante ômicron obtidos a partir das análises da triagem virtual ^{511,512}. As estruturas 3D das moléculas no estado protonado em pH fisiológico e de todos os estereoisômeros e tautômeros foram previstas automaticamente usando o software *Marvin Sketch*. Todas as estruturas químicas convertidas também foram analisadas por docking molecular utilizando o *AutoDock Tools 2.4.0*, e os resultados foram comparados com os do docking da molécula líder.

6.4.7 Análise de drug-likeness e farmacocinética (ADME)

Para os ligantes que apresentaram maior afinidade de ligação pela proteína *Spike (RBD)* foram realizadas previsões das características de semelhança ao fármaco (*drug-likeness*) e dos parâmetros farmacocinéticos [absorção, distribuição,

metabolismo, excreção (ADME)] utilizando a plataforma online denominada *ADMETLAB2*, que utiliza um algoritmo de *machine learning* (redes neurais artificiais ou aprendizado profundo) com precisão > 85%⁵¹³. As previsões de semelhança com medicamentos foram realizadas usando produtos farmacêuticos da Pfizer e regras do triângulo dourado. Pela Metodologia da Pfizer, compostos com alto LogP (>3) e baixo TPSA. Os ligantes *drug-like* e que tiveram melhores previsões ADME foram selecionados para a etapa subsequente, a análise de toxicidade.

6.4.8 Predição da toxicidade aguda e crônica

Os ligantes que mostrarem resultados promissores na etapa anterior (*drug-likeness* e ADME) também tiveram suas toxicidades agudas e crônicas previstas usando a plataforma online *PROTOX II*. Foram previstos quatro grupos diferentes de toxicidade: (i) Toxicidade de órgãos: Hepatotoxicidade; (ii) toxicidades endpoints: carcinogenicidade, mutagenicidade e citotoxicidade; (iii) Vias de sinalização do receptor nuclear Tox21: receptor de aril hidrocarboneto (AhR); Receptor de Andrógeno (AR); Domínio de Ligação ao Ligante do Receptor de Andrógeno (AR-LBD); Aromatase, Receptor Alfa de Estrogênio (ER), Domínio de Ligação ao Ligante do Receptor de Estrogênio (ER-LBD), Receptor Gama Ativado por Proliferador de Peroxissoma (PPAR-Gama); (iv) Vias de resposta ao estresse Tox21: fator nuclear (derivado de eritróide 2)-like 2/elemento responsivo a antioxidante (nrf2/ARE); Elemento de resposta do fator de choque térmico (HSE), potencial de membrana mitocondrial (MMP), fosfoproteína (supressor de tumor) p53 e domínio AAA da família ATPase contendo proteína 5 (ATAD5). Os ligantes que mostraram resultados promissores da análise de toxicidade foram selecionados para a análise posterior, as simulações de dinâmica molecular.

6.4.9 Simulações de dinâmica molecular

Os complexos ligante-proteína que apresentaram resultados promissores nas análises de docking, *drug-likeness* e previsões ADMET foram submetidos a simulações de dinâmica molecular (DM) com o uso do software *GROMACS 2020.3*⁵¹⁴ e campo de força CHARMM36 (*Chemistry at Harvard Macromolecular Mechanics*)⁵¹⁵⁻⁵¹⁷. Podemos descrever a DM empregada por meio de 5 etapas a saber:

Preparação dos componentes do sistema, Minimização da energia, Equilibração para as condições físicas desejadas, Produção da dinâmica e Análise dos dados. Os medicamentos molnupinavir e remdesivir foram também usados como referência para comparar os resultados das simulações de dinâmica molecular entre nossos ligantes com o *RBD* do SARS-CoV-2.

A etapa de preparação iniciou-se com os ligantes e a proteína *Spike* sendo separados para que em seguida fosse realizada a parametrização dos ligantes através do programa CGenFF⁵¹⁸. Ligantes e proteína *Spike*, já novamente complexados e no formato de entrada GRO, foram inseridos em um caixa cúbica com distância mínima de 1 nm das bordas. Utilizou-se o modelo de representação da água TIP3P e o modelo spc216 para preenchimento desta caixa com o solvente. Íons cloreto (Cl⁻) e sódio (Na⁺), em concentrações fisiológicas de 0,15 mol/L, foram adicionados em quantidade suficiente para neutralizar as cargas residuais. Na etapa de Minimização foi aplicado o algoritmo de otimização *Steepest Descents* até que a energia potencial do sistema seja negativa e da ordem de grandeza de 10⁶ a 10⁵⁵¹⁹.

Todos os sistemas foram submetidos a dois estágios de equilibração dinâmica por um período de 1 ns, o primeiro em condições de número de mols, volume e temperatura constantes (NVT) e o segundo em condições de número de mols, pressão e temperatura constantes (NPT). O constrangimento das vibrações holomônicas das ligações de hidrogênio foi realizado pelo método linear *constraint solver* (*LINC*). Os demais parâmetros aplicados na equilibração foram: Intervalo de integração de 2 fs, Integrador *Leap-Frog*, Temperatura de 310 K para os dois grupos de acoplamento (proteína-ligante/Na-Cl-Solvente), condições de contorno periódicas (3D-PCB), interpolação cúbica e 1,2 nm de corte de interação eletrostática PME (*Particle Mesh Ewald*)⁵²⁰.

Por fim, simulações de DM para a etapa de produção foram realizadas para 100 ns, onde foram avaliadas as seguintes métricas de DM: RMSD, flutuação quadrática média (RMSF), ligação de hidrogênio, raio de giração (Rg), área superficial acessível ao solvente (SASA), temperatura, pressão, densidade e potencial⁵²¹⁻⁵²³. É fundamental ressaltar que nas simulações de DM (incluindo o cálculo das energias livres de MM/PBSA e MM/GBSA) foram utilizados 2.000 frames. Os resultados de DM foram plotados no software *Graph Prism*. O software VMD (dinâmica molecular visual) foi utilizado para visualizar os resultados das simulações de DM⁵²⁴.

6.4.10 Cálculo de energia livre de ligação MM/GBSA

Os cálculos de ligação MM/GBSA (*Molecular mechanics/generalized Born surface area*) do complexo proteína-ligante foram realizados usando o *gmx-mmpbsa* no software *GROMACS* usando as equações 1, 2 e 3 ⁵²⁵.

$$\Delta G_{\text{bind}} = \Delta H - T\Delta S = \Delta E_{\text{MM}} - T\Delta S + \Delta G_{\text{sol}} \quad (\text{Eq. 1})$$

$$\Delta G_{\text{MM}} = \Delta E_{\text{vdw}} = \Delta E_{\text{interna}} + \Delta E_{\text{ele}} \quad (\text{Eq. 2})$$

$$\Delta G_{\text{sol}} = \Delta G_{\text{SA}} + \Delta G_{\text{GB}} \quad (\text{Eq. 3})$$

Onde: $T\Delta S$, ΔE_{MM} e ΔG_{sol} são, respectivamente, a entropia conformacional, a energia MM da fase gasosa e a energia livre de solvatação (a soma da contribuição apolar ΔG_{SA} e a contribuição polar ΔG_{GB}). ΔE_{MM} contém energias diédricas, eletrostáticas ΔE_{ele} e $\Delta E_{\text{interna}}$ da ligação, energia de Van Der Waals ΔE_{vdw} e ângulo. O cálculo da entropia pode ser omitido se nenhuma mudança estrutural for causada por ligações no processo de simulação DM.

6.4.11 Cálculo de energia livre de ligação MM/PBSA

O método MM/PBSA (*Molecular mechanics/Poisson–Boltzmann surface area*) foi utilizado para determinar a energia livre de ligação das trajetórias de simulação de DM ⁵²⁶. Os principais parâmetros que controlam o cálculo MM/PBSA incluem (i) tensão superficial para estimar a energia de solvatação apolar usando a área superficial avaliável do solvente, 0,054; (ii) constante dielétrica interna, 1, e constante dielétrica externa, 80. As equações 4 e 5 abaixo foram usadas para calcular o MM/PBSA ⁵²⁶.

$$\Delta G_{\text{bind}} = \Delta G_{\text{complex (minimizado)}} - [\Delta G_{\text{ligand (minimizado)}} + \Delta G_{\text{receptor (minimizado)}}] \quad (\text{Eq. 4})$$

$$\Delta G_{\text{bind}} = \Delta G_{\text{MM}} + \Delta G_{\text{PB}} + \Delta G_{\text{SA}} - T\Delta S \quad (\text{Eq. 5})$$

onde: $T\Delta S$ = contribuição de entropia; ΔG_{MM} = soma entre as interações eletrostática e de *Vander Waals*; ΔG_{SA} = energia de solvatação apolar; ΔG_{PB} = energia de solvatação polar.

6.4.11 Simulações de pós – dinâmica molecular

Para um estudo mais profundado da dinâmica do comportamento dinâmico do complexo formado entre o ligante e a proteína RBD da variante ômicron, foram realizadas novas simulações pós-dinâmica moleculares, onde foram calculadas as seguintes métricas: a distância (Å) do centro de massa por quadro do complexo proteína-ligante por frame e as energias livres de ligação (*MM/PBSA* e *MM /GBSA*) por frame. É fundamental destacar que os cálculos pós-DM foram realizados utilizando um tempo de 20 nanosegundos, sendo este tempo cuidadosamente escolhido de acordo com estudos científicos previamente publicados na literatura^{527,528}.

6.4.12 Desenvolvimento de modelos de Inteligência artificial e machine learning baseados em QSAR-3D

A Figura 6.1 representa o fluxo de execução desta etapa do trabalho. Em resumo, elaboramos modelos de *machine learning* baseados em QSAR-3D com o objetivo de prever a bioatividade anti-SARS-CoV-2 de compostos que apresentaram resultados promissores em análises de docking molecular, dinâmica molecular e previsões ADMET. O estudo aderiu ao guia da Organização para a Cooperação e Desenvolvimento Econômico (OCDE), abrangendo as seguintes etapas: (i) compilação de um conjunto de dados com um ponto final; (ii) análise exploratória desses dados; (iii) aplicação de diferentes algoritmos de *machine learning* supervisionado; (iv) avaliação do desempenho dos modelos de *machine learning* utilizando métricas específicas; (v) interpretação mecanística dos modelos, incluindo a análise dos descritores de Lipinski; (vi) aplicação dos algoritmos em um banco de dados externo.

6.4.12.1 Descrição do banco de dados utilizado para Machine learning: ChEMBL Database

Neste estudo, foram utilizados conjuntos de dados do banco *ChEMBL* 32 (<https://www.ebi.ac.uk/chembl/>) para desenvolver modelos de *machine learning* visando prever compostos com atividade contra o vírus SARS-CoV-2. O *ChEMBL* é um banco de dados público do Reino Unido, contendo informações de bioatividade de mais de 2,4 milhões de compostos, similares a medicamentos, distribuídos em 211 conjuntos de dados. Ele integra dados químicos, de bioatividade e genômicos, facilitando a tradução de informações genômicas em possíveis medicamentos. Além disso, inclui cerca de 86.000 publicações científicas, 1,5 milhão de ensaios, dados de 15.000 alvos terapêuticos, 6.700 mecanismos de ação de medicamentos, 2.000 tipos de células, 43.000 indicações de medicamentos e 759 tecidos biológicos. Todos esses dados passam por seleção manual e curadoria por especialistas.

O conjunto experimental do *ChEMBL* consistiu em 10.086 compostos, cada um com 37.665 pontos de dados de bioatividade anti-SARS-CoV-2. Os parâmetros incluíam EC50, MIC, porcentagem de atividade, Ki, porcentagem de inibição, IC50, entre outros. Optou-se por investigar o IC50 (medido em nM), pois estava disponível para a maioria dos compostos (10.063). Após a remoção de compostos duplicados, sem anotações IC50 ou sem código *SMILES*, restaram 10.057 compostos para a próxima etapa: o pré-processamento dos dados. Todos os códigos *SMILES* dos compostos foram tratados usando *Python*.

6.4.12.2 Pré-processamento dos dados para machine learning

O pré-processamento do conjunto de dados final (n=10,063) de compostos bioativos contra o vírus SARS-CoV-2 foi realizado utilizando a biblioteca *Padelpy* na linguagem *Python* (disponível em <https://pypi.org/project/padelpy/>). Esse processo envolveu a remoção de sais, a padronização de tautômeros e a conversão dos valores de IC50 em pIC50 (escala logarítmica). Além disso, os valores de IC50 foram categorizados em três grupos de compostos: compostos ativos (IC50 < 100 nM), compostos com atividade intermediária (IC50 entre 100-1000 nM) e compostos inativos (IC50 > 1000 nM)⁵²⁹. Após esta etapa, foram calculados os descritores de impressão digital (usando descritores *PubChem*) das moléculas. Eles consistem em

um conjunto de códigos binários 881 que descrevem a impressão digital e o espaço químico 3D de cada molécula específica. Existem 12 tipos diferentes de descritores de impressão digital (por exemplo, *PubChem*, CDK, subestrutura etc.)⁵³⁰. Neste estudo, a impressão digital das moléculas foi calculada utilizando o descritor *PubChem*, que é o mais utilizado para estudos de *machine learning* baseados em QSAR⁵³¹.

6.4.12.3 Análise exploratória univariada do espaço químico: descritores de Lipinski

A Análise do espaço químico foi realizada usando os descritores da regra dos cinco de Lipinski. Esses descritores abrangem peso molecular (MW), contagem de aceptores de ligações de hidrogênio (nHBAcc), contagem de doadores de ligações de hidrogênio (nHBDon) e o logaritmo do coeficiente de partição octanol/água (AlogP). De acordo com as recomendações farmacêuticas da Pfizer, para que um composto químico possua propriedades semelhantes às de um medicamento, devem ser atendidos critérios específicos: $PM \leq 500$ g/mol; $AlogP < 5$; $nHBDon < 10$ e $nHBAcc < 5$. Para calcular esses descritores, foi utilizada a biblioteca RDKit em *Python* (<https://www.rdkit.org/>)⁵³².

6.4.12.4 Análise exploratória multivariada do espaço químico: Modelo PCA

A análise exploratória do espaço químico multivariado foi também realizada usando os descritores de *PubChem* via modelo PCA. Nesta análise, compostos ativos, inativos e aqueles com bioatividade intermediária foram discriminados usando o algoritmo de *machine learning*, PCA. O objetivo dessa análise foi investigar se existia diferenças significativas no espaço químico das três classes de bioatividade dos compostos, com vista a reunir *insights* para o desenvolvimento de modelos ainda mais complexos de predição. Todos os modelos PCA foi desenvolvendo seguindo os mesmos procedimentos descritos nos capítulos IV e V da presente tese de doutorado.

6.4.12.5 Seleção das variáveis

A multicolinearidade, uma preocupação significativa em modelos de regressão, refere-se à intercorrelação entre descritores (recursos), levando a um viés amplificado

e à complexidade do modelo no *machine learning*. Para resolver esse problema, um filtro de variância foi aplicado, removendo descritores moleculares (variáveis) com variabilidade limitada (variância $<0,1$) do conjunto de dados. O objetivo foi obter um subconjunto condensado de descritores *PubChem*. Todo esse processo foi implementado utilizando a linguagem de programação *Python* (Figura 6.1) ⁵³³.

6.4.12.6 Divisão de dados de treinamento e de teste

A divisão de dados para *machine learning* empregou o método *Holdout* para reduzir preconceitos na seleção de dados de treinamento, teste e validação (Figura 6.1). Este método segmentou o conjunto de dados de 10,063 compostos bioativos em duas fases: alocando 70% para treinamento de modelo e reservando 30% para fins de teste ⁵³⁴.

6.3.12.7 Machine learning

Todo o projeto, incluindo a fase de *machine learning*, foi executado em *Python*. Modelos de regressão foram implantados para prever a atividade biológica (expressa como pIC50) contra o SARS-CoV-2, com base na estrutura química dos 10,063 compostos bioativos representados pelos descritores *PubChem*. Nesta análise, a variável resposta foi pIC50, enquanto os descritores de impressão digital *PubChem* serviram como variáveis preditoras. As bibliotecas *Scikit-Learn* e *LazyPredict* facilitaram a construção de 40 algoritmos distintos de *machine learning*, com o objetivo de selecionar os cinco principais modelos que apresentam desempenho preditivo superior. A avaliação dos modelos de *machine learning* desenvolvidos contou com diversas métricas, incluindo coeficiente de determinação (R^2), erro quadrático médio (*MSE*), raiz do erro quadrático médio (*RMSE*) e erro médio absoluto (*MAE*) ^{535,536}. R^2 mede a capacidade preditiva do modelo de regressão numa escala de zero a um; mais próximo de um significa maior precisão preditiva. Por outro lado, *MAE*, *MSE* e *RMSE* são métricas de erro, onde um modelo de regressão bem ajustado visa minimizar seus valores ⁵³⁵.

6.4.12.7 Investigação dos descritores moleculares mais importantes na bioatividade anti-SARS-CoV-2 usando SHAP values.

A análise de recursos cruciais nos modelos de *machine learning* foi conduzida por meio do método de valores *SHAP* (*SHapley Additive exPlanations*). Inicializando com o objeto 'explicador', este método possibilitou o escrutínio de indivíduos ou grupos de compostos bioativos. A etapa subsequente envolveu o cálculo dos valores médios de cada recurso por meio da função *SoftMax*. A avaliação do impacto global dos descritores (recursos) de impressão digital *PubChem* foi alcançada por meio de várias visualizações, incluindo gráficos *SHAP*, gráficos *Beeswarm*, gráficos de resumo e gráficos de violino. Além disso, para discernir as partes específicas de uma molécula (características) que influenciam significativamente ou não a atividade biológica (pIC50), foram gerados gráficos de barras e gráficos em cascata^{537,538}.

6.5 RESULTADOS

6.5.1 Estudo I: Naringenina-4'-glicuronídeo como novo candidato a fármaco contra a variante Omicron da COVID-19: um estudo baseado em docking molecular, dinâmica molecular, MM/PBSA e MM/GBSA

6.5.1.1 Identificação dos aminoácidos mutantes da região de domínio de ligação da proteína Spike do SARS-CoV-2

Uma análise atenciosa das sequências *FASTA* dos aminoácidos da proteína *Spike* (*RBD*) do tipo selvagem (PDB ID 6M17) e da variante Ômicron (PDB ID 7T9L) permitiu identificar os seguintes resíduos de aminoácidos mutantes: Asp339, Leu371, Pro373, Phe375, Asn417, Lys440, Ser446, Asn477, Lys478, Ala484, Arg493, Ser496, Arg498, Tyr501 e His505. Destes, 6 resíduos estavam diretamente envolvidos na interação S1-ACE-2, nomeadamente, Asn477, Arg493, Ser496, Arg498, Tyr501 e His505. As análises de modelagem por docking molecular e simulações de dinâmica molecular têm como objetivo principal investigar a afinidade dos ligantes com os aminoácidos da *RBD*. No entanto, essas análises estiveram muito focadas nos quinze aminoácidos mutantes da variante ômicron que foram listados.

6.5.1.2 Triagem virtual e docking molecular: proteína Spike (RBD) da variante ômicron do vírus SARS-CoV-2

De um total de 71 mil ligantes, selecionamos os quatro principais ligantes com melhor afinidade de ligação pela proteína *Spike (RBD)* da variante ômicron do SARS-CoV-2 na triagem virtual (*PyRx* - triagem virtual) e das análises de docking molecular (ferramentas *AutoDock* e *AutoDockVina*): ID ZINC000045789238 (naringenina -4'-O-glicuronídeo) (Figura 6.3-E), ID ZINC000003995616 (ergolóide) (Figura 6.3-F), ID ZINC000004098448 (ohioensina A) (Figura 6.3-G) e ID ZINC000008662732 (prunetrina) (Figura 6.3-H) (Tabela 6.1).

Tabela 6.1. Triagem virtual e docking molecular dos quatro principais compostos naturais que tiveram maior afinidade pela proteína *Spike (RBD)* do SARS-CoV-2 da mutação ômicron.

Ligante	PyRx-virtual		AutoDock Tools		AutodockVina	
	screening					
ZINC ID	(kcal/mol)	RMSD	(kcal/mol)	RMSD	(kcal/mol)	RMSD
ZINC000004098448	-8.3	0.0	-8.9	1.1	-8.6	1.1
ZINC000045789238	-8.5	0.0	-8.8	1.3	-8.9	1.4
ZINC000008662732	-8.2	0.0	-8.7	1.1	-8.5	1.2
ZINC000003995616	-8.7	0.0	-8.3	1.4	-8.4	1.2

Nota: Todos os valores de energia de ligação e RMSD foram baseados na conformação mais estável do ligante após o acoplamento. As afinidades de ligação são expressas como as energias de ligação mais baixas obtidas em kcal/mol (<-7 kcal/mol). A triagem virtual foi obtida por *PyRx-AutodockVina*; O docking molecular foi obtido por *AutoDock1.5.6* e *AutodockVina*. Naringenina-4'-O-Glucuronídeo (ZINC000045789238), ergoloide (ZINC000003995616), ohioensina A (ZINC000004098448) e prunetrina (ZINC000008662732). **Fonte:** O Autor (2024).

Todos os quatro ligantes interagiram com os resíduos de aminoácidos da região *RBD* da variante Omicron, sendo o ligante naringenina-4'-O-glicuronídeo (Figura 6.3-E) aquele que estabeleceu o maior número de ligações de hidrogênio ($n = 5$) com os resíduos de aminoácidos da região *RBD* da mutação Omicron (Ser494, Ser496, Thr500, Thr5051 e Thr505). Prunetrina (Figura 6.3-H) foi o ligante com o segundo maior número de ligações de hidrogênio (Arg408, Asp417, Ser496 e His505), conforme representado na Figura 6.3.

Nove conformações foram obtidas nas análises de docking para cada um dos quatro ligantes selecionados. Apenas a conformação com maior estabilidade de ligação com o alvo molecular foi selecionada (ou seja, conformação com valor RMSD inferior a 2) (Tabela 6.1). Neste estudo, os coeficientes de variação das energias de ligação da proteína *Spike* (*RBD* da mutação ômicron com os quatro principais compostos promissores obtidos nos diferentes softwares (*PyRx*, *AutoDock Tools* e *AutoDockVina*) foram <2%, mostrando a confiabilidade e precisão das análises de docking molecular (Tabela 6.1). Na Figura 6.4 são mostrados os resultados de docking molecular de remdesivir (fármaco controle).

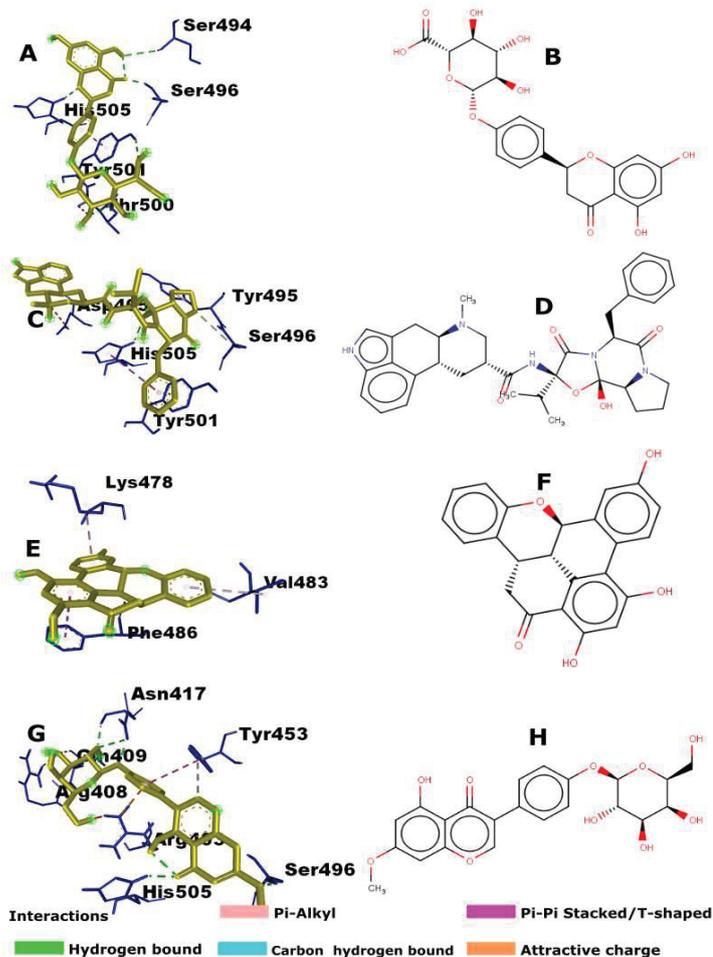


Figura 6.3. Modelagem de docking molecular dos quatro principais ligantes. O acoplamento dos ligantes naringenina-4'-O-Glucuronídeo (ZINC000045789238), ergolóide (ZINC000003995616), ohioensina A (ZINC000004098448) e prunetrina (ZINC000008662732) são mostrados nas Figuras A, C, E e G, respectivamente. Estruturas amarelas e azuis representam os ligantes e a proteína *Spike* (S1), respectivamente. As interações do complexo proteína-ligante *Spike* são mostradas em linhas tracejadas, onde as linhas verde, azul, roxa e rosa representam ligações de hidrogênio, cargas atrativas, pi-pi empilhadas/em forma de PiPi-T e pi-alkil, respectivamente. As Figuras B, D, F e H representam as estruturas químicas de naringenina-4'-O-Glucuronídeo (ZINC000045789238), ergoloide (ZINC000003995616), ohioensina A (ZINC000004098448) e prunetrin (ZINC000008662732), respectivamente, são mostradas. **Fonte:** O Autor (2024).

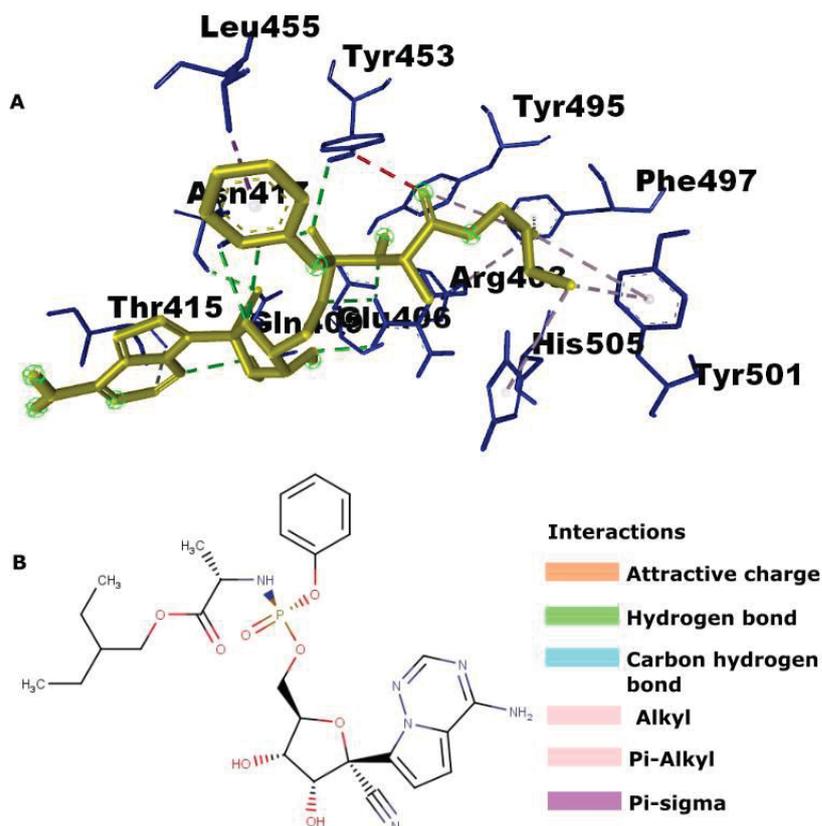


Figura 6.4. Modelagem de docking molecular do remdesivir (medicamento de referência). As interações do complexo proteína-ligante *Spike* são mostradas em linhas tracejadas. Estruturas amarelas e azuis representam o remdesivir e a proteína *Spike* (S1), respectivamente. As cores laranja, verde, rosa claro e roxo representam as ligações do tipo cargas de tração, ligações de hidrogênio, ligação carbono-hidrogênio, alquil, pi-alquil e pi-sigma, respectivamente. **Fonte:** O Autor (2024).

A Tabela 6.2 compara as interações do complexo proteína RBD (ômicron)-ligante a partir de triagem virtual e análises de docking molecular. Podemos observar que em ambas as análises (triagem virtual e docking molecular), a naringenina-4'-O-Glucuronídeo (ZINC000045789238) formou ligações de hidrogênio envolvendo os aminoácidos Ser496, Thr500 e His505 do RBD da proteína *Spike* (S1). O ligante ergoloide (ZINC000003995616) interagiu com os aminoácidos Ser496 (ligação carbono-hidrogênio) e Tyr501 (em forma de pi-pi) da região RBD da proteína *Spike* (S1). O ligante ohioensina A (ZINC000004098448) estabeleceu interações em forma de T pi-alquil e Pi-Pi com os aminoácidos Lys478, Phe486 e Val483, respectivamente.

Finalmente, o ligante prunetrina (ZINC000008662732) interagiu com os aminoácidos Arg408, Asn417 e Tyr453 do RBD, produzindo ligações de hidrogênio, ligações de hidrogênio e ligações carbono-hidrogênio, respectivamente. Podemos observar uma semelhança nas interações do complexo proteína-ligante na triagem virtual e nas análises de docking molecular.

Tabela 6.2. Comparação de interações do complexo ligante da proteína *Spike (RBD)* da variante ômicron do SARS-CoV-2 a partir de triagem virtual e análises de docking molecular

Modelagem	ZINC ID do composto	Aminoácidos RBD envolvidos na interação	Tipo de interação
Virtual screening (PyRx)	ZINC000045789238	Ser496, Thr500, His505	Ligação de hidrogênio
	ZINC000003995616	Ser496, Tyr495, Tyr501	ligação carbono-hidrogênio, pi-alquil e em forma de pi-pi
	ZINC000004098448	Asn477, Lys478, Ala484, Phe486	ligação pi-alquil, em forma de pi-pi e carbono-hidrogênio
	ZINC000008662732	Arg408, Asn417, Tyr453	Ligação de hidrogênio e ligação carbono-hidrogênio
Docking molecular (Autodock tools)	ZINC000045789238	Ser494, Ser496, Thr500, Thr501, His505	Ligação de hidrogênio
	ZINC000003995616	Asp405, Ser496, Tyr495, Tyr501,	Ligação de hidrogênio, carga atrativa, pi-alquil, ligação carbono-hidrogênio e formato pi-pi
	ZINC000004098448	Lys478, Val483, Phe486	pi-alquil, pi-pi em forma de T
	ZINC000008662732	Arg403, Arg408, Gln409, Asn417, Tyr453, Ser496, His505,	Ligação de hidrogênio, ligação carbono-hidrogênio
Docking Molecular (Autodock Vina)	ZINC000045789238	Ser496, Thr500, Thr501	Ligação de hidrogênio
	ZINC000003995616	Ser496, Tyr495, Tyr501	ligação carbono-hidrogênio, pi-alquil e em forma de pi-pi
	ZINC000004098448	Lys478, Phe486, Val483	ligação pi-alquil, em forma de pi-pi e carbono-hidrogênio
	ZINC000008662732	Arg408, Tyr453, His505	Ligação de hidrogênio e ligação carbono-hidrogênio

Nota: Naringenina-4'-O-Glucuronídeo (ZINC000045789238), ergolóide (ZINC000003995616), ohioensina A (ZINC000004098448) e prunetrina (ZINC000008662732). **Fonte:** O Autor (2024).

De acordo com as previsões feitas na plataforma ADMETLAB2 (Tabela 6.3), todos os quatro compostos avaliados (naringenina-4'-O-glicuronídeo, Ohioensina A, prunetrin e ergolóide) apresentavam características semelhantes a medicamentos de acordo com a regra e o triângulo dourado da *Pfizer*. A Tabela 6.3 mostra os resultados das previsões dos parâmetros farmacocinéticos. Dos quatro compostos, apenas naringenina-4'-O-glicuronídeo e ohioensina A tiveram maior probabilidade de absorção oral e aumentada e boa biodisponibilidade. No entanto, na distribuição, apenas o naringenina-4'-O-glicuronídeo apresentou os quatro parâmetros de distribuição (ligação às proteínas plasmáticas, distribuição de volume, barreira hematoencefálica e FU) simultaneamente dentro dos valores da faixa aceitável. Foi demonstrado que a naringenina-4'-O-glicuronídeo tem maior probabilidade de ser um substrato do CYP2C9. Na eliminação, todas as quatro moléculas apresentaram valores de depuração dentro da faixa tolerada.

Tabela 6.3. Predição dos parâmetros farmacocinéticos dos quatro principais ligantes que tiveram a maior afinidade pela proteína *Spike* (S1) RBD da mutação ômicron

	Naringenina-4'-O-glucuronide	Ohioensin A	Prunetrin	Ergolóide	
Parâmetro farmacocinético	Prob/valor*	Prob/valor*	Prob/valor*	Prob/valor*	Observação
Absorção					
Absorção Intestinal Humana (%)	70-90	70-90	0,0-10	0,0-10	-----
F (20% Biodisponibilidade) (%)	70-90	70-90	10-30	0,0-10	-----
Distribuição					
Ligação às proteínas plasmáticas (%)	87,12	92,757	87,730	93,306	Ótimo: < 90
Distribuição de Volume (L/kg)	0,37	1,857	0,741	3,095	Ótimo: 0,04 - 20
Barreira hematoencefálica (%)	0,0-10	0,0-10	8,802	30-50	
Fração livre (%)	12,89	4,387		3,056	Baixo<5; médio :5-20; Alto>20
Metabolismo					
Inibidor do CYP1A2 (%)	0,0-10	10-30	30-50	0,0-10	-----
Substrato do CYP1A2 (%)	0,0-10	70-90	10-30	0,0-10	-----
Inibidor do CYP2C19 (%)	0,0-10	10-30	0,0-10	50-70	-----
Substrato CYP2C19 (%)	0,0-10	90-100	0,0-10	90-100	-----

Inibidor do CYP2C9 (%)	0,0-10	0,0-10	10-30	90-100	-----
Substrato CYP2C9 (%)	70 – 90	10-30	70 – 90	10-30	-----
Inibidor do CYP2D6 (%)	0,0-10	0,0-10	30-50	0,0-10	-----
Substrato CYP2D6 (%)	0,0-10	10-30	30-50	10-30	-----
Inibidor do CYP3A4 (%)	0,0-10	10-30	10-30	90-100	-----
Substrato do CYP3A4 (%)	0,0-10	90-100	0,0-10	90-100	-----
Excreção					
Taxa de eliminação (mL/min/kg)	5,295	6,145	5,987	14,675	Baixo<5; Médio:5-15; Alto>15
Tempo de meia vida (h)	0,808	0,165	0,586	0,727	Curto<3; Longo>3

Nota: *Para os endpoints de classificação, os valores de probabilidade de predição são transformados em seis símbolos: 0-10% (- - -), 10-30% (- -), 30-50% (-), 50-70% (+), 70-90% (+ +) e 90-100% (+ + +). **Fonte:** O Autor (2024).

6.5.1.3 Predição da toxicidade aguda e crônica

A Tabela 6.4 mostra os resultados das previsões de toxicidade aguda e crônica obtidas na plataforma *PROTOX II*. Dentre os quatro compostos analisados, apenas o ligante ohioensina A foi o que apresentou algum tipo de toxicidade (citotoxicidade, ER, ER-LBD, MMP e fosfoproteína p53). Em contraste, o restante dos três ligantes não apresentou nenhum tipo de toxicidade. Os resultados de toxicidade aguda mostraram que todos os compostos apresentaram dose letal superior a 2.000 mg/kg e foram classificados como compostos de toxicidade classe V, que é a classe de compostos com pouca probabilidade de causar danos agudos.

Tabela 6.4. Predição de toxicidade aguda e crônica dos quatro principais ligantes que tiveram a maior afinidade pela proteína *Spike* (S1) RBD da mutação ômicron

Type of toxicity	Naringenin-4'-O-glucuronide		Ohioensin A		Prunetrin		Ergolóide	
	Predição	P	Predição	P	Predição	P	Predição	P
Hepatotoxicidade	Inativo	0,74	Inativo	0,71	Inativo	0,83	Inativo	0,96
Carcinogenicidade	Inativo	0,60	Inativo	0,58	Inativo	0,90	Inativo	0,52
Mutagenicidade	Inativo	0,68	Inativo	0,62	Inativo	0,65	Inativo	0,90
Citotoxicidade	Inativo	0,84	Ativo	0,68	Inativo	0,58	Inativo	0,54
AhR	Inativo	0,81	Inativo	0,59	Inativo	0,62	Inativo	0,99

Receptor andrógeno de	Inativo	0,98	Inativo	0,91	Inativo	0,98	Inativo	0,99
AR-LBD	Inativo	0,96	Inativo	0,95	Inativo	1,0	Inativo	1,0
Aromatase	Inativo	0,96	Inativo	0,79	Inativo	0,99	Inativo	0,94
pronto-socorro	Inativo	0,86	Ativo	0,65	Inativo	0,97	Inativo	0,98
ER-LBD	Inativo	0,93	Ativo	0,52	Inativo	0,99	Inativo	1,0
PPAR-Gama	Inativo	0,91	Inativo	0,62	Inativo	0,99	Inativo	0,98
nrf2/ARE	Inativo	0,90	Inativo	0,91	Inativo	0,99	Inativo	0,93
HSE	Inativo	0,90	Inativo	0,91	Inativo	0,99	Inativo	0,93
MMP	Inativo	0,65	Ativo	0,72	Inativo	0,99	Inativo	0,86
Fosfoproteína p53	Inativo	0,80	Ativo	0,55	Inativo	0,86	Inativo	0,98
ATAD5	Inativo	0,94	Inativo	0,75	Inativo	0,99	Inativo	0,99
Toxicidade aguda (DL50)	2300 mg/kg		2000 mg/kg		5000 mg/kg		2000 mg/kg	

Nota: AhR; Receptor de Aril Hidrocarboneto; AR-LBD: Domínio de Ligação ao Ligante do Receptor de Andrógênio; ER: Receptor Alfa de Estrogênio; ER-LBD: Domínio de Ligação ao Ligante do Receptor de Estrogênio; PPAR-Gama: Receptor Gama Ativado por Proliferador de Peroxissoma; nrf2/ARE: Fator nuclear (derivado de eritróide 2) semelhante a 2/elemento responsivo a antioxidante; HSE: Elemento de resposta do fator de choque térmico; MMP: Potencial de Membrana Mitocondrial; ATAD5: Proteína 5 contendo o domínio AAA da família ATPase; P: probabilidade. **Fonte:** O Autor (2024).

6.5.1.4 Simulações de dinâmica molecular

As simulações de DM para cada complexo de glicoproteína *Spike* (S1) da variante Omicron com os quatro ligantes (ZINC000045789238, ZINC000004098448, ZINC000008662732 e ZINC000013374469) foram conduzidas por 100 ns. Os resultados da análise de *RMSD* são mostrados na Figura 6.5. Todos os quatro complexos proteína-ligante apresentaram valores de *RMSD* inferiores a 0,3, mostrando que houve estabilidade de ligação entre os resíduos de aminoácidos da proteína *Spike* (S1) da região RBD durante todo o período de dinâmica: complexo ZINC000045789238-*Spike*(S1) *RMSD* 0,124 – 0,250 nm e *RMSD* médio = 0,242 nm; Complexo ZINC000004098448-*Spike*(S1) *RMSD* 0,122 – 0,251 nm e *RMSD* médio = 0,232 nm; Complexo ZINC000008662732-*Spike*(S1) *RMSD* 0,116 – 0,285 nm e *RMSD* médio = 0,246; ZINC000003995616-*Spike*(S1) *RMSD* (0,125 – 0,308 nm) e *RMSD* médio = 0,264 nm (Figura 3).

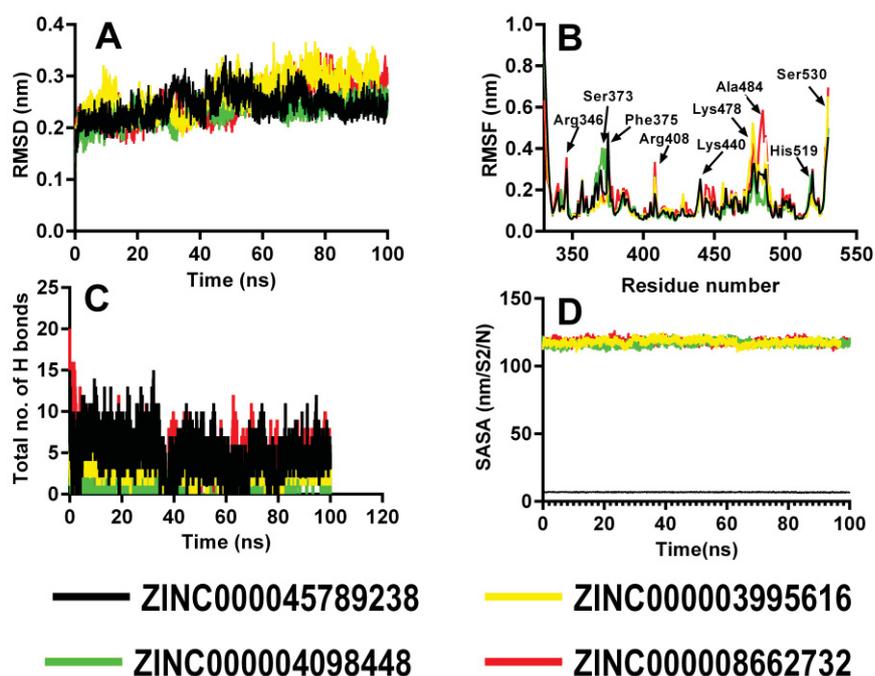


Figura 6.5. Simulações de dinâmica molecular dos quatro principais ligantes. As Figuras A, B, C e D mostram o perfil de RMSD, RMSF, ligação de hidrogênio total e acessibilidade ao solvente para as simulações de dinâmica molecular no tempo de 100 ns. Os complexos naringenina-4'-O-Glucuronídeo (ZINC000045789238)-*Spike*, ergoloid (ZINC000003995616)-*Spike*, ohioensin A (ZINC000004098448)-*Spike* e prunetrin (ZINC000008662732)-*Spike* são coloridos em preto, amarelo, verde e vermelho, respectivamente. **Fonte:** O Autor (2024).

Os resultados da simulação de DM do remdesivir (medicamento de referência) são mostrados na Figura 6.6. O valor médio de *RMSD* do complexo de proteína *Spike* remdesivir-*RBD* foi de 0,2417 nm, mostrando também estabilidade de ligação do complexo formado ($RMSD < 0,3$), e esses resultados são semelhantes aos encontrados com os quatro ligantes em nosso estudo (Fig. 6.6-A).

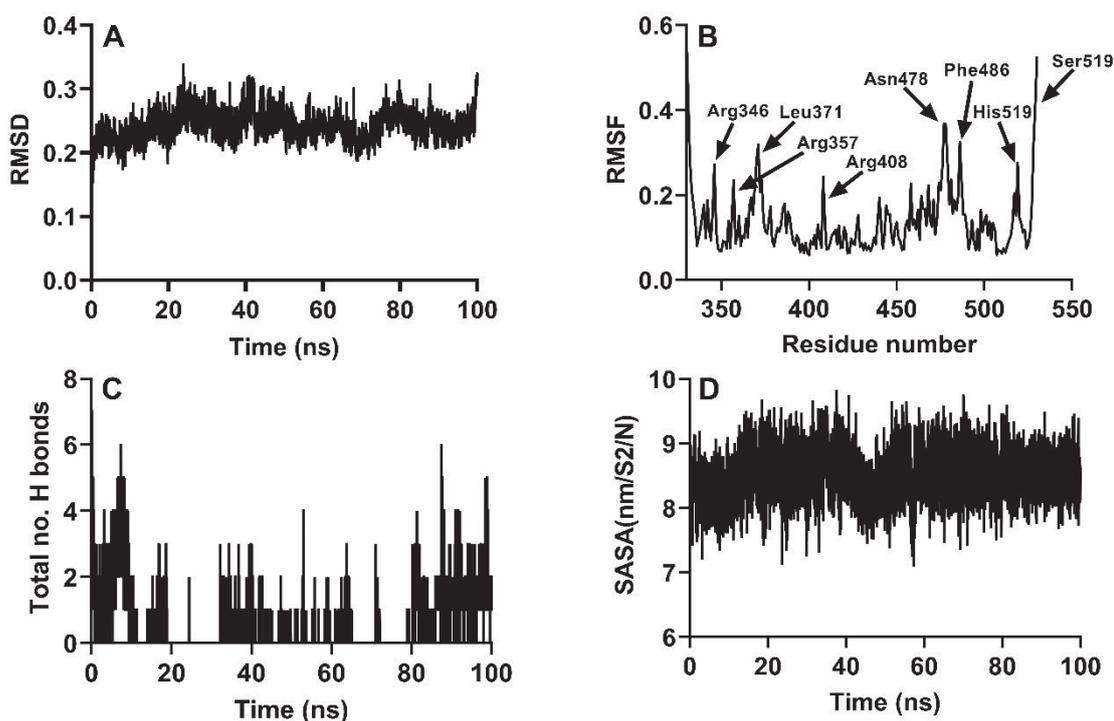


Figura 6.6. Simulações de dinâmica molecular do remdesivir (medicamento de referência). As Figuras A, B, C e D mostram o perfil de *RMSD*, *RMSF*, ligação de hidrogênio total e acessibilidade ao solvente para as simulações de dinâmica molecular no tempo de 100 ns. **Fonte:** O Autor (2024).

O comportamento dinâmico dos resíduos de aminoácidos da glicoproteína *Spike* (S1) também foi estudado através do cálculo dos valores de *RMSF*. Embora todos os resíduos de aminoácidos da glicoproteína *Spike* apresentassem flutuações dentro do limite tolerado ($RMSF < 1,3$ nm), os resíduos de aminoácidos da mutação Ômicron produziram as maiores flutuações, especialmente Phe375, Lys440, Lys478 e Phe375 (Figura 6.6). Para o complexo ZINC000045789238-*Spike*, cinco resíduos de aminoácidos foram identificados nesta análise, três pertencentes à mutação Ômicron (Phe375, Lys440, Lys478) e dois resíduos não mutantes (His519 e Ser530). No complexo ZINC000003995616-*Spike* (S1), os resíduos de aminoácidos que sofreram maior flutuação foram Arg408, Lys478, Ser530, His519 e Ser530. Para o complexo ZINC000004098448-*Spike* (S1), sete aminoácidos apresentaram mais alterações, quatro deles envolvidos na mutação Ômicron (Leu371, Proina373, Lys440, Lys478) e três resíduos não mutantes (Arg408, His519 e Ser530). Para o complexo ZINC000008662732-*Spike*, foram identificados três aminoácidos mutantes (Lys478,

Ala484 e Phe375) e cinco aminoácidos não mutantes (Pro330, Arg346, Arg408, Ala486 e Ser530). A Figura 6.6-B mostra os resultados de *RMSF* para remdesivir, e podemos ver que cinco resíduos de aminoácidos *RBD* (Arg346, Arg357, Arg408, His519 e Ser519) que apresentaram valores de *RMSF* mais elevados (maior flutuação) foram os mesmos resíduos de aminoácidos que também apresentaram maior flutuações nos quatro ligantes identificados neste estudo.

O complexo ZINC000045789238-*Spike* apresentou o maior número de interações de ligações de hidrogênio (total = 46.007 ligações de hidrogênio), com uma média de 4.601 ligações por nanosegundo de dinâmica, e os seguintes resíduos de aminoácidos estabelecendo essas ligações: Asn417, Ser494, Ser496, Arg403, Arg408 e His505. O complexo ZINC000008662732 – *Spike* (S1) apresentou o segundo maior número de ligações de hidrogênio (total = 25.847 ligações de hidrogênio), com média de 259 ligações por nanosegundo, seguido pelo complexo ZINC000003995616 – *Spike* (S1) (total = 21.750, média = 218) e finalmente o complexo ZINC000004098448 – *Spike* (S1) (total = 8607, média = 86). No caso do complexo remdesivir-proteína *Spike* (Figura 6.6-C), houve formação de 6.074 ligações de hidrogênio em todo o período da dinâmica molecular, sendo que os aminoácidos *RBD* e três resíduos de aminoácidos *RBD* (Ser494, Ser496, Arg408) foram identificaram que também estavam envolvidos na ligação com nossos quatro ligantes.

6.5.1.5 Cálculo das energias livres de ligação *MM/PBSA* e *MM/GBSA*

A Tabela 6.5 mostra o cálculo resumido das energias livres de ligação médias de *MM/PBSA* e *MM/GBSA* do complexo ΔG , receptor ΔG (proteína *Spike*) e ligante ΔG , bem como a diferença de energia ΔG (Complexo - Receptor - Ligante). Todos os quatro ligantes apresentaram valores médios (-) de *MM/GBSA* para o complexo ΔG e a diferença ΔG (Complexo - Receptor - Ligante). No entanto, apenas o ligante ZINC000045789238 apresentou uma (-) média de energia livre de ligação *MM/GBSA* para o ΔG (Complexo - Receptor – Ligante), que foi de -3,74 kcal/mol, significando que este complexo foi o único a estabelecer ligação com energia favorável ao alvo molecular. Os três ligantes restantes apresentaram valores médios de *MM/PBSA* >0. Os resultados de *MM/PBSA* e *MM/GBSA* do remdesivir são semelhantes aos encontrados pela nossa molécula promissora (naringenina-4'-O-glicuronídeo), ou

seja, valores de *MM/PBSA* e *MM/GBSA* menores que zero (Tabela 6.5). Tal semelhança consolida nossos resultados que sugerem que o fitoquímico naringenina-4'-O-glicuronídeo é um potencial candidato a medicamento para o tratamento da COVID-19.

Tabela 6.5. Energia livre de ligação de *MM/PBSA* e *MM/GBSA* do complexo entre a proteína *Spike (RBD)* da variante ômicron e os quatro ligantes e remdesivir (fármaco controle)

	MM/PBSA			
ZINC ID (ligante)	ΔG Complexo	ΔG Receptor	ΔG Ligante	ΔG (Complexo - Receptor - Ligante)
ZINC000045789238	-1983.33	-1993.60	14.01	-3.74
ZINC000004098448	-1961.07	-2010.66	47.64	1.95
ZINC000008662732	-1875.03	-1989.50	106.42	8.04
ZINC000003995616	-1933.84	-2003.13	63.19	6.11
Remdesivir	-1721.55	-1919.07	202.82	-5.30
	MM/GBSA			
ZINC ID (ligante)	ΔG Complex	ΔG Receptor	ΔG Ligand	ΔG (Complex - Receptor - Ligand)
ZINC000045789238	-2539.84	-2543.37	19.18	-15.65
ZINC000004098448	-2522.03	-2562.51	51.83	-11.35
ZINC000008662732	-2428.50	-2532.94	112.70	-8.25
ZINC000003995616	-2496.10	-2549.13	66.14	-13.10
Remdesivir	-2276.65	-2469.59	201.68	-8.73

Nota: Naringenina-4'-O-Glucuronide (ZINC000045789238), ergolóide (ZINC000003995616), ohioensin A (ZINC000004098448) and prunetrin (ZINC000008662732). **Fonte:** O Autor (2024).

6.5.1.6 Simulações Pós-dinâmica molecular

A Figura 6.7 mostra os resultados dos cálculos de simulação de pós-dinâmica molecular (pós-MD). Podemos observar que pelo resultado dos cálculos das distâncias (Å) do centro de massas (CoM) por quadro mostrado na Figura 5A, o complexo ZINC000045789238-*Spike* foi o único com distâncias de CoM menores. Seu perfil de CoM é semelhante ao perfil de CoM do complexo remdesivir-*Spike* (medicamento de controle) em todos os momentos de simulação pós-DM. Cálculos de energias livres de ligação de MM/GBSA por quadro (Figura 6.7-C), todos os quatro ligantes testados e remdesivir (medicamento de controle) mostraram boa estabilidade de ligação [(-) (MM/GBSA)]. No entanto, para o cálculo de energia livre de MM /PBSA por quadro (Figura 6.7-B), apenas o ligante ZINC000045789238 e o remdesivir (medicamento controle) foram os que apresentaram energia livre de ligação estável com a proteína *Spike* do SARS-CoV-2 [(-) (MM/PBSA)].

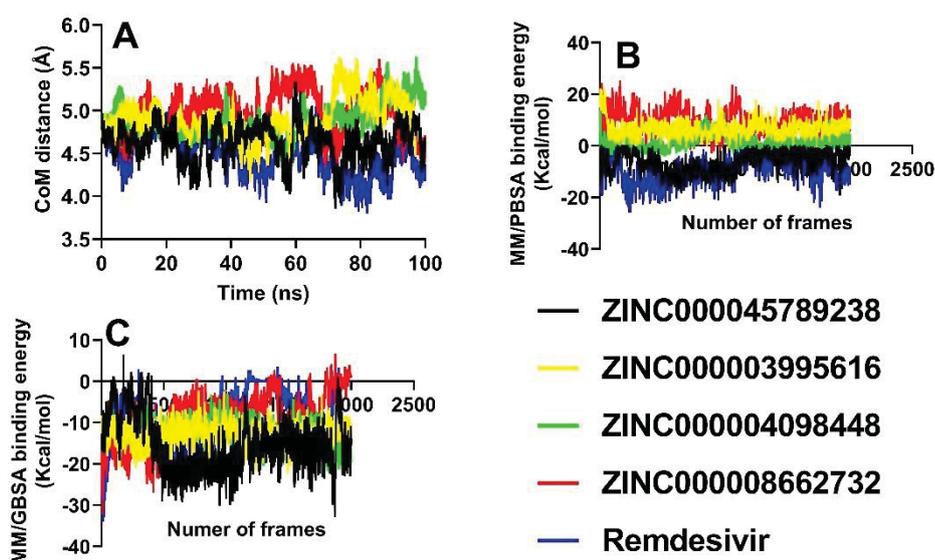


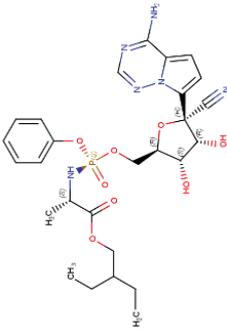
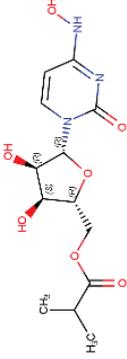
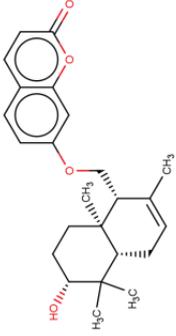
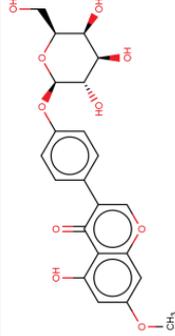
Figura 6.7. Simulações de pós-dinâmica molecular (Pós-DM) dos quatro principais ligantes e remdesivir (fármaco de controle). As Figuras A, B e C mostram o perfil do centro de massa (CoM) por quadro, energia livre de ligação *MM/PBSA* por quadro e

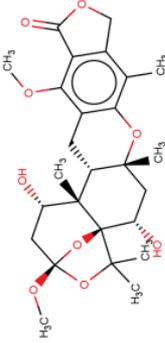
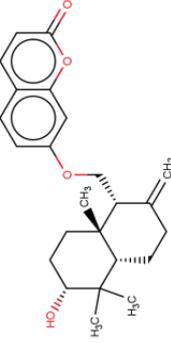
energia livre de ligação MM/GBSA por quadro para as simulações pós-MD em um tempo de 20 ns. Os complexos naringenina-4'-O-Glucuronídeo (ZINC000045789238)-*Spike*, ergoloid (ZINC000003995616)-*Spike*, ohioensina A (ZINC000004098448)-*Spike*, prunetrin (ZINC000008662732)-*Spike* e remdesivir-*Spike* são colorido em preto, amarelo, verde, vermelho e azul, respectivamente. **Fonte:** O Autor (2024).

6.5.2 Estudo II: Triagem virtual, docking molecular, análise de drug-likeness e predições ADMET: proteína *Spike* (RBD) do SARS-CoV-2 tipo selvagem

Os resultados da triagem virtual e da análise de docking molecular dos 171 mil ligantes com a proteína *Spike* (RBD) do SARS-CoV-2 do tipo selvagem são mostrados na tabela 6.6. Dos 34 compostos identificados na triagem virtual como tendo maior afinidade pela RBD, apenas 4 compostos, BNC5 (feselol), BN14, BN15 e BN27 apresentaram melhores resultados nas predições dos parâmetros farmacocinéticos e de toxicidade (Tabela 6.6). Em termos de semelhanças estruturais, os quatro compostos pertencem à classe das cumarinas (benzopironas), pois possuem um anel benzênico fundido a uma lactona (heterosídeos). Outra semelhança estrutural é que os quatro compostos apresentam o anel benzopirona que forma uma ligação éter com um grupo álcool cíclico polihidroxilado. Além disso, estes quatro compostos tinham afinidade de ligação semelhante ao pico (S1) com os medicamentos antivirais actualmente utilizados no tratamento da COVID-19, nomeadamente, remdesivir e molnupinavir (Tabela 6.6). Os quatro compostos bioativos naturais (BNC5, BNC14, BN15 e BN27) e os antivirais padrões (remdesivir e molnupinavir) tiveram afinidade para bloquear os seguintes resíduos de aminoácidos da região S1 da proteína *Spike*: R454, W436, N437, F464, E516, G482, F456, F374, S373, T430, R509 e V367 (Tabela 6.6). As interações químicas mais importantes foram as ligações de hidrogênio (Tabela 6.6).

Tabela 6.6. Comparação dos resultados de docking molecular dos ligantes com maior afinidade pela proteína Spike (RBD) do SARS-CoV-2 tipo selvagem com os fármacos controle (molnupiravir e remdesivir)

Name/ZINC ID	Molar weight (g mol ⁻¹)	Chemical structure	Energia de afinidade (kcal mol ⁻¹)	Amino ácido	Interações químicas
Remdesivir*	602,6		-8,5	F456, R454, W436, N437, R509, V367, F464, E516	Ligações de hidrogênio, Pi-alquil, Pi-pi em forma de T, ligação carbono-hidrogênio
Molnupiravir*	329,31		-8,4	G482, F456, F374, S373, T430, R509, V367	Ligações de hidrogênio, Pi-alquil, Pi-pi em forma de T, ligação carbono-hidrogênio
BNC5 (ZINC000000704424)	382,5		-8,5	G482, F456, R454	Ligações de hidrogênio
BNC14 (ZINC000008662732)	446,4		-8,7	D364, V367, L368, W436, N437, N343, F374, S373	Ligações de hidrogênio, Pi-alquil, Pi-pi em forma de T, ligação carbono-hidrogênio

BNC15 (ZINC0000013374469)	490,5		-8,4	T430, R355, F464, E516	Limites de hidrogênio, em forma de Pi-pi T
BNC27 (ZINC000003022621)	382,5		-8,2	R509, V367	Ligações de hidrogênio, Pi- alquil

Nota: * Esses compostos são medicamentos antivirais tradicionalmente utilizados para o tratamento da COVID-19. Neste estudo, eles foram usados como grupo controle. **Fonte:** O Autor (2024).

6.5.2.1 Análise de *drug-likeness*

Na tabela 6.7 são mostrados os resultados das análises de *drug-likeness*. De acordo com a regra de Lipinski, uma molécula é considerada semelhante a um medicamento se atender a pelo menos três dos quatro critérios estabelecidos; ao passo que, pela regra de Veber, o medicamento deveria atender a todos os três critérios pré-definidos. Em nosso estudo, o consenso entre as duas regras foi utilizado para definir uma molécula como tendo semelhança com um medicamento. Dos 34 produtos naturais, 24 (70,5%) apresentavam características de semelhança com medicamentos (tabela 6.7). Assim, os 10 compostos restantes não foram considerados para estudos posteriores.

Tabela 6.7. Análise de similaridade medicamentosa (druglikeness) dos principais compostos naturais com maior afinidade com a glicoproteína Spike (RBD) do SARS-CoV-2 tipo selvagem

Produto natural	Regras dos cinco de Lipinski					Regra de Veber					Drug-like consensual
	MW<500 (g/mol)	LogP≤5	DLH ≤5	ALH ≤10	Drug-like?	NLR ≤10	TPSA ≤140	ALH e DLH ≤12	Drug-like?		
NBC1	449,62	3,594	4	4	Sim (100%)	7	106,86	8	Sim (100%)	Drug-like	
NBC2	470,606	3,495	2	6	Sim (100%)	2	96,36	8	Sim (100%)	Drug-like	
NBC3	470,694	6,413	2	3	Sim (75%)	1	74,6	5	Sim (100%)	Drug-like	
NBC4	456,711	7,234	2	2	Sim (75%)	1	57,53	4	Sim (100%)	Drug-like	
NBC5	382,5	4,941	1	4	Sim (100%)	3	59,67	5	Sim (100%)	Drug-like	
NBC6	482,441	2,363	5	10	Sim (100%)	4	155,14	15	Não (33,33%)	Não Drug-like	
NBC7	446,408	0,353	5	10	Sim (100%)	5	159,05	15	Não (33,33%)	Não Drug-like	
NBC8	372,376	4,371	3	5	Sim (100%)	0	86,99	8	Sim (100%)	Drug-like	
NBC9	458,473	4,607	3	7	Sim (100%)	4	153,69	10	Não (66,67%)	Não Drug-like	
NBC10	446,364	0,142	6	10	Sim (75%)	4	187,12	14	Não (33,33%)	Não Drug-like	
NBC11	482,441	2,369	5	10	Sim (100%)	4	155,14	15	Não (33,33%)	Não Drug-like	
NBC12	384,516	4,258	1	4	Sim (100%)	1	62,97	5	Sim (100%)	Drug-like	
NBC13	464,568	8,45	2	2	Sim (75%)	5	48,91	4	Sim (100%)	Drug-like	
NBC14	327,181	4,092	2	1	Sim (100%)	0	44,89	3	Sim (100%)	Drug-like	

NBC15	490,549	2,384	2	9	Sim (100%)	2	112,91	11	Sim (100%)	Drug-like
NBC16	316,353	2,778	0	5	Sim (100%)	2	61,83	5	Sim (100%)	Drug-like
NBC17	326,304	3,247	3	6	Sim (100%)	0	100,13	9	Sim (100%)	Drug-like
NBC18	482,441	2,369	5	10	Sim (100%)	4	155,14	15	Não (33,33%)	Não Drug-like
NBC19	515,713	2,368	5	6	Sim (75%)	8	144,16	11	Não (66,67%)	Não Drug-like
NBC20	327,181	4,092	2	1	Sim (100%)	0	44,89	3	Sim (100%)	Drug-like
NBC21	376,496	5,123	1	3	Sim (75%)	3	46,53	4	Sim (100%)	Drug-like
NBC22	611,743	2,717	3	6	Sim (75%)	6	118,21	9	Sim (100%)	Drug-like
NBC23	314,425	3,72	2	3	Sim (100%)	1	53,6	5	Sim (100%)	Drug-like
NBC24	356,374	2,51	1	6	Sim (100%)	4	66,38	7	Sim (100%)	Drug-like
NBC25	418,398	0,277	5	9	Sim (100%)	4	145,91	14	Não (33,33%)	Não Drug-like
NBC26	380,484	5,15	0	4	Sim (75%)	3	56,51	4	Sim (100%)	Drug-like
NBC27	382,5	4,941	1	4	Sim (100%)	3	59,67	5	Sim (100%)	Drug-like
NBC28	330,468	4,45	0	3	Sim (100%)	2	43,37	3	Sim (100%)	Drug-like
NBC29	596,724	6,156	3	8	Não (50%)	8	94,86	11	Sim (100%)	Não Drug-like
NBC30	324,288	3,006	1	6	Sim (100%)	1	78,13	7	Sim (100%)	Drug-like
NBC31	397,647	5,655	1	2	Sim (75%)	0	23,47	3	Sim (100%)	Drug-like

NBC32	354,358	3,219	0	6	Sim (100%)	2	55,38	6	Sim (100%)	Drug-like
NBC33	442,684	7,154	2	2	Sim (75%)	5	57,53	4	Sim (100%)	Drug-like
NBC34	448,38	0,075	6	10	Sim (75%)	4	183,21	16	Não (33,33%)	Não Drug-like

Nota: *DLH*: Doadores de ligações de hidrogênio; *ALH*: Aceitadores de ligações de hidrogênio; *NLR*: Número de ligações rotacionáveis; *TPSA*: área de superfície polar topográfica; *MW*: peso molecular. **Fonte:** O Autor (2024).

6.5.2.2 Predição dos parâmetros farmacocinéticos

A Tabela 6.8 mostra a análise farmacocinética (absorção, distribuição, metabolismo e excreção) dos 24 compostos naturais que apresentaram simultaneamente maior afinidade com a espícula S1 e apresentaram características de semelhança ao fármaco. Destes, 21 (80,7%) apresentaram probabilidade significativa de absorção intestinal humana e biodisponibilidade de 20-30%. Embora todos os compostos apresentassem valores de volume de distribuição dentro da faixa recomendada (0,04-20 L kg⁻¹), os compostos NBC3, NBC5, NBC12, NBC14, NBC15, NBC21, NBC27 e NBC33 foram os únicos com menor probabilidade de atravessar o sistema hematoencefálico humano. barreira (ou seja, evitando a toxicidade do sistema nervoso central). A fração das moléculas de NBC3, NBC5, NBC12, NBC14, NBC15, NBC21, NBC27 e NBC33 que seriam transportadas pelas proteínas plasmáticas durante o processo de distribuição foi estimada dentro da faixa recomendada (<90%). Em relação à fração de moléculas não ligadas às proteínas plasmáticas, apenas seis compostos (NBC3, NBC5, NBC12, NBC14, NBC15 e NBC33) estavam dentro da faixa recomendada ($F_u > 5\%$).¹⁴

De acordo com os resultados das previsões feitas na plataforma *in silico* (online) *ADMETlab 2.0* mostradas na Tabela 6.8, os compostos NBC3, NBC5, NBC12, NBC14, NBC15, NBC27 e NBC33 tiveram maior probabilidade de atuar como substratos do citocromo P-450 enzimas (CYP1A2, CYP3A4, CYP2C9, CYP2C19 e CYP2D6), o que significa que podem ser biotransformadas em metabólitos solúveis, que são facilmente eliminados pelo organismo. Apenas os compostos NBC5, NBC12, NBC14, NBC15, NBC27 e NBC34 apresentaram valores de depuração (> 5 mL/min/kg) e meia-vida de eliminação ($T_{1/2} > 0,5$ h) dentro dos intervalos recomendados (ou seja, propriedades farmacocinéticas aceitáveis)

Tabela 6.8. Predição de parâmetros farmacocinéticos dos compostos naturais que simultaneamente apresentaram melhores afinidades de ligação com S1 e possuem características de semelhança com medicamentos

	NBC1	NBC2	NBC3	NBC4	NBC5	NBC8	NBC12	NBC13	NBC14	NBC15	NBC16	NBC17	NBC18	NBC19
Absorção														
Inibidor de P _{gp}	++	+++	-	---	+++	---	---	+++	---	+	---	---	++	---
Substrato P _{gp}	-	---	---	---	+++	+++	+++	+++	+	+++	---	---	---	---
Absorção Intestinal Humana	-	---	---	---	---	---	---	---	---	---	---	---	---	++
F (20% de biodisponibilidade)	-	---	---	+++	+++	++	+++	+++	---	-	---	---	---	+++
F (30% de biodisponibilidade)	++	--	+++	+	+++	+++	+++	--	++	+	+++	---	++	+++
Distribuição														
Ligação às proteínas plasmáticas (PPB)	92,3	65,052	91,564	97,647	82,677	92,757	92,959	104,156	87,730	79,789	96,099	95,683	95,154	89,336
Distribuição de Volume (L/kg)	0,448	1,267	0,671	0,784	1,557	1,857	1,798	0,449	0,741	1,604	1,113	0,694	0,646	0,541
Barreira Hematoencefálica	---	+++	--	---	---	---	---	---	---	++	---	---	---	---
UF (%)	4,4	23,592	7,613	2,938	7,249	4,387	5,196	0,503	8,802	6,209	6,653	7,730	8,125	3,056
Metabolismo														
Inibidor do CYP1A2	---	---	---	---	-	---	---	+++	-	---	++	+++	---	---
Substrato CYP1A2	--	-	++	-	-	++	+	---	---	+++	+	-	---	+
Inibidor do CYP2C19	---	---	---	---	-	---	---	-	---	---	-	---	---	---
Substrato CYP2C19	+	+	++	+++	---	+++	+	---	---	++	---	---	---	+
Inibidor do CYP2C9	---	---	---	---	+	+	-	---	---	---	++	++	++	---
Substrato CYP2C9	+	---	---	+++	+++	+++	---	++	++	---	---	++	++	---
Inibidor do CYP2D6	---	---	---	---	-	---	---	---	+	---	---	+	---	---
Substrato CYP2D6	--	---	---	-	++	---	+	+	+	---	---	---	---	---
Inibidor do CYP3A4	---	++	-	---	-	-	++	---	---	---	+	-	++	---
Substrato CYP3A4	---	-	+	---	-	+++	-	---	---	++	---	---	---	---
Eliminação														
Taxa de eliminação (mL/min/kg)	3,880	7,711	2,517	2,144	11,432	6,145	15,114	7,447	5,987	13,935	12,883	2,849	7,004	2,164
Tempo de meia vida (T 1/2)	0,891	0,751	0,079	0,134	0,065	0,165	0,174	0,078	0,586	0,423	0,086	0,717	0,303	0,599
Absorção														
Inibidor de P _{gp}	+++	+++	+++	---	+	+++	+++	+++	+++	+++	---	---	++	---
Substrato P _{gp}	---	---	+	-	---	---	+++	---	---	+++	---	---	---	---
Absorção Intestinal Humana	---	---	---	---	---	---	---	---	---	---	---	---	---	---
F (20% de biodisponibilidade)	---	+++	---	---	---	++	---	++	+	---	---	---	---	---
F (30% de biodisponibilidade)	---	+++	+++	+++	---	+++	+++	---	+++	++	-	---	---	---
Distribuição														
Ligação às proteínas plasmáticas (PPB)	96,704	98,781	93,306	97,459	97,839	91,405	85,520	89,998	92,023	92,961	81,130	87,719	96,274	-----

Distribuição de Volume (L/kg)	1,099	0,672	3,095	1,569	1,232	1,049	1,391	1,010	1,261	1,038	1,105	1,719	0,752	-----
Barreira Hematoencefálica	++	--	-	--	--	---	---	--	---	--	---	-	---	-----
Fração não ligada (%)	1,305	1,225	3,056	3,049	2,530	6,015	3,570	3,269	6,389	7,014	16,157	5,062	2,187	-----
Metabolismo														-----
Inibidor do CYP1A2	+++	-	---	---	+++	+	-	---	++	-	+++	---	+	-----
Substrato CYP1A2	-	--	---	--	--	++	+	---	++	-	++	--	-	-----
Inibidor do CYP2C19	++	-	+	--	+++	+	-	---	++	-	+	--	+++	-----
Substrato CYP2C19	--	--	+++	+	--	+	-	+++	+	---	-	++	--	-----
Inibidor do CYP2C9	++	+	+++	--	++	+	-	---	+	+	++	---	+	-----
Substrato CYP2C9	++	++	--	--	++	+++	++	--	++	++	++	---	+	-----
Inibidor do CYP2D6	+	---	---	++	+++	+	-	---	+	-	++	+++	+++	-----
Substrato CYP2D6	++	-	--	--	+++	+++	++	---	++	++	--	++	+++	-----
Inibidor do CYP3A4	++	---	+++	+++	+++	+	---	+	+	---	+	++	+++	-----
Substrato CYP3A4	--	-	+++	-	+	-	-	-	-	-	--	++	-	-----
Eliminação														-----
Taxa de eliminação (mL/min/kg)	1,633	13,434	14,675	11,093	17,590	11,596	9,815	7,885	16,472	9,452	2,825	14,885	16,448	-----
Tempo de meia vida (T 1/2)	0,518	0,102	0,727	0,549	0,282	0,224	0,074	0,515	0,197	0,057	0,413	0,078	0,150	-----

Nota: Para os endpoints de classificação, os valores de probabilidade de predição são transformados em seis símbolos: 0-0,1(---), 0,1-0,3(--), 0,3-0,5(-), 0,5-0,7(+), 0,7-0,9(++), 0,9-1,0(+++). **Fonte:** O Autor (2024).

6.5.2.3 *Predição da toxicidade aguda e crônica*

Dados os resultados farmacocinéticos, os compostos NBC5, NBC12, NBC14, NBC15, NBC27 e NBC33 foram os únicos submetidos para as previsões de toxicidade. NBC12 mostrou uma alta probabilidade de causar dois tipos de toxicidade por meio de vias de sinalização de receptores nucleares (receptor de estrogênio alfa e receptor gama ativado por proliferador de peroxissoma) e vias de resposta ao estresse (potencial de membrana mitocondrial e supressor de tumor de fosfoproteína p53. NBC33 tinha maior probabilidade de causar carcinogenicidade, mutagenicidade e toxicidade por vias de resposta ao estresse [fator nuclear (derivado de eritróide 2) - like 2/elemento responsivo a antioxidante (nrf2/ARE) e elemento de resposta ao fator de choque térmico (HSE)]. Os compostos NBC5, NBC14, NBC15 e NBC27 foram os únicos aqueles que não apresentaram probabilidade de causar qualquer tipo de toxicidade avaliada (Tabela 6.9)

A dose letal prevista (dose de toxicidade aguda) de NBC5, NBC14, NBC15 e NBC27 foi estimada em 3.200 mg kg⁻¹, 5.000 mg kg⁻¹, 3.000 mg kg⁻¹ e 3.200 mg kg⁻¹, respectivamente (Tabela 6.9), mostrando que esses compostos apresentam baixa toxicidade (ou seja, o LD50 é superior a 500 mg kg⁻¹), o que significa que são candidatos promissores para avaliação em ensaios pré-clínicos.

Tabela 6.9. Predição de toxicidade aguda e toxicidade crônica dos seis compostos que tiveram os melhores resultados farmacocinéticos

	NBC5	NBC12	NBC14	NBC15	NBC27	NBC33
Grupo de toxicidade	Predição	Predição	Predição	Predição	Predição	Predição
Orgão-específica	-----	-----	-----	-----	-----	-----
Hepatotoxicidade	Inativo	Inativo	Inativo	Inativo	Inativo	Inativo
Toxicidade cardíaca	Inativo	Inativo	Inativo	Inativo	Inativo	Inativo
Ponto final de toxicidade	-----	-----	-----	-----	-----	-----
Carcinogenicidade	Inativo	Inativo	Inativo	Inativo	Inativo	Ativo
Mutagenicidade	Inativo	Inativo	Inativo	Inativo	Inativo	Ativo
Citotoxicidade	Inativo	Inativo	Inativo	Inativo	Inativo	Inativo
Vias de sinalização de receptores nucleares						
Receptor de hidrocarboneto de arila (AhR)	Inativo	Inativo	Inativo	Inativo	Inativo	Inativo
Receptor de andrógeno (AR)	Inativo	Inativo	Inativo	Inativo	Inativo	Inativo
Domínio de ligação do ligante do receptor de andrógeno (AR-LBD)	Inativo	Inativo	Inativo	Inativo	Inativo	Inativo
Aromatase	Inativo	Inativo	Inativo	Inativo	Inativo	Inativo
Receptor de Estrogênio Alfa (ER)	Inativo	Ativo	Inativo	Inativo	Inativo	Inativo
Domínio de Ligação ao Ligante do Receptor de Estrogênio (ER-LBD)	Inativo	Inativo	Inativo	Inativo	Inativo	Inativo
Receptor gama ativado por proliferador de peroxissoma (PPAR-gama)	Inativo	Ativo	Inativo	Inativo	Inativo	Inativo
Vias de resposta ao estresse	-----	-----	-----	-----	-----	-----
Fator nuclear (derivado de eritróide 2) semelhante a 2/elemento responsivo a antioxidante (nrf2/ARE)	Inativo	Inativo	Inativo	Inativo	Inativo	Ativo
Elemento de resposta do fator de choque térmico (HSE)	Inativo	Inativo	Inativo	Inativo	Inativo	Ativo
Potencial de membrana mitocondrial (MMP)	Inativo	Ativo	Inativo	Inativo	Inativo	Inativo
Fosfoproteína (supressor de tumor) p53	Inativo	Ativo	Inativo	Inativo	Inativo	Inativo
Proteína 5 contendo o domínio AAA da família ATPase (ATAD5)	Inativo	Inativo	Inativo	Inativo	Inativo	Inativo
LD50 (mg/kg)	3200	59800	5000	3000	3200	1000

As poses (conformações mais estáveis resultantes dos resultados de docking molecular) de NBC5, NBC14, NBC15 e NBC27 ligadas à proteína *Spike S1* do SARS-CoV-2 são mostradas na Figura 6.7. As ligações do tipo ligação de hidrogênio foram as mais importantes na interação ligante-proteína alvo (Figura 6.8).

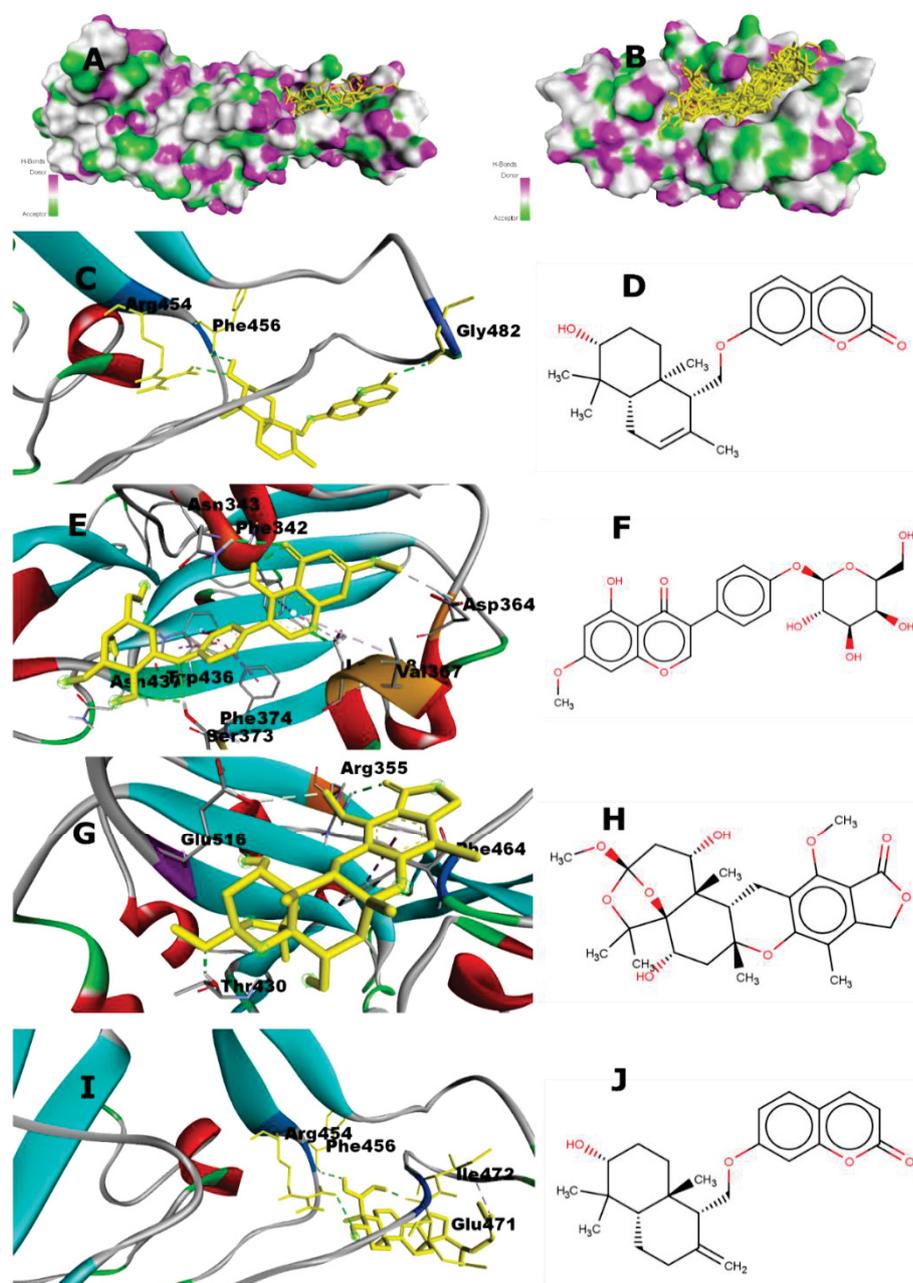


Figura 6.7. Resultados da análise de docking de poses de compostos com maior afinidade com a espícula (S1) do SARS-CoV-2. São mostrados apenas os docking dos ligantes que apresentaram semelhança com o fármaco, e que também apresentaram melhores resultados na análise ADMET (NBC5, NBC14, NBC15 e NBC27). (A): todos os quatro ligantes (NBC5, NBC14, NBC15 e NBC27) são mostrados ancorados na mesma cavidade da superfície da espiga (S1), de acordo com ligações de hidrogênio. (B): todos os estereoisômeros, todos os tautômeros dominantes e todas as microespécies no estado de protonação em pH fisiológico (pH = 7,4) dos quatro ligantes que se mostraram promissores contra COVID-19 (NBC5, NBC14, NBC15 e NBC27) são ligados na mesma cavidade da espiga, mostrando a grande seletividade para o sítio S1. O acoplamento do ligante NBC5 é ilustrado em (C), e a estrutura química do ligante NBC5 é ilustrada em (D). O acoplamento do ligante NBC14 é ilustrado em (E), e a estrutura química do ligante NBC14 é ilustrada em (F). O acoplamento do ligante NBC15 é ilustrado em (G), e a estrutura química do ligante NBC15 é mostrada em (H). O acoplamento do ligante NBC27 é ilustrado em (I), e a estrutura química do ligante NBC27 é mostrada em (J). **Fonte:** O Autor (2024).

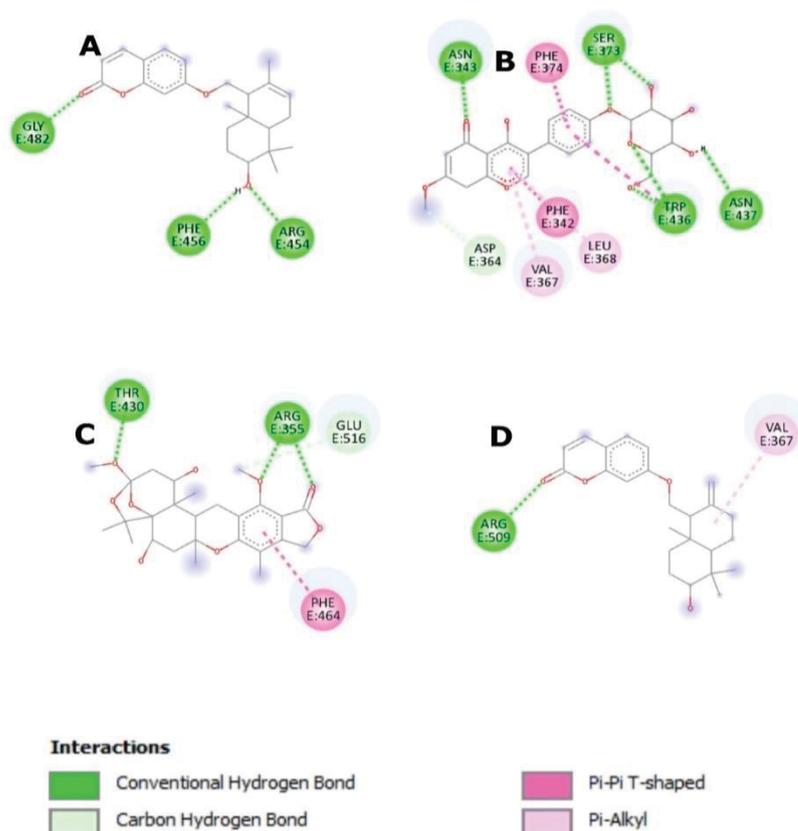


Figura 6.8. Estruturas 2D de interações proteína-ligante. São mostradas apenas as estruturas dos ligantes mais promissores contra a proteína *Spike* do SARS-CoV-2 (NBC5, NBC14, NBC15 e NBC27). As interações dos ligantes NBC5, NBC14, NBC15 e NBC27 são mostradas nas Figuras A, B, C e D, respectivamente. **Fonte:** O Autor (2024).

6.5.3 Validação dos resultados das análises *in silico* do estudo I e II via *Machine learning*

A análise de triagem virtual de 171 mil compostos naturais por meio de métodos *in silico* (docking molecular, dinâmica molecular, *drug-likeness* e ADMET) permitiu a identificação de naringenina-4-O-glucoronido como sendo potencial fármaco para tratamento de COVID-19, via inibição da proteína *Spike (RBD)* da variante ômicron. Nós também identificamos o fitoquímico e outros três compostos (ZINC000008662732, ZINC000013374469 e ZINC000003022621) mostraram-se ser potenciais inibidores da proteína *Spike (RBD)* tipo selvagem. Todos esses quatro foram validados usando modelos de inteligência artificiais baseados em *machine learning* visando prever a atividade farmacológica (IC_{50}) desses fitoquímicos contra o próprio vírus SARS-CoV-2.

6.5.3.1 Análise do espaço químico: análise univariada

A **Tabela 6.10** mostra a análise descritiva de todos 10,057 compostos incluídos no estudo. A análise descritiva univariada desses compostos em relação aos descritores de Lipinski são mostrados na tabela 6.10. Nesta análise, podemos observar que a mediana dos valores do peso molecular foi menor que 500 (mediana=333.39), $\log P$ mediano =2,94, grupos aceptores de ligações de hidrogênio mediano foi igual 1 e aceptores de ligações de hidrogênio foi igual a 3. Esses resultados mostram que em média os 10,057 compostos incluídos no estudo tiveram características *drug-like*, ou seja, cumprem a regra dos cinco de Leminski ($PM < 500$, $\log P < 5$, $DLH < 5$ e $ALH < 10$). Importante também em destacar que a mediana da

bioatividade contra ARS-CoV-2 de toda população molecular foi em torno 1952,62 nM ($1,95 \cdot 10^{-6}$ M).

Tabela 6.10. Análise descritiva das variáveis da regra dos 5 de Lipinski de todos 10,057 compostos incluídos no estudo.

Parâmetro	PM	LogP	DLH	ALH	IC50 (nM)
média	368,69	2,75	1,95	4,94	46304,05
DP	216,12	2,43	2,91	3,74	1416861,73
min	59,06	-17,40	0,0	0,0	0,16
25%	260,31	1,47	1,0	3,0	19952,62
50%	339,39	2,94	1,0	4,0	19952,62
75%	433,03	4,24	2,0	6,0	20000,0
máx	4491,94	17,85	63,0	60,0	100000000,0

Nota: Legenda: PM: peso molecular, DLH: grupos doadores de ligações de hidrogênio, ALH: grupos aceptores de ligações de hidrogênio. **Fonte:** O Autor (2024).

A Figura 6.9 mostra os resultados da análise do espaço químico dos dados dos compostos usando os descritores de Lipinski. Podemos observar que, os valores de pIC50 dos compostos ativos, inativos e aqueles com bioatividade intermediária contra o vírus SARS-CoV-2 foram estatisticamente diferentes ($p = 1,6 \cdot 10^{-149}$). O threshold dos valores de pIC50 entre os dois grupos de compostos foi de 6, onde compostos ativos apresentam $pIC50 > 6$ e compostos inativos (ou com bioatividade intermediária) apresentam $pIC50 < 6$. Outras diferenças observadas entre os três grupos de compostos (ativos, inativos e bioatividade intermediária), foi em relação ao peso molecular ($p = 3,46 \cdot 10^{-23}$), $AlogP$ ($p = 0,00941$), grupos doadores ($p = 3,2 \cdot 10^{-25}$) e aceptores ($p = 1,2 \cdot 10^{-21}$) de ligações de hidrogênio (nHBDon). Esses resultados mostram que realmente o espaço químico entre os compostos com bioatividade e aqueles com bioatividade intermediária ou sem bioatividade contra o vírus SARS-CoV-2, são realmente diferentes. Esses resultados, são também condizentes com a literatura^{539,540}. Esses *insights* serviram de justificativa para o desenvolvimento de

modelos de machine learning baseados em QSAR para validação da bioatividade anti-SARS-CoV-2 daqueles compostos identificados pelos métodos de in silico nos tópicos anteriores do presente capítulo da tese.

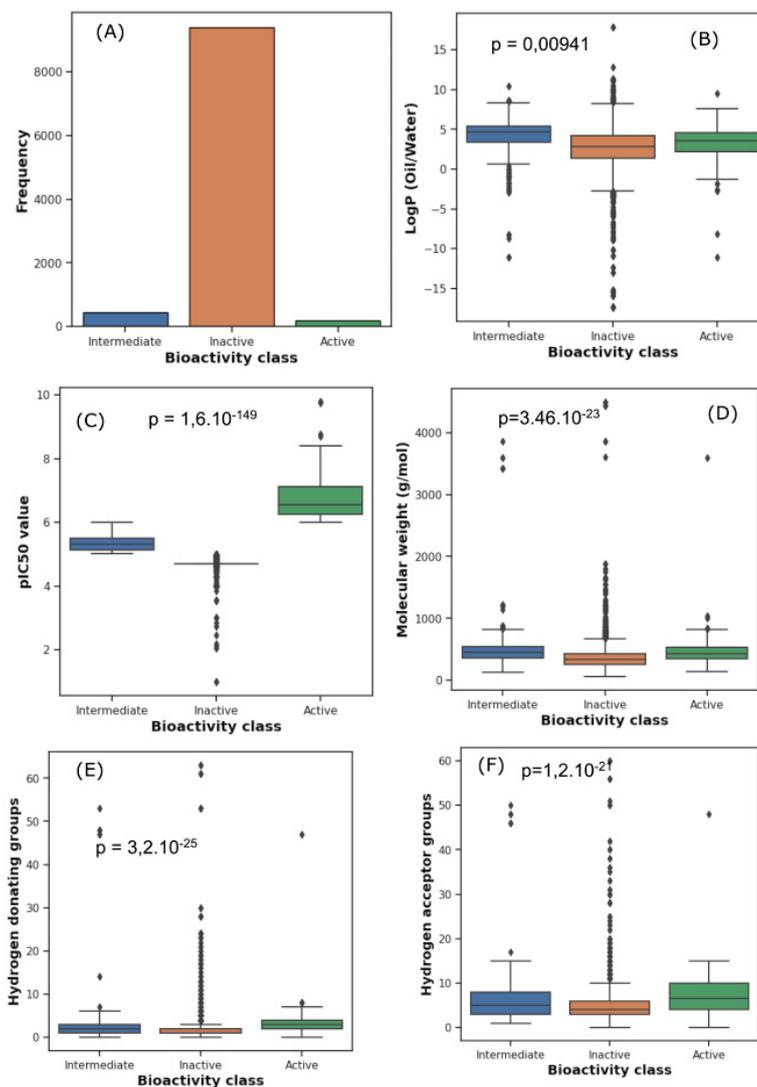


Figura 6.9. Análise do espaço químico usando os descritores de Lipinski. Em (A) são mostrados a classe dos compostos bioativos. Houve uma diferença significativa das classes de bioatividade segundo o LogP (B), pIC50 (C), peso molecular (D) grupos doadores (E) e aceptores de ligações de hidrogênio (F). **Fonte:** O Autor (2024).

6.5.3.2 Análise do espaço químico: análise multivariada

A análise do espaço químico foi também realizada usando o modelo PCA, com os dados pre-processados pelo método de *StandardScaler*. O PCA foi capaz de discriminar o grupo de compostos ativos ($IC_{50} < 100 \text{ nM}$), inativos ($IC_{50} > 10,000 \text{ nM}$) e aqueles com atividade intermediária (IC_{50} entre $100\text{-}10000 \text{ nM}$) contra o vírus SARS-CoV-2.

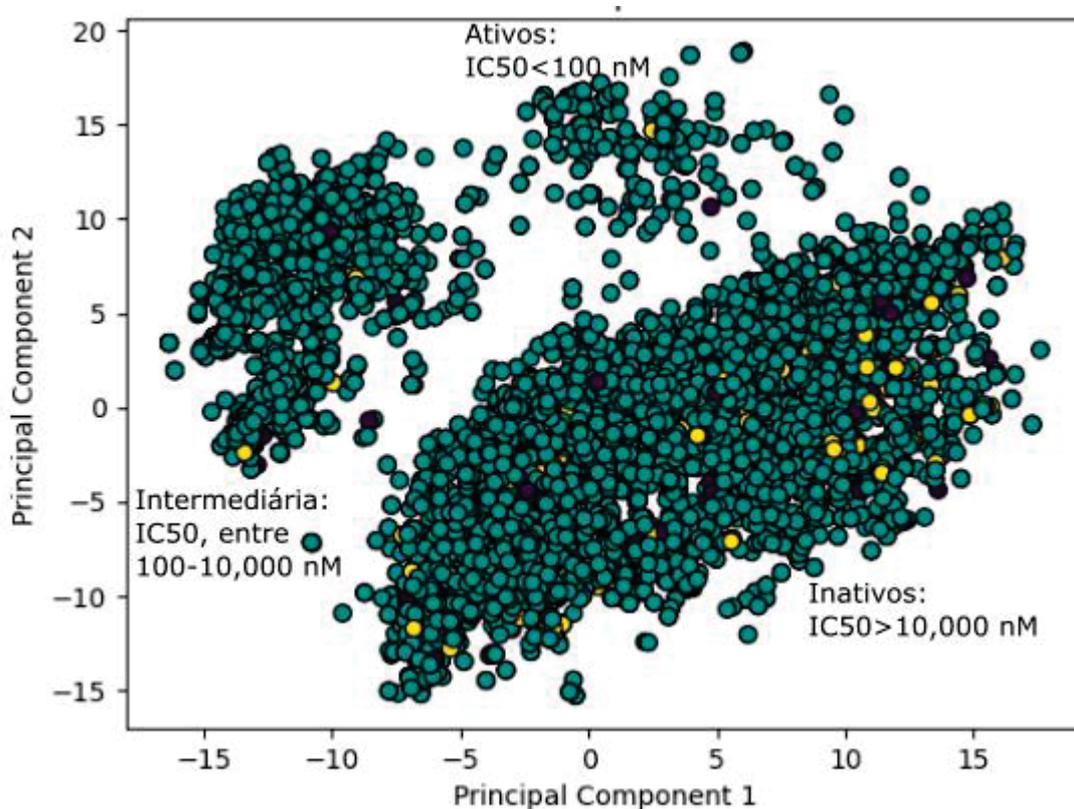


Figura 6.10. Modelo PCA. Discriminação entre os compostos ativos ($IC_{50} < 100 \text{ nM}$), inativos (IC_{50} entre $100\text{-}10000 \text{ nM}$) e com atividade intermediária (IC_{50} entre $100\text{-}10000 \text{ nM}$) contra o vírus SARS-CoV-2. **Fonte:** O Autor (2024).

6.5.3.3 Machine learning e QSAR

Um total de trinta e dois (32) diferentes algoritmos de machine learning foram treinados e validados, visando prever a bioatividade (pIC_{50}) a partir dos descritores moleculares de *PubChem*. Importante destacar que dos 881 descritores de *PubChem* calculados⁵⁴¹, apenas 204 deles foram selecionados para o treinamento dos modelos de machine learning, o que permitiu minimizar os problemas de multicolinearidade nos modelos de machine learning treinados⁵⁴²⁻⁵⁴⁴. Dos 32 algoritmos treinados e testados, foram selecionados os top três algoritmos com melhores desempenho preditivo, nomeadamente: *XGBoost*, *Random Forest*, *Histogram based Gradient Boosting*, *LGBM*, *Bagging* e *KNN*. A Tabela 6.11 mostra os resultados do desempenho dos top três modelos de *machine learning*, após terem os seus hiperparâmetros otimizados.

Tabela 6.11 Desempenho preditivo dos modelos de machine learning após avaliação da performance preditiva dos modelos top três de machine learning, após a otimização dos seus hiper-parâmetros

Modelo	R ²	MSE	RMSE	MAE
Random Forest	0,8950	0,6562	0,8101	0,5735
Histogram based Gradient Boosting	0,8081	0,6466	0,8041	0,5950
XGB regression	0,9276	0,8681	0,9317	0,6154

Nota: *MSE*: mean square error; *RMSE*: root mean square error; *MAE*: mean absolute error. **Fonte:** O Autor (2024).

A Figura 6.11 mostra as curvas dos valores experimentais e preditos da bioatividade (pIC_{50}) dos compostos contra a proteína o vírus SARS-CoV-2 para o tratamento potencial de COVID-19. Podemos observar que os valores de R² variaram entre 0,80-0,92 mostrando que todos cinco modelos de *machine learning* tiveram uma boa performance na predição da bioatividade contra o SARS-CoV-2 a partir da estrutura química das moléculas analisadas.

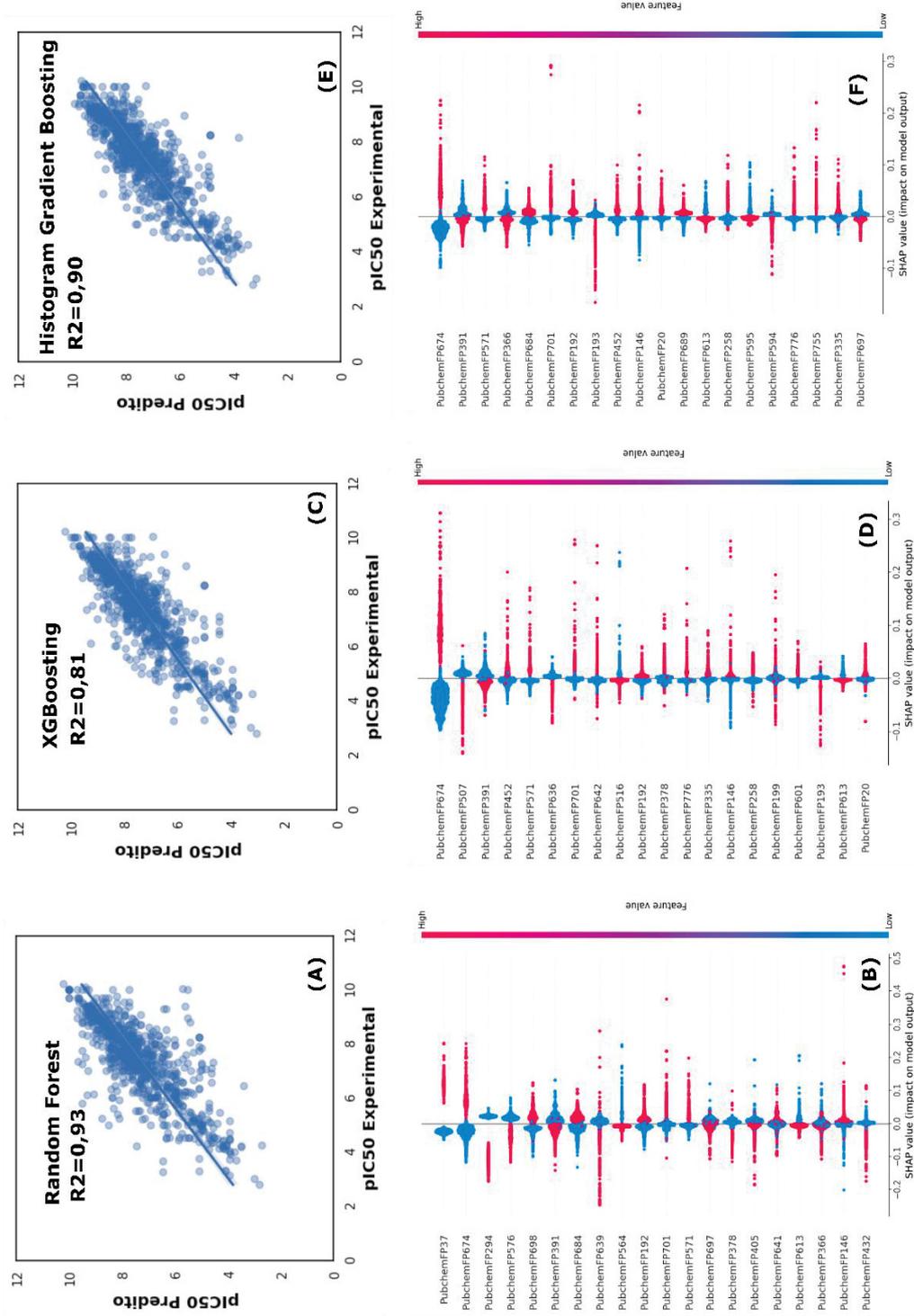


Figura 6.11. Em (A), (C) e (E) os valores da bioatividade experimental e predita dos modelos de *Random forest (RF)*, *Xgboost* e *Hisogradient Boosting (HGB)*, respectivamente. Em (B), (D) e (E) são mostradas as variáveis (grupos funcionais) mais importantes dos compostos na predição bioatividade contra o -SARS-CoV-2 nos modelos RF, Xgboost e HGB, respectivamente.

A tabela 6.12 mostra os valores de bioatividade (pIC50) predita dos cinco compostos promissores usando os algoritmos de *machine learning*, *Random Forest*, *XGBoost* e *Histogram gradient Boosting*. Importante destacar que conforme a análise exploratória dos descritores de lipinski, foram considerados compostos bioativos aqueles que apresentavam a bioatividade $Pic50 > 6$. Assim, os compostos que tiveram bioatividade significativa contra o vírus SARS-CoV-2 são ZINC000045789238 e ZINC000000704424 cujos valores de pIC50 médios de 6,35 (CV=1,14%) e 6,38(1.56%), respectivamente. Os fármacos controles remdesivir e molnupinavir também mostraram resultados promissores com seus valores de pIC50 médios sendo maiores que (tabela 6.12), portanto validando os nossos resultados. O restante dos compostos não teve bioatividade significativa ($pIC50 > 6$).

Tabela 6.12. Valores de bioatividade (pIC50) dos cinco compostos mais promissores identificados por dinâmica molecular e dos fármacos controles (remdesivir e molnupinavir) preditos pelos algoritmos Random Forest, XGBoost e Histogram gradient Boosting usando a abordagem QSAR-3D

Composto	Random Forest	XGBoost	Histogram gradient Boosting	Média	DP	CV(%)
ZINC000045789238	6,32	6,45	6,28	6,35	0,07	1,14
ZINC000000704424	6,41	6,49	6,25	6,38	0,10	1,56
ZINC000008662732	5,85	5,59	5,76	5,73	0,11	1,88
ZINC000013374469	5,63	5,49	5,70	5,61	0,09	1,56
ZINC000003022621	5,17	5,20	5,22	5,20	0,02	0,40
Remdesivir	6,88	6,87	6,91	6,89	0,02	0,25
Molnupinavir	7,20	6,99	6,95	7,05	0,11	1,56

Nota: DP: desvio padrão; CV: Coeficiente de variação. **Fonte:** O Autor (2024).

6.6 DISCUSSÃO

Este estudo avaliou 171 mil produtos naturais como potenciais candidatos a fármacos para o tratamento da COVID-19 pelo mecanismo de inibição da proteína *Spike* (RBD) do vírus SARS-CoV-2 tipo selvagem e da mutação ômicron, usando simulações de triagem virtual baseados em docking molecular, análise de drug-likeness, predições ADMET, simulações de dinâmica molecular (incluindo os cálculos de energia livres de ligação MM/PBSA e MM/GBSA). Na análise, o ligante naringenina-4'-O-glucuronídeo (ID ZINC000045789238) mostrou-se ser promissor na

inibição a RBD da variante omicron, ao passo que os ligantes ZINC000000704424, ZINC000008662732 e ZINC000013374469 forma promissores para o tratamento da COVID-19 via inibição da RBD tipo selvagem. Os resultados das análises foram validados por modelos de *machine learning* baseados em QSAR, onde demonstraram que os quatro ligantes foram capazes de prever a bioatividade contra o vírus SARS-CoV-2 dos quatro ligantes a partir das suas estruturas químicas. Segundo esses algoritmos, os valores de bioatividade (IC_{50}) contra SARS-CoV-2 dos quatro compostos variaram entre 10-83 nM.

6.6.1 Estudo I: Naringenina-4'-glicuronídeo como novo candidato a fármaco contra a variante Omicron da COVID-19: um estudo baseado em docking molecular, dinâmica molecular, MM/PBSA e MM/GBSA

A variante ômicron do SARS-Cov-2 é atualmente uma preocupação significativa na luta contra a COVID-19, dados os seus padrões mutacionais^{500,545}. Sabe-se que a proteína *Spike* (S1), especificamente a região *RBD*, é responsável pelo reconhecimento do vírus nas células hospedeiras (*ACE-2* humana)⁵⁴⁶. No caso da proteína *Spike* (S1) da variante ômicron, foram relatadas diversas mutações dos resíduos de aminoácidos da região *RBD* (Asp339, Asp339, Pro373, Phe375, Asn417, Lys440, Ser446, Asn477, Lys478, Ala484, Arg493, Ser496, Ar498, TyR501 e His505)^{547,548}. Além de proporcionar maior velocidade de disseminação ao vírus, essas mutações desafiam o desenvolvimento ou aprimoramento de vacinas e tratamentos que possam aumentar a imunidade dos indivíduos^{549,550}. Existem também evidências científicas que mostram vários casos de reinfecção por Omicron, sendo necessária a administração de vacinas de reforço⁵⁵⁰.

Vários estudos de docking e simulações de dinâmica molecular destinados a identificar novos inibidores contra diferentes proteínas alvo do SARS-CoV-2 estão disponíveis na literatura⁵⁵¹⁻⁵⁵³. Por exemplo, Gangadevi (2022) identificou o kobofenol A como um potencial inibidor da proteína *Spike RBD* (S1), onde os aminoácidos Glu375 e Thr347 estabeleceram ligações de hidrogênio com a proteína *Spike*⁵⁵³. Duas possíveis razões explicam as diferenças nos resíduos de aminoácidos envolvidos na interação do complexo proteína-ligante entre nosso estudo e o estudo de Gangadevi (2022): primeiro, porque Gangadevi (2022) utilizou a proteína *Spike* do

SARS-CoV- 2 tipos selvagens (PDB ID 6M0J) (Lan et al., 2020), ao contrário do nosso estudo onde utilizamos a proteína *Spike* do SARS-CoV-2 que sofreu mutação omícron (PDB ID 7T9L) ⁵⁴⁸. A explicação a seguir é que no estudo de Gangadevi (2022), o acoplamento e a dinâmica molecular dos ligantes foram feitos com o complexo de espícula RBD/ACE-2. Os ligantes foram ligados em um sítio diferente do RBD ⁵⁵³. Em contraste, em nosso estudo, isolamos a proteína *Spike ACE-2* e procuramos investigar ligantes que inibem aminoácidos específicos de *RBD* envolvidos na ligação de *ACE-2* (Asp339, Asp339, Pro373, Phe375, Asn417, Lys440, Ser446, Asn477, Lys478 , Ala484, Arg493, Ser496, Ar498, Tyr501 e His505) ⁵⁴⁸. Por outro lado, os aminoácidos RBD identificados no nosso também foram encontrados no estudo de Kumar (2021), onde ensaios de dinâmica molecular do medicamento favipiravir mostraram resultados promissores na inibição do RBD da proteína *Spike* do aminoácido bloqueador do SARS-CoV-2, ácidos Arg408 (HB), Gln409 (hidrofóbico) por ligações de hidrogênio e interações hidrofóbicas, respectivamente ⁵⁵⁴. Além disso, no estudo de Kumar (2022), o ácido isoclorogênico apresentou maior afinidade de ligação com os aminoácidos Tyr453 e Tyr453 do RBD via ligações de hidrogênio ⁵⁵⁴.

Gostaríamos também de ressaltar que realizamos análises de docking molecular e simulações de dinâmica molecular utilizando remdesivir como medicamento de referência, e os resultados encontrados foram semelhantes aos encontrados por nossos quatro ligantes promissores (Naringenin-4'-O-Glucuronide, ergoloid ohioensin A e prunetrin).

Uma forma de garantir a confiabilidade e a precisão dos resultados do acoplamento é realizar o acoplamento consensual utilizando diferentes softwares ^{555,556}. Usando três softwares diferentes (*PyRx*, ferramentas *AutoDock* e *AutoDockVina*), pudemos fornecer resultados confiáveis, com valores de energia de ligação relatados dos quatro principais ligantes <2%. Um total de nove conformações (poses) foram geradas, e selecionamos aquela com maior estabilidade de ligação (ou seja, valores *RMSD* <2,0), o que revela a precisão dos resultados de docking.

Interações não covalentes, como ligações de hidrogênio, desempenham um papel essencial no reconhecimento do ligante ao alvo molecular e na estabilização da ligação do complexo alvo-ligante. Além disso, muitos estudos quantitativos de relações estrutura-atividade (*QSAR*) demonstram que as ligações de hidrogênio são essenciais na modelagem de uma atividade alvo específica ⁵⁵⁷. Em nosso estudo, a análise das interações de docking mostrou que o ligante naringenina-4'-O-

glicuronídeo (ID ZINC000045789238) produziu o maior número de ligações de hidrogênio ($n = 6$) com resíduos de aminoácidos na região *RBD* da proteína *Spike* (S1) da mutação ômicron (Ser494, Ser496, Thr500, Thr5051 e Thr505). Prunetrina (ID ZINC000008662732) foi o ligante com o segundo maior número de ligações de hidrogênio (Arg408, Asp417, Ser496 e His505).

Um gráfico de RMSD versus tempo de simulação é comumente usado para estabelecer o período de equilíbrio, estabilidade e qualidade de ligação do complexo molecular ligante-alvo em análises de DM. No geral, em nosso estudo, todos os quatro ligantes formaram um complexo ligante-proteína com valores médios de RMSD inferiores a 0,3 nm ao longo do tempo de dinâmica de 100 ns, o que é considerado aceitável^{130,558,559}.

O *RMSF* é utilizado para analisar a flutuação (ou estabilidade) de átomos (ou grupos de átomos) do complexo receptor apo-ligante e dos complexos proteína-ligante em geral, como pode ser observado em alguns estudos da literatura envolvendo o complexo *RBD* do pico proteico do SARS-CoV-2 e ligantes⁵⁶⁰⁻⁵⁶². Em nosso estudo, os valores de *RMSF* também foram determinados para compreender a estabilidade/flutuações dos quatro complexos^{563,564}. Embora todos os resíduos de aminoácidos da glicoproteína *Spike* (S1) apresentassem flutuações dentro do limite tolerado ($RMSF < 1,3$ nm), como esperado⁵⁶⁵, os resíduos da mutação ômicron envolvidos na interação com o ligante (Phe375, Lys440, Lys478 e Phe375) apresentaram flutuações importantes. Estudos anteriores mostraram que esses resíduos mutantes são responsáveis por transições dinâmicas de ordem-desordem durante a ligação da proteína *Spike* (S1) com *ACE-2* humana, o que pode explicar os números obtidos em nossas análises⁵⁴⁸.

Em nosso estudo, todos os quatro complexos apresentaram estabilidade na acessibilidade ao solvente e compatibilidade com o complexo proteína-ligante ao longo do período de simulação dinâmica (100 ns)⁵⁶⁵, conforme demonstrado no gráfico SASA. O complexo naringenina-4'-O-glicuronídeo-espiga (S1) apresentou o maior número de ligações de hidrogênio cujos resultados se correlacionam com os das análises de docking molecular.

MM/PBSA e *MM/GBSA* são dois métodos popularmente utilizados em simulações de DM por possuírem alta precisão no cálculo da energia livre de ligação entre um ligante e alvo molecular em toda a trajetória de DM, sendo fortemente recomendados por cientistas da área⁵⁶⁶. Em nossas análises, apenas o complexo

naringenina-4'-O-glicuronídeo-*Spike* (S1) apresentou resultados (-) para energias livres de ligação de *MM/PBSA* e *MM/GBSA*, mostrando maior estabilidade de ligação com a proteína *Spike* (S1) e, portanto, sendo um candidato promissor a medicamento contra infecções por SARS-CoV-2. Por outro lado, embora os ligantes restantes, ZINC000003995616, ZINC000004098448 e ZINC000008662732, tivessem valores de equilíbrio dinâmico aceitáveis ($RMSD < 0,3$ nm), eles não devem ser considerados candidatos a medicamentos para esta indicação devido à falta de estabilidade de ligação à proteína *Spike* ($MM/PBSA > 0$).

A naringenina é um fitoquímico da classe dos flavonoides conhecido por ter potente atividade antiviral e imunomoduladora, inclusive contra o vírus SARS-Cov-2^{567,568}. No entanto, não foram encontrados estudos na literatura relativos à atividade contra SARS-Cov-2 do composto naringenina-4'-O-glicuronídeo identificado em nossa pesquisa. No entanto, como a naringenina-4'-O-glicuronídeo é um derivado da naringenina, provavelmente tem ação contra o SARS-CoV-2. Estudos *in vitro* e *in vivo* são necessários para consolidar nossos achados.

O alto custo computacional na realização de simulações de DM foi a principal limitação deste estudo. Embora tenhamos utilizado 100 ns para as simulações de DM, conforme geralmente recomendado na literatura⁵⁶⁹⁻⁵⁷², reconhecemos que um aumento no tempo DM poderia gerar mais informações, como o enovelamento de proteínas e o comportamento da estrutura secundária (hélice alfa e folhas beta pregueadas).

6.6.2 Estudo II: Triagem virtual, docking molecular, análise de drug-likeness e predições ADMET: proteína *Spike* (RBD) do SARS-CoV-2 tipo selvagem

O composto NBC5 é comumente conhecido como feselol, um produto natural encontrado em plantas do gênero *Ferula* (por exemplo, *Ferula gummosa* Boiss. e *Ferula galbaniflua* Boiss.)⁵⁷³. No entanto, os estudos sobre as atividades biológicas desta substância são limitados na literatura. Estão disponíveis apenas estudos *in vitro* mostrando efeitos antimicrobianos contra *P. aeruginosa*, *S. epidermidis* e *S. aureus* e atividade antiparasitária contra *P. falciparum*^{574,575}. Estudos *in vitro* mostram que a combinação de feselol com antineoplásicos potencializou os efeitos anticancerígenos; isso pode ser explicado devido à sua capacidade de inibir a glicoproteína P, que é a principal proteína responsável pelo mecanismo de resistência de muitos fármacos

(inclusive anticancerígenos), favorecendo o aumento das taxas de absorção e biodisponibilidade desses fármacos, e conseqüentemente obtendo a atividade terapêutica desejada ^{576,577}. Não foram encontrados estudos sobre as atividades antivirais do feselol. Da mesma forma, até onde sabemos, não existem estudos que avaliem os efeitos dos compostos naturais NBC14, NBC15 e NBC27, sugerindo a necessidade de avaliações adicionais das atividades antivirais biológicas destas substâncias naturais.

Do ponto de vista molecular, a atividade antiviral dos compostos fitoquímicos NBC5, NBC14, NBC15 e NBC27 pode ser justificada pelo fato de possuírem grupos carboxila e grupos hidroxila cujos átomos de oxigênio e hidrogênio se ligam intermolecularmente, por ligações de hidrogênio com os resíduos dos aminoácidos da proteína *Spike* (S1) do SARS-CoV-2, a saber, Gly482, F454, Arg454, N343, Ser373, W436, N437, Thr430, Arg355, F456, E471 e Arg454 ⁵⁷⁸.

A determinação estrutural dos ligantes em seus estados estereoisoméricos e tautoméricos tanto em pH fisiológico (pH = 7,4) é muito importante em um estudo de docking, pois simularia as condições do organismo humano ⁵⁷⁹. Esta análise é viável apenas em situações em que o número de ligantes acoplados é pequeno, pois essa determinação é realizada manualmente, fazendo um ligante por vez através de software específico (por exemplo, *Chemdraw* ou *Marvinsketch*) ^{580,581}. Em situações em que existem milhares (ou mesmo milhões) de moléculas a serem docados, que é o caso do nosso estudo (utilizamos 171 mil ligantes), não é viável realizar a determinação de ligantes em pH fisiológico (pH = 7,4), devido ao grande volume de ligantes existentes na base de dados, inclusive existem vários estudos semelhantes na literatura ⁵⁸²⁻⁵⁸⁴. Nessa situação de maior número de ligantes (milhares ou milhões de ligantes), a primeira tarefa a ser realizada é a triagem virtual, que é um processo que consiste em investigar quais os compostos têm uma maior afinidade de ligação com o alvo molecular. Após a identificação dos ligantes com maior afinidade com o alvo molecular (que geralmente são em pequeno número), é então realizada uma análise de docking utilizando critérios muito rigorosos, que incluem a determinação das estruturas dos ligantes em pH = 7,4, a influência da tautomeria, estereoisomeria, semelhança com drogas, entre outros. Em nosso estudo, seguimos a mesma estratégia; onde inicialmente realizamos uma triagem virtual de 171 mil ligantes, na qual selecionamos 34 compostos. Dos 34 ligantes foram determinadas suas estruturas em pH = 7,4, seus estados estereoisomérico e tautomérico e em seguida

realizada uma nova análise de docking consensual (usando dois programas *AutoDock Tools* e *AutoDock Vina*) e *machine learning* para validar os resultados. Demonstramos adicionalmente através de estados estereoisoméricos, tautoméricos e de protonação (em pH fisiológico) ^{585,586}, que NBC5, NBC14, NBC15 e NBC27 poderiam ser usados como drogas racêmicas, o que significa que nenhuma tecnologia avançada é necessária para o isolamento de estereoisômeros. Medicamentos enantiomericamente puros (por exemplo, naproxeno, labetalol, varfarina) exigem altos custos para desenvolvimento e tecnologias de produção (enantiômero puro com atividade biológica), o que pode ser uma barreira importante na maioria dos países ⁵⁸⁷.

A toxicidade renal é uma das toxicidades muito importantes que devem ser avaliadas em compostos que possuem maior solubilidade em água, como é o caso dos derivados cumarínicos identificados em nosso estudo (por exemplo, feselol) ⁵⁸⁸. No entanto, as poucas plataformas online de *machine learning*, disponíveis na literatura que realizam essas previsões apresentam baixa precisão preditiva. Além disso, existem apenas alguns estudos recentes que desenvolveram modelos de *machine learning* para prever a toxicidade renal com maior precisão, mas os autores não desenvolveram uma aplicação para os modelos serem utilizados na prática, como pode ser visto no recente estudo de Gong ⁵⁸⁸. Estas foram as razões pelas quais não foi possível fazer previsões in silico da toxicidade renal dos compostos cumarínicos identificados neste estudo. Isso constituiu uma das limitações do nosso estudo.

6.7 CONCLUSÃO

A pesquisa realizada neste capítulo VI da tese de doutorado empregou uma abordagem abrangente, utilizando uma variedade de técnicas de triagem para avaliar compostos naturais em potencial para o tratamento da COVID-19. As análises de docking molecular, previsões ADME, simulações de dinâmica molecular e o emprego de algoritmos de machine learning destacaram candidatos promissores, tanto para a variante ômicron quanto para o tipo selvagem do SARS-CoV-2.

Os resultados revelaram o naringenina-4'-O-glicuronídeo como um composto particularmente promissor para inibir os resíduos de aminoácidos mutantes na *RBD* da variante Omicron, enquanto o feselol), NBC14, NBC15 e NBC27 mostrou-se eficaz na inibição da *RBD* do tipo selvagem.

A contribuição dos algoritmos de *machine learning*, como Random Forest, XGBoost e Histogram Gradient Boosting, foi crucial na identificação desses cinco ligantes com atividade farmacológica contra o vírus SARS-CoV-2.

No entanto, é importante ressaltar a necessidade de estudos complementares, tanto *in vitro* quanto *in vivo*, para validar essas descobertas. A realização desses estudos aprofundará nossa compreensão da eficácia e do potencial terapêutico desses compostos naturais, fortalecendo assim o desenvolvimento de medicamentos eficazes para o combate à COVID-19.

**7 CAPÍTULO VII - ABORDAGEM INTEGRATIVA REVELA CANDIDATOS
MULTIFACETADOS A FÁRMACOS PARA O TRATAMENTO SIMULTÂNEO DE
COVID-19, HEPATITE, DENGUE E HIV USANDO INTELIGÊNCIA ARTIFICIAL
MULTI-TARGET, E POLIFARMACOLOGIA**

7.1 RESUMO

Neste estudo, desenvolvemos modelos de *machine learning* baseados em QSAR para prever a bioatividade de compostos contra vírus SARS-CoV-2, hepatite B, hepatite C, dengue e HIV. Utilizamos um conjunto de dados de 19765 compostos da base ChEMBL, treinando 40 algoritmos de ML e avaliando seu desempenho. Os top cinco melhores modelos foram aplicados ao método *SHAP values* para entender os descritores relacionados à bioatividade *multi-target*. Triamos um banco de dados externo (*Human Metabolome Database*, n=220 mil compostos) para identificar compostos bioativos, os quais foram submetidos a análises de docking e simulações de dinâmica molecular. Os resultados destacam a eficácia dos modelos, identificando três compostos promissores: nummularine B, telaprevir e entecavir, este último já aprovado pelo FDA para hepatites B e C. Propomos o reposicionamento desses fármacos para COVID-19 e dengue. A descoberta do nummularine B como novo candidato a fármaco *multi-target* é relevante. Os modelos baseados em árvores de decisão mostraram vantagens na predição de bioatividade. Os mecanismos de ação propostos fornecem *insights* importantes. Este estudo destaca o potencial da integração de *machine learning* na descoberta de fármacos *multi-target*, oferecendo uma abordagem promissora para terapias eficazes contra doenças virais.

Palavras-chave: *Machine learning*, QSAR, vírus, COVID-19, hepatite B, hepatite C, dengue, HIV, bioatividade, docking molecular, dinâmica molecular, FDA, reposicionamento de fármacos

7.2 INTRODUÇÃO

A estratégia Global do Sector da Saúde da OMS (OD3 3.3) apelou a eliminação das epidemias de doenças infecciosas até em 2030, incluindo HIV/AIDS, hepatite, dengue e COVID-19, através de prevenção, testes e tratamento aumentados ^{589,590}. A descoberta de novos tratamentos eficazes e abrangentes para HIV/AIDS, hepatite, dengue e COVID-19 faz parte dos objetivos da OD3 3.3 da OMS. Isso inclui o desenvolvimento de terapias que não apenas combatam uma única doença, mas que também possam direcionar múltiplos patógenos simultaneamente, abordando a interconexão e a sobreposição de sintomas e complicações entre essas enfermidades ⁵⁹⁰.

Em relação ao HIV, uma doença ainda sem cura, a OMS estima que até em 2023, o HIV tinha ceifado 40,4 milhões [32,9–51,3 milhões] de vidas, com transmissão contínua em todos os países do mundo; com alguns países a reportar tendências crescentes de novas infecções quando anteriormente estavam em declínio. No final de 2022, havia cerca de 39,0 milhões [33,1–45,7 milhões] de pessoas que viviam com HIV, dois terços das quais (25,6 milhões) vivem na Região Africana da OMS. Em 2022, 630 000 [480 000–880 000] pessoas morreram de causas relacionadas com o HIV e 1,3 milhões [1,0–1,7 milhões] de pessoas contraíram o HIV ⁵⁹⁰.

Por outro lado, as hepatites B e C e a dengue são doenças virais que fazem parte de uma lista de vinte doenças consideradas negligenciadas que afetam principalmente populações em países de baixa e média renda. Elas recebem esse rótulo porque historicamente têm recebido menos atenção em termos de pesquisa, desenvolvimento de tratamentos e investimento em comparação com outras doenças mais visíveis globalmente ⁵⁸⁹. Segundo a OMS estima que globalmente 328 milhões de pessoas estejam cronicamente infectadas pelo HBV ou pelo HCV, e cerca de 1,3 milhões de pessoas morreram por hepatite em todo mundo, dos quais 720000 ocorreram em o estágio de cirrose e em sua maioria eram pessoas não tratadas ^{591,592}. A dengue, uma doença viral também sem nenhum medicamento eficaz para sua cura, a OMS estima que 100 a 400 milhões de infecções ocorrendo a cada ano ao nível mundial e cerca da metade da população mundial corre agora o risco de contrair dengue ⁵⁹⁰. Por fim, a COVID-19 que continua sendo um problema de saúde pública, apesar de grandes avanços na imunização, com a OMS estimando 773 milhões de

casos de COVID-19 e 7 milhões de mortes em todo mundo²⁶. Isso mostra que a busca por soluções terapêuticas eficazes e abrangentes diante de múltiplas doenças virais continua mandatório e representa um desafio crucial na medicina moderna⁵⁹³.

A intersecção entre COVID-19, Hepatite, dengue e HIV destaca a necessidade premente de estratégias inovadoras que possam abordar essas enfermidades de forma simultânea e eficaz. Neste contexto, o uso combinado de técnicas avançadas de *machine learning*, polifarmacologia e metabolômica emergem como ferramentas promissoras para identificar novos candidatos a medicamentos capazes de alvejar múltiplos alvos patológicos^{594,595}. Este estudo propõe uma abordagem integrativa e pioneira, explorando a interdisciplinaridade dessas áreas para revelar potenciais compostos farmacológicos que possam revolucionar o tratamento dessas enfermidades, oferecendo perspectivas inovadoras para o enfrentamento dessas grandes questões de saúde global.

7.3 OBJETIVOS

7.3.1 *Objetivo geral:*

- Investigar a eficácia de modelos de machine learning baseados em QSAR na predição da bioatividade de compostos contra vírus SARS-CoV-2, hepatite B, hepatite C, dengue e HIV, visando identificar candidatos multifacetados a fármacos para o tratamento simultâneo dessas doenças.

7.3.2 *Objetivos específicos:*

- Desenvolver modelos de machine learning utilizando dados experimentais da base ChEMBL para prever a bioatividade de compostos contra os vírus mencionados;
- Avaliar o desempenho dos modelos treinados, utilizando métricas como R^2 , MSE, RMSE e MAE;
- Aplicar o método SHAP values nos melhores modelos para identificar os descritores relacionados à bioatividade *multi-target*;

- Triar o banco de dados externo (*Human Metabolome Database*) para identificar compostos bioativos e realizar análises de docking e simulações de dinâmica molecular;
- Identificar compostos promissores com potencial terapêutico para o tratamento simultâneo das doenças estudadas;
- Propor o reposicionamento dos fármacos identificados para o tratamento de COVID-19, dengue, hepatite e dengue;
- Investigar o potencial do nummularine B como novo candidato a fármaco multi-target;
- Propor mecanismos de ação para os compostos identificados, fornecendo *insights* importantes para o desenvolvimento de terapias eficazes contra doenças virais.

7.4 MATERIAL E MÉTODOS

7.4.1 Fluxograma do estudo

A Figura 7.1 ilustra o fluxo de execução deste estudo. Em síntese, desenvolvemos modelos de *machine learning multi-target* baseados em QSAR, metabolômica e polifarmacologia, com o propósito de prever e analisar compostos bioativos para tratamento simultâneo de HIV, COVID-19, dengue, hepatite B e hepatite C. O estudo seguiu o guia da Organização para a Cooperação e Desenvolvimento Econômico (OCDE), abrangendo as etapas: (i) compilação de um conjunto de dados com um ponto final; (ii) análise exploratória desses dados; (iii) aplicação de diferentes algoritmos de *machine learning* supervisionado; (iv) utilização de métricas para avaliar o desempenho dos modelos de *machine learning*; (v) interpretação mecanística dos modelos (análise dos descritores de Lipinski); (vi) aplicação dos algoritmos em um banco de dados externo. (vii) validação dos ligantes *multi-target* via modelagem por docking molecular e simulações de dinâmica molecular, incluindo o cálculo das energias livre de ligação MM/GBSA. Para assegurar transparência e reprodutibilidade dos modelos de *machine learning* (QSAR), disponibilizamos todos os arquivos *Jupyter Notebook* do *Python* utilizados, assim como o banco de dados, no seguinte link do GitHub.

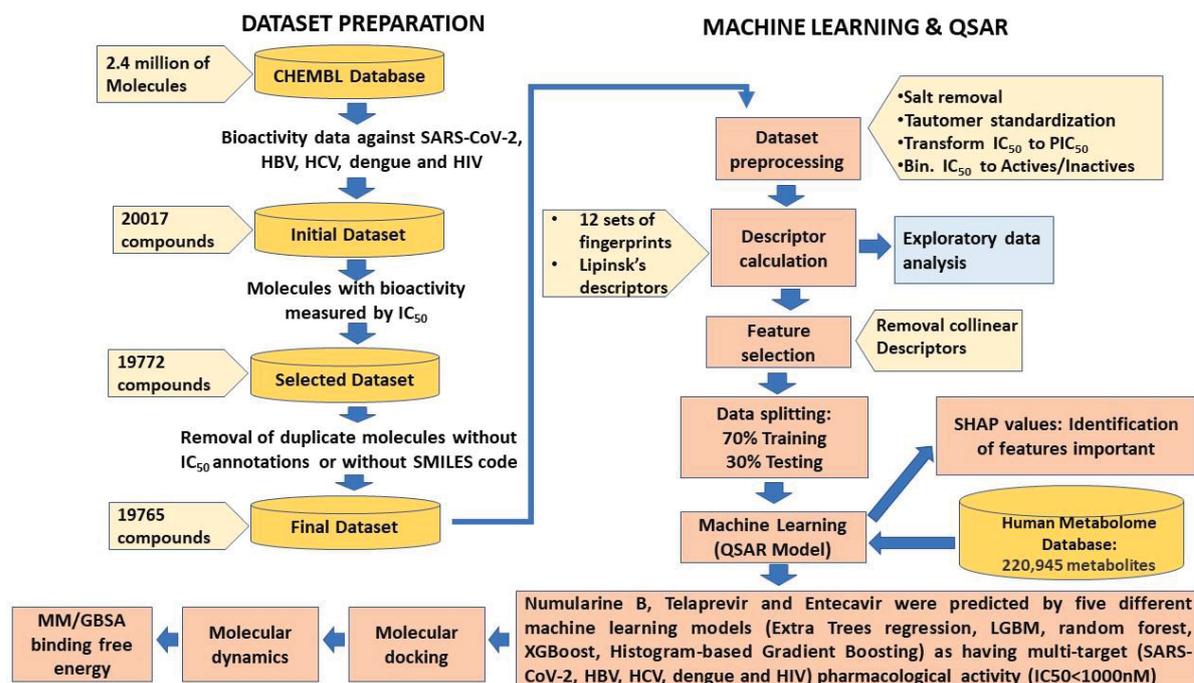


Figura 7.1. Fluxograma usado para coleta e limpeza dos dados, treinamento e validação dos modelos de machine learning baseados em QSAR para predição de candidatos a fármacos com bioatividade multi-target (SARS-CoV-2, HBV, HCV, Dengue e HIV). **Fonte:** O Autor (2024).

7.4.2 Descrição do banco de dados utilizado para Machine learning: ChEMBL Database

Os conjuntos de dados utilizados neste estudo para desenvolver modelos de *machine learning* para prever compostos que exibem bioatividade simultânea contra os vírus HIV, SARS-CoV-2, dengue, hepatite B e C foram provenientes do banco de dados ChEMBL 32 (<https://www.ebi.ac.uk/chembl/>). ChEMBL é um banco de dados britânico acessível ao público que contém dados de bioatividade de mais de 2,4 milhões de compostos que possuem propriedades semelhantes a medicamentos, espalhados por 211 conjuntos de dados. Este banco de dados consolida informações químicas, dados de bioatividade e dados genômicos para facilitar a tradução de informações genômicas em potenciais novos medicamentos. Além disso, o ChEMBL incorpora aproximadamente 86.000 publicações científicas, 1,5 milhão de ensaios, dados para 15.000 alvos terapêuticos, 6.700 mecanismos de ação de medicamentos, 2.000 células, 43.000 indicações de medicamentos e 759 tecidos biológicos. É crucial observar que todos esses dados passam por uma meticulosa seleção manual e curadoria por especialistas do domínio.

Os conjuntos de dados experimentais do *ChEMBL* abrangendo o conjunto de dados de compostos que inibem HIV ($n = 2.961$), SARS-CoV-2 ($n = 10086$), dengue (1.237), hepatite B ($n = 3063$) e hepatite C ($n = 2670$) foram combinados em um conjunto de dados consolidado de 20017 compostos e cada um desses compostos tinham 46 dados diferentes dados de bioatividade.

Todas os códigos *SMILES* dos compostos foram curadas usando em linguagem *Python*. O conjunto de dados inicial continha vários parâmetros de bioatividade, tais como EC_{50} , MIC, porcentagem de atividade, K_i , porcentagem de inibição, IC_{50} etc. O IC_{50} (medido na escala nM) foi selecionado para investigação adicional, pois é o parâmetro de bioatividade que estava disponível na maioria dos compostos ($n = 20017$ compostos). Posteriormente, foi realizada a remoção de compostos duplicados, compostos sem anotações IC_{50} ou compostos que não possuíam código *SMILES*, tendo sido eliminados 245 compostos e o *dataset* final foi formado por 19772 compostos bioativos. Assim, o conjunto de dados final foi formado por 19772 compostos bioativos, que foram utilizados para a próxima etapa do estudo, o pré-processamento dos dados.

7.4.3 Pré-processamento dos dados para machine learning

O pré-processamento do conjunto de dados final ($n=9772$) de compostos bioativos contra os vírus HIV, SARS-Cov-2, dengue e hepatites B e C, foi feito utilizando a biblioteca *PadelPy* na linguagem *Python* (<https://pypi.org/project/padelpy/>). Este processo consistiu na remoção de sais, padronização de tautômeros, conversão de IC_{50} em pIC_{50} (escala logarítmica). A categorização dos valores de IC_{50} também foi realizada em três grupos de compostos: compostos ativos: $IC_{50} < 100$ nM; compostos com atividade intermediária: IC_{50} entre 100-1000 nM; compostos inativos, $IC_{50} > 1000$ nM ⁵²⁹. Após esta etapa, foram calculados os descritores de impressão digital (usando descritores *PubChem*) das moléculas. Eles consistem em um conjunto de códigos binários 881 que descrevem a impressão digital e o espaço químico 3D de cada molécula específica. Existem 12 tipos diferentes de descritores de impressão digital (por exemplo, *PubChem*, *CDK*, subestrutura etc.) ⁵³⁰. Neste estudo, a impressão digital das moléculas foi calculada utilizando o descritor *PubChem*, que é o mais utilizado para estudos de *machine learning* baseados em QSAR ⁵³¹.

Descritores adicionais foram calculados para aderir à regra dos cinco de Lipinski. Esses descritores abrangem peso molecular, contagem de aceitadores de ligações de hidrogênio (nHBAcc), contagem de doadores de ligações de hidrogênio (nHBDon) e o logaritmo do coeficiente de partição octanol/água (AlogP). De acordo com as recomendações da *Pfizer*, para que um composto químico possua propriedades semelhantes às de um medicamento, devem ser atendidos critérios específicos: $PM \leq 500$ g/mol; $AlogP < 5$; $nHBDon < 10$ e $nHBAcc < 5$. Para calcular esses descritores, foi utilizada a biblioteca *RDKit* em *Python* (<https://www.rdkit.org/>)⁵³².

7.4.4 Seleção de recursos

A multicolinearidade, uma preocupação significativa em modelos de regressão, refere-se à intercorrelação entre descritores (recursos), levando a um viés amplificado e à complexidade do modelo no *machine learning*. Para resolver esse problema, um filtro de variância foi aplicado, removendo descritores (recursos) com variabilidade limitada (variância $< 0,1$) do conjunto de dados. O objetivo foi obter um subconjunto condensado de descritores *PubChem*. Todo esse processo foi implementado utilizando a linguagem de programação *Python* (Figura 7.1)⁵³³.

7.4.5 Divisão de dados de treinamento e de teste

A divisão de dados para *machine learning* empregou o método Holdout para reduzir preconceitos na seleção de dados de treinamento, teste e validação (Figura 7.1). Este método segmentou o conjunto de dados de 19.772 compostos bioativos em duas fases: alocando 70% para treinamento de modelo e reservando 30% para fins de teste⁵³⁴.

7.4.6 Inteligência artificial e machine learning

Todo o projeto, incluindo a fase de *machine learning*, foi executado em *Python*. Modelos de regressão foram implantados para prever a atividade biológica (expressa como pIC50) simultaneamente contra os vírus HIV, SARS-CoV-2, dengue e hepatite B e C, com base na estrutura química de 19.772 compostos bioativos representados pelos descritores *PubChem*. Nesta análise, a variável resposta foi pIC50, enquanto os

descritores de impressão digital *PubChem* serviram como variáveis preditoras. As bibliotecas *Scikit-Learn* e *LazyPredict* facilitaram a construção de 40 algoritmos distintos de *machine learning*, com o objetivo de selecionar os cinco principais modelos que apresentam desempenho preditivo superior. A avaliação dos modelos de ML desenvolvidos contou com diversas métricas, incluindo coeficiente de determinação (R^2), erro quadrático médio (*MSE*), raiz do erro quadrático médio (*RMSE*) e erro médio absoluto (*MAE*)^{535,536}. R^2 mede a capacidade preditiva do modelo de regressão numa escala de zero a um; mais próximo de um significa maior precisão preditiva. Por outro lado, *MAE*, *MSE* e *RMSE* são métricas de erro, onde um modelo de regressão bem ajustado visa minimizar seus valores⁵³⁵.

7.4.7 Investigação dos descritores moleculares mais importantes na atividade biológica multi-target usando SHAP values

A análise de recursos cruciais nos modelos de *machine learning* foi conduzida por meio do método de *SHAP values* (*SHapley Additive exPlanations*). Inicializando com o objeto 'explicador', este método possibilitou o escrutínio de indivíduos ou grupos de compostos bioativos. A etapa subsequente envolveu o cálculo dos valores médios de cada recurso por meio da função SoftMax. A avaliação do impacto global dos descritores (recursos) de impressão digital *PubChem* foi alcançada por meio de várias visualizações, incluindo gráficos SHAP, gráficos Beeswarm, gráficos de resumo e gráficos de violino. Além disso, para discernir as partes específicas de uma molécula (características) que influenciam significativamente ou não a atividade biológica (*pIC50*), foram gerados gráficos de barras e gráficos em cascata^{537,538}.

7.4.8 Validação dos resultados de machine learning multi-target via modelagem por docking molecular e simulações de dinâmica molecular

Os compostos que tiveram resultados promissores como candidatos a fármacos multi-target (bioatividade contra os vírus SARS-CoV-2, HBV, HCV, dengue e HIV) foram submetidos nas análises de docking molecular e simulações de dinâmica molecular (incluindo os cálculos de energias livres de ligação *MM/GBSA*, com vista a obter *insights* sobre sua afinidade de ligação com as diferentes proteínas desses diferentes vírus. O tipo e a descrição das proteínas alvos de cada vírus, que foram

usadas para investigar a sua afinidade com os candidatos a fármacos *multi-target* são mostrados na tabela 7.1.

Tabela 7.1. Descrição das características e funções biológicas das proteínas alvos dos vírus SARS-CoV-2, dengue, hepatite B, hepatite C e HIV

Organismo	Descrição das proteínas alvos e dos ligantes	
Virus SARS-CoV-2	Alvo	3CL-Pro, 3CLP, Main Protease, MPro
	PDB ID	8EYJ
	Artigo científico	DOI: 10.1038/s41467-023-37035-5
	Importância Biológica da proteína alvo	A MPro é essencial para a replicação do SARS-CoV-2, clivando proteínas virais para formar proteínas necessárias à replicação e montagem do vírus
	Ligante controle (padrão)	Nirmatrelvir(NIR) Inibidor de 3C-like protease (3CLpro)
Vírus da hepatite B - HBV	Alvo	NS5B, Complex Polymerase
	PDB ID	4TN2
	Artigo científico	https://doi.org/10.1016/j.bmcl.2014.06.031
	Padrão	Entecavir
	Importância Biológica da proteína alvo	A complexa polimerase desempenha um papel fundamental na replicação do vírus HBV. Essa enzima, também conhecida como DNA polimerase viral, é responsável por catalisar a síntese do DNA viral a partir de um RNA intermediário. Especificamente, a enzima realiza a transcrição reversa do RNA viral em DNA de cadeia dupla, que é então incorporado ao genoma do hospedeiro
Vírus da hepatite C - HCV	Alvo	HCV NS5B RNA Polymerase/RNA-Directed RNA Polymerase
	PDB_ID	3VQS
	Artigo científico	DOI: https://doi.org/10.1128/aac.00312-12
	Padrão	Telaprevir
	Importância Biológica da proteína alvo	A NS5B é uma RNA polimerase dependente de RNA do HCV, responsável pela síntese do genoma viral de RNA em cópias de RNA complementares, que são então usadas para produzir novos genomas virais. Esse processo é fundamental para a replicação do HCV e para a produção de partículas virais infecciosas.
Virus da dengue	Alvo	Serine protease NS3
	PDB_ID	2M9P
	Artigo científico	Artigo em processo de revisão por pares
	Padrão	Telaprevir
	Importância Biológica da proteína alvo	A serina protease NS3 é crucial para a replicação do vírus da dengue, clivando proteínas virais e modulando a resposta imune do hospedeiro.
Virus HIV	Alvo	HIV-1 Reverse Transcriptase
	PDB_ID	5XN1
	Artigo científico	https://doi.org/10.1038/s41598-018-19602-9
	Padrão	Entecavir
	Importância Biológica da proteína alvo	A HIV-1 Reverse Transcriptase é crucial para a replicação do vírus HIV, convertendo seu RNA em DNA e permitindo sua integração no genoma do hospedeiro

Fonte: O Autor (2024)

7.4.9 Docking molecular

Nas análises de docking molecular realizadas no software *AutoDock Tools*, adotamos os seguintes parâmetros da *Grid-Box*: (i) exaustividade de 100; (ii) otimização das coordenadas do centro da caixa de grade para $x = 151,223$, $y = 108,563$ e $z = 21,821$; e (iii) otimização das dimensões da caixa de grade para $x = 40$ Å, $y = 40$ Å, $z = 40$ Å. Os complexos ligante-proteína com maior afinidade de ligação (kcal/mol) foram escolhidos para análises subsequentes. A busca pelas conformações (poses) dos ligantes com maior estabilidade de ligação as diferentes proteínas dos cinco vírus (vírus SARS-CoV-2, HBV, HCV, dengue e HIV) foi conduzida usando uma função de pontuação chamada soma de energias parametrizadas, juntamente com um algoritmo de busca aleatória conhecido como algoritmo genético. Para validar os resultados obtidos nas análises de docking triagem virtual, comparamos os valores de desvio médio quadrático (*RMSD*) e as energias de ligação entre as diferentes proteínas targets dos vírus com o seu fármaco controle listado na tabela 7.1. O software *Discovery Studio* foi usado para visualizar os resultados das análises de docking molecular.

7.4.10 Simulações de dinâmica molecular

Os ligantes com resultados promissores no docking molecular foram submetidos a simulações de dinâmica molecular com a proteína as proteínas alvas dos cinco diferentes vírus estudados (vírus SARS-CoV-2, HBV, HCV, dengue e HIV), usando sempre como comparador, os fármacos controles listados na tabela 7.1. Os testes de DM utilizaram um modelo de água cúbica de 20 Å, que incluía a proteína CCR-5, íons de cloreto e sódio para neutralizar as cargas. A otimização das geometrias envolveu etapas sequenciais de minimização, aquecimento e pressurização antes da simulação de equilíbrio no ensemble microcanônico (*NVE*). As simulações seguiram parâmetros específicos, incluindo passos de tempo de 2 fs e corte de interação eletrostática de 9,0 Å. As etapas de equilíbrio dinâmico foram conduzidas em dois períodos de 10 ns, variando entre temperatura constante (*NVT*) a 310 K e pressão contínua (*NPT*) a 1,0 atmosfera, dependendo da estabilização entre 100-200 ns. O campo de força Charm-36 (versão de julho de 2021) foi usado com interpolação cúbica, reajuste de ligações de hidrogênio usando o método *LINCS*, e as

simulações de DM foram realizadas por 100 ns. As métricas avaliadas incluíram *RMSD*, área de superfície acessível ao solvente (*SASA*) e cálculo de energia livre de ligação *MM/GBSA*, com representação gráfica dos resultados no software *Graph Prism*.

7.4.11 Cálculo de energia livre de ligação *MM/GBSA*

Os cálculos de ligação *MM/GBSA* (*Molecular mechanics/generalized Born surface area*) do complexo proteína-ligante entre os diferentes proteínas alvos dos diferentes vírus (vírus SARS-CoV-2, HBV, HCV, dengue e HIV) e os ligantes selecionados nas etapas anteriores foram realizados usando o *gmx-mmpbsa* no *GROMACS* usando as equações 1, 2 e 3.

$$\Delta G_{\text{bind}} = \Delta H - T\Delta S = \Delta E_{\text{MM}} - T\Delta S + \Delta G_{\text{sol}} \quad (\text{Eq. 1})$$

$$\Delta G_{\text{MM}} = \Delta E_{\text{vdw}} = \Delta E_{\text{interna}} + \Delta E_{\text{ele}} \quad (\text{Eq. 2})$$

$$\Delta G_{\text{sol}} = \Delta G_{\text{SA}} + \Delta G_{\text{GB}} \quad (\text{Eq. 3})$$

Onde: $T\Delta S$, ΔE_{MM} e ΔG_{sol} são, respectivamente, a entropia conformacional, a energia MM da fase gasosa e a energia livre de solvatação (a soma da contribuição apolar ΔG_{SA} e a contribuição polar ΔG_{GB}). ΔE_{MM} contém energias diédricas, eletrostáticas ΔE_{ele} e $\Delta E_{\text{interna}}$ da ligação, energia de *Van Der Waals* ΔE_{vdw} e ângulo. O cálculo da entropia pode ser omitido se nenhuma mudança estrutural for causada por ligações no processo de simulação de DM.

7.5 RESULTADOS

7.5.1 Análise do espaço químico

O espaço químico conjunto de dados dos 19772 compostos bioativos que com atividade antiviral contra o SARS-CoV-2, HIV, dengue e hepatite B e C foi explorado visando a obter *insights* sobre a relação quantitativa estrutura atividade (*QSAR*)

utilizando os quatro descritores da regra de 5 Lipinski (AlogP, MW, nHBDon e nHBAcc). AlogP é um parâmetro que define a medição da absorção do medicamento pelas membranas celulares. Quanto menor o AlogP, melhor será a absorção do fármaco através da bicamada lipídica das membranas. O PM é um parâmetro que também está associado à absorção de medicamentos, e quanto menor o peso molecular de um medicamento, mais rápida será sua absorção. nHBDon e nHBAcc são usados para expressar o número de grupos doadores e aceptores de ligações de hidrogênio. Em geral, quanto maior o número de ligações de hidrogênio entre um fármaco e o seu receptor, maior será a sua bioatividade.

A análise descritiva dos descritores da regra dos cinco de Lipinski de todos 19772 compostos químicos incluídos no estudo, mostrou que a mediana dos valores de PM= 412.28 g/mol; LogP=3.63; nHBDon =1 e nHBAcc=5. Podemos concluir que em média toda população molecular analisada atendia a regra dos cinco de Lipinski (MW<500; LogP<5; nHBDon =5 e nHBAcc=10) ou seja, os compostos tinham uma maior probabilidade de uma boa biodisponibilidade quando administrados pela via oral, ou simplesmente eram compostos *drug-like*.

A Figura 7.2 apresenta os resultados da análise do espaço químico dos 19772 compostos utilizando os descritores de Lipinski. Podemos observar que, segundo o teste de *Kruskal Wallis*, os três grupos de compostos (ativos, inativos e compostos com bioatividade intermediária) foram estatisticamente diferentes quanto a sua bioatividade (pIC50) como é mostrado na Figura em (A). Pode se observar que pelo gráfico de boxplot dos três grupos de compostos, o limiar dos valores de pIC50 que separa entre os compostos ativos com os restantes dois grupos (inativos e atividade intermediária) é pIC50 = 6.0, onde os compostos ativos apresentam pIC50>6,0 e os compostos inativos ou de atividade intermediária, apresentam um, pIC50<6,0 (p=0,00).

Outras diferenças observadas entre os três grupos de compostos (ativo e inativo) foram em relação ao peso molecular (p=0,00), AlogP (p=0,00) e grupos aceptores de ligações de hidrogênio (p=0,00). O único descritor da regra dos cinco de Lipinski que não foi observada diferenças significativas entre os três grupos de composto é o grupo doadores de ligações de hidrogênio (p=0,89). As diferenças estatísticas aqui observadas mostram que existe uma grande lacuna no espaço químico entre compostos com bioatividade *multi-target* (IC50<1000nM) contra os vírus SARS-CoV-2, HIV, dengue e hepatite B e C, e aqueles sem nenhuma bioatividade

($IC_{50} > 10000$ nM) ou com bioatividade intermediária (IC_{50} entre 1000-10000 nM). Esses *insights* fundamentaram o desenvolvimento de modelos de *machine learning* baseados em *QSAR multi-target* para investigar potenciais medicamentos para o tratamento simultâneo das infecções por SARS-CoV-2, HIV, dengue e hepatite B e C.

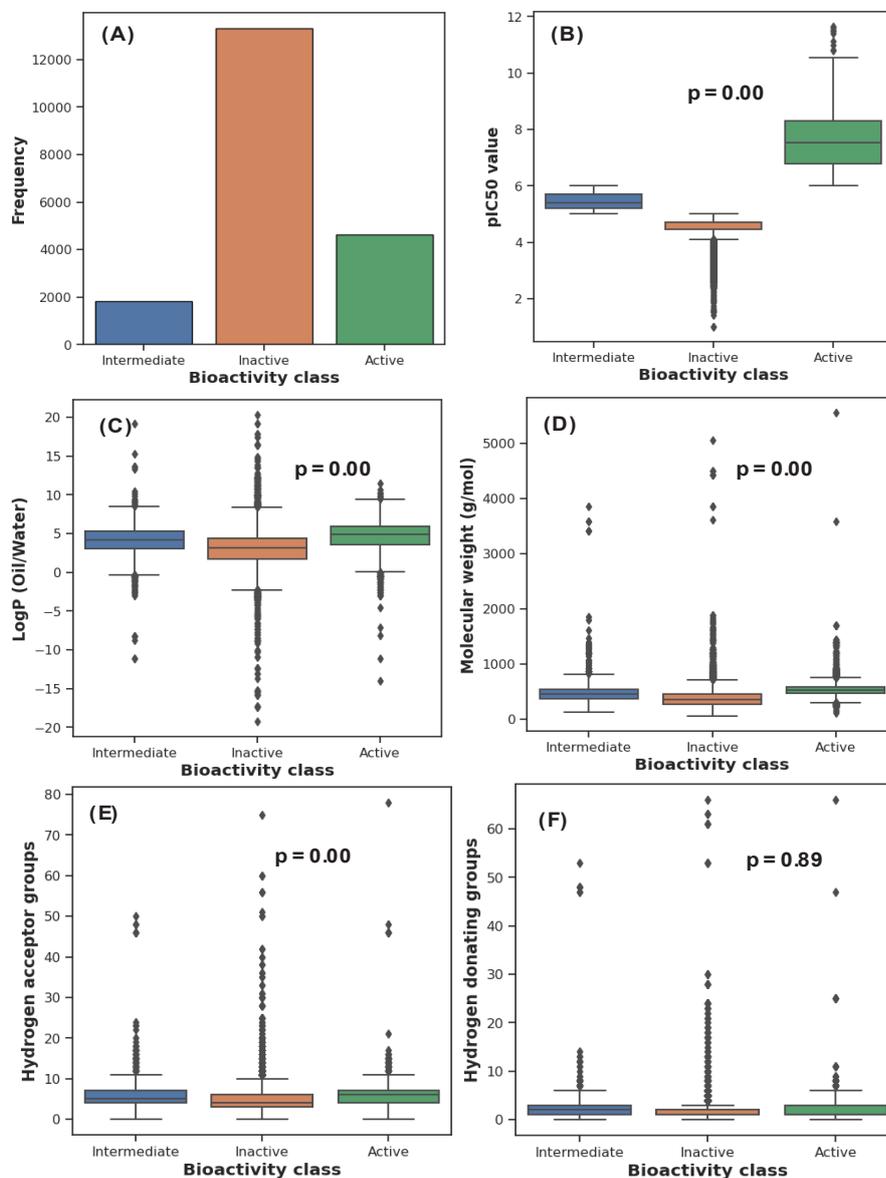


Figura 7.2. Análise exploratória dos compostos bioativos usados para o desenvolvimento de modelos de *machine learning* para predição de bioatividade de compostos anti SARS-CoV-2, HIV, HBV, HCV e dengue.

7.5.2 Machine learning

A análise das variáveis (descritores) permitiu selecionar apenas 277 descritores moleculares dos 881 descritores PubChem que estavam no banco de dados (604 variáveis eliminadas), e essas 277 descritores foram usados para o treinamento e

validação e teste dos 40 modelos de *machine learning*. permitiu minimizar os problemas de multicolinearidade nos modelos de *machine learning* treinados. Na sequência, um total de quarenta ($n=40$) diferentes algoritmos de regressão de *machine learning* foram treinados, avaliados e testados, visando prever a atividade biológica (pIC_{50}) *multi-target* contra os vírus SARS-CoV-2, HIV, dengue e hepatite B e C, a partir dos descritores moleculares *PubChem* dos 19772 compostos químicos bioativos. Dos 40 algoritmos treinados e testados, os 5 algoritmos com melhor desempenho preditivo foram *Extra Trees regression*, *random Forest (RF)*, *Extreme Gradient Boosting (XGBoost)*, e *Histogram-based Gradient Boosting*, cujos valores de coeficiente de determinação foram de (R^2) 0,94; 0,92; 0,89; 0,80 e 0,80, respectivamente (Tabela 7.2). Outras métricas de desempenho (MSE, RMSE e MAE) desses cinco melhores algoritmos de *machine learning multi-target* são também mostradas na tabela 1. A Figura 7.3 apresenta o gráfico dos valores experimentais e preditos da bioatividade (pIC_{50}) dos compostos contra os vírus SARS-CoV-2, HIV, dengue e hepatite B e C.

Tabela 7.2. Top 5 melhores modelos de machine learning multi-target para predição dos compostos com bioatividade simultânea contra o SARS-CoV-2, HIV, dengue e hepatite B e C

Modelo de regressão multi-target	R^2	MSE	RMSE	MAE
Extra Trees Regression	0,94	0,57	0,76	0,41
Random Forest Regression	0,92	0,42	0,65	0,39
Extreme Gradient Boosting	0,89	0,43	0,65	0,43
Histogram-based Gradient Boosting	0,80	0,51	0,71	0,47
Light Gradient Boosting Model	0,80	0,50	0,71	0,47

Nota: MSE: mean square error; RMSE: root mean square error; MAE: mean absolute error. **Fonte:** O Autor (2024)

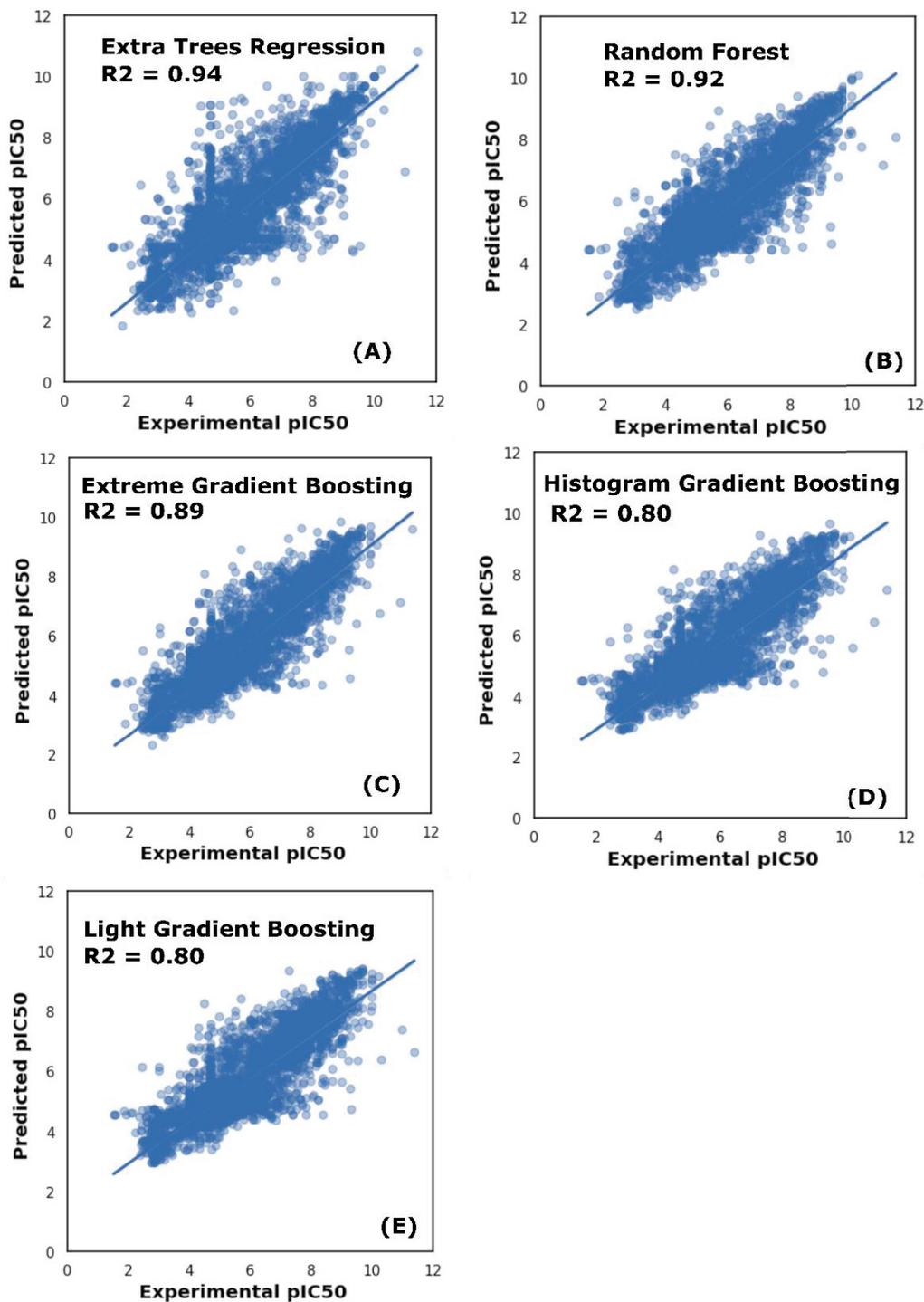


Figura 7.3. Gráfico dos valores experimentais e preditos da bioatividade antiviral multi-target (HIV, SARS-CoV-2, dengue, e hepatite B e C) baseados na análise de relação estrutura atividade quantitativa (QSAR) dos 19 mil compostos químicos usando os top cinco melhores modelos de machine learning. Em (A), (B), (C), (D), e (E) é referente aos modelos *Extra Trees regression*, *random Forest (RF)*, *Extreme Gradient Boosting (XGBoost)*, *Histogram-based Gradient Boosting*, respectivamente. **Fonte:** O Autor (2024).

7.5.3 Utilização dos cinco modelos de machine learning de melhor desempenho em um conjunto de dados externo: Human Metabolome Database

Os cinco modelos de *machine learning* com melhor desempenho preditivo (*Extra Trees*, *RF*, *XGBoost*, *HBGB* e *LGBM*) foram utilizados para prever a bioatividade simultânea contra SARS-CoV-2, HIV, dengue, hepatite B e C, em um conjunto de 113.682 compostos do *Human Metabolome Database (HMDB)*. Na Tabela 7.3, são apresentados os valores previstos dos top 10 compostos com os maiores níveis de bioatividade (altos valores de pIC₅₀).

Com base na Tabela 7.3, observamos que apenas três compostos apresentaram bioatividade antiviral *multi-target* significativa (IC₅₀ < 1000 nM ou pIC₅₀ > 6), nomeadamente: Entecavir (HMDB0014585) com pIC₅₀ médio = 7.79 ou IC₅₀ médio = 16.22 nM; Nummularine B (HMDB0029334) com pIC₅₀ médio = 6.34 ou IC₅₀ = 457.08 nM; e Telaprevir (HMDB0015616) com pIC₅₀ médio = 6.14 e IC₅₀ = 724.44 nM.

É relevante destacar que os coeficientes de variação dos resultados de predição de bioatividade nos cinco modelos de *machine learning* diferem em menos de 2%, demonstrando a alta precisão e robustez desses algoritmos na predição da bioatividade simultânea contra os vírus HIV, SARS-CoV-2, dengue, hepatite B e C.

Nummularine B emerge como um composto bioativo, indicando seu potencial como candidato para o tratamento das doenças virais mencionadas. Entretanto, Entecavir e Telaprevir são fármacos já aprovados pela FDA, sugerindo que podem ser potenciais candidatos para o reposicionamento no tratamento simultâneo de SARS-CoV-2, HIV, dengue, hepatite B e C.

Outro aspecto a ser destacado é que a definição de valores de pIC₅₀ > 6 como importantes para a predição de bioatividade foi estabelecida na análise do espaço químico dos compostos, conforme mostrado na Figura 7.2-B. Por outro lado, os valores de IC₅₀ < 1000 nM como limite de corte para considerar um composto como bioativo foram baseados na literatura e estão descritos na seção de materiais e métodos.

Tabela 7.3. Predição da bioatividade *multi-target* (HIV, SARS-CoV-2, dengue, hepatite B e C) para os 113.682 compostos do *Human Metabolome Database*, utilizando modelos de machine learning como *Extra Trees regression*, *Random Forest (RF)*, *Extreme Gradient Boosting (XGBoost)* e *Histogram-based Gradient Boosting*. Apresentamos apenas os 10 principais compostos que exibiram valores elevados de bioatividade média (média de pIC50). Para uma compreensão mais clara dos resultados, os valores médios de pIC50 foram convertidos para IC50 (nM)

HMDB ID	Nome trivial	Valores pIC50 previstos por algoritmos de <i>machine learning</i>					Média de pIC50	Média de IC ₅₀ (nM)	CV (%)
		EXTRA	HISTO	LGBM	RF	XGB			
HMDB0014585	Entecavir	7,89	7,81	7,76	7,64	7,85	7,79	16,22	1,13
HMDB0029334	Nummularine B	6,36	6,21	6,54	6,29	6,29	6,34	457,08	1,74
HMDB0015616	Telaprevir	6,18	6,11	6,16	6,20	6,04	6,14	724,44	0,96
HMDB0255505	M8-Nelfinavir	5,37	5,28	5,24	5,36	5,12	5,27	5370,31	1,75
HMDB0014584	Gemcitabina	5,28	5,27	5,14	5,26	5,23	5,23	5888,43	0,93
HMDB0015603	Ximelagatran	5,21	5,31	5,11	5,27	5,13	5,21	6165,95	1,53
HMDB0304782	Histidinil-triptofano	4,88	5,10	5,10	4,92	4,98	5,00	10000	1,82
HMDB0028916	Isoleucil-serina	4,93	4,95	5,13	4,93	5,01	4,99	10232,92	1,49
HMDB0013617	Lipoil-GMP	5,09	5,02	4,84	5,00	4,99	4,99	10232,92	1,63
HMDB0013024	Neurotensin 11-13	4,88	4,99	5,02	4,88	5,07	4,97	10715,19	1,53

Nota: EXTRA; extra trees regression; HISTO: Histogram-based Gradient Boosting; LGBM: Light Gradient Boosting Model; RF: random forest; XGB: Extreme Gradient Boosting. **Fonte:** O Autor (2024).

Na Figura 7.4 são mostradas as estruturas 3D dos três compostos promissores para o tratamento simultâneo de HIV, SARS-CoV-2, dengue, hepatite B e C.

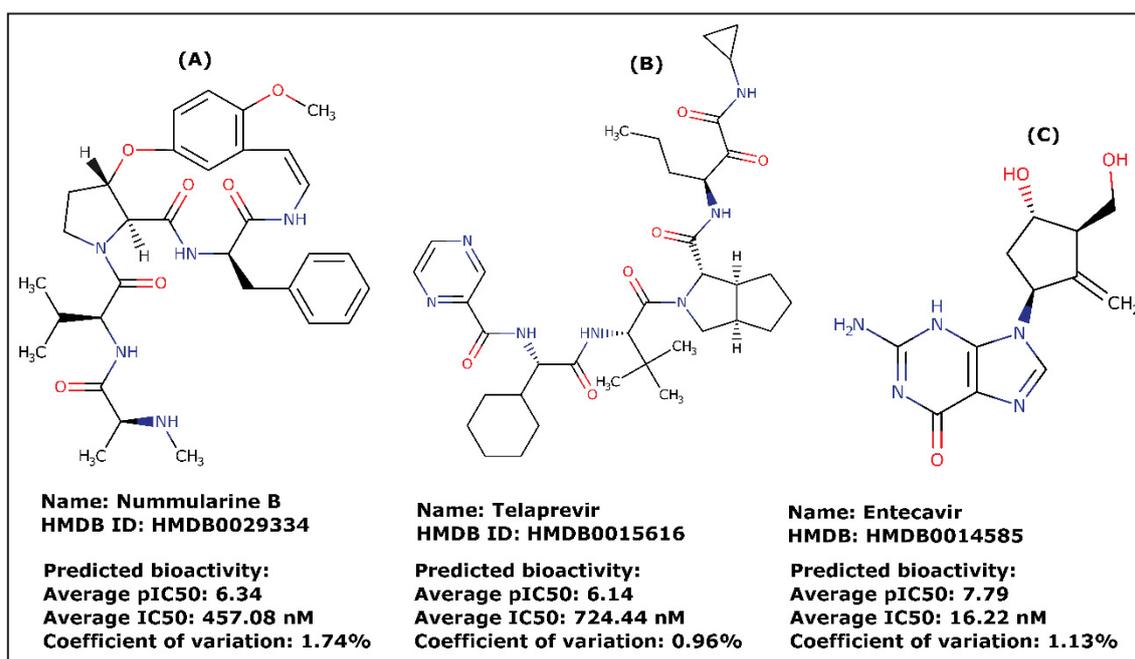


Figura 7.4. Estrutura química de Nummularine B (A), Telaprevir(B) e Entecavir (C), com seus respectivos valores médios de bioatividade *multi-target* anti SARS-CoV-2, HBV, HCV, Dengue e HIV. **Fonte:** O Autor (2024).

7.5.4 Análises de docking molecular

Na figura 7.5 são mostrados os resultados de docking molecular entre os nummularine B, telaprevir e Entecavir, com a protease principal do SARS-CoV-2 (Figura 4A), complex isomerase do HBV (figura 4B), RNA polimerase de HCV (figura 4C), a serina protease de NS3 da dengue (figura 4D), e a transcriptase reversa de HIV (figura 4E). Podemos observar que todos três ligantes tiveram energias de afinidade favorável com os suas proteínas alvos.

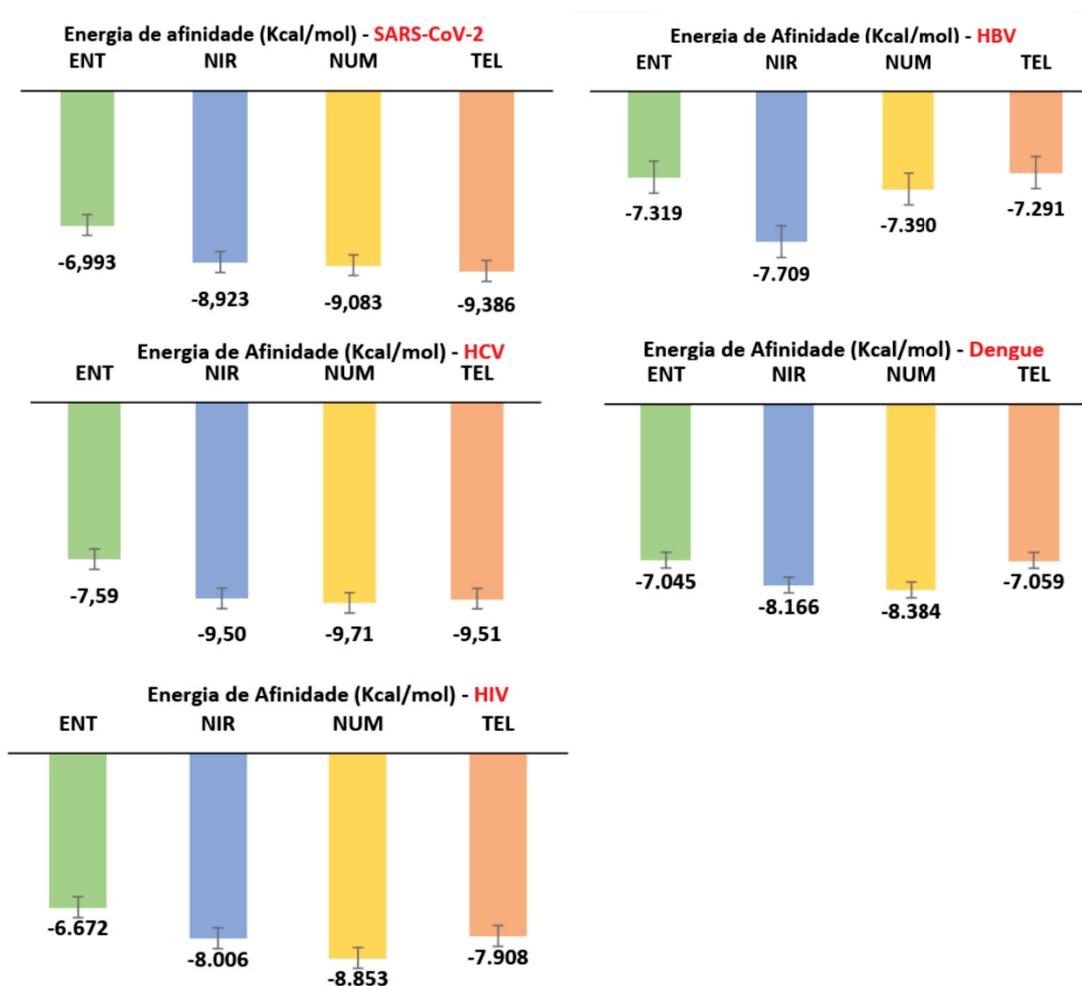


Figura 7.5 – Análise de docking molecular. Gráficos de energia livre de ligação entre nummularine B, telaprevir e Entecavir com as proteina salvos dos virus SARS-CoV-2, HBV, HCV, dengue e HIV. **Fonte:** O Autor (2024).

7.5.5 Simulações de dinâmica molecular

Os resultados das energias livres de ligação MM/GBSA para cada complexo formado entre os nummularine B, telaprevir e Entecavir, com a protease principal do SARS-CoV-2 (Figura 7.6-A), complex isomerase do HBV (Figura 7.6-B), RNA polimerase de HCV (Figura 7.6-C), a serina protease de NS3 da dengue (Figura 7.6-D), e a transcriptase reversa de HIV (Figura 7.6-E), obtidas por simulações de dinâmica molecular são mostrados abaixo. Em todas as figuras, podemos observar que todos os complexos proteína-ligante formou uma interação favorável em todo período de 2000 pico segundos de simulação de dinâmica molecular, ou seja, todos os três ligantes apresentaram valores médios (-) de MM/GBSA para o complexo ΔG e a diferença ΔG (Complexo - Receptor - Ligante). Esses resultados validam os achados observados nas análises de *machine learning* baseados em QSAR e docking molecular, sugerindo que os ligantes nummularine B, telaprevir e Entecavir são potenciais candidatos a fármacos para o tratamento simultaneo de COVID-19, hepatite B, hepatite C, dengue e HIV/AIDS.

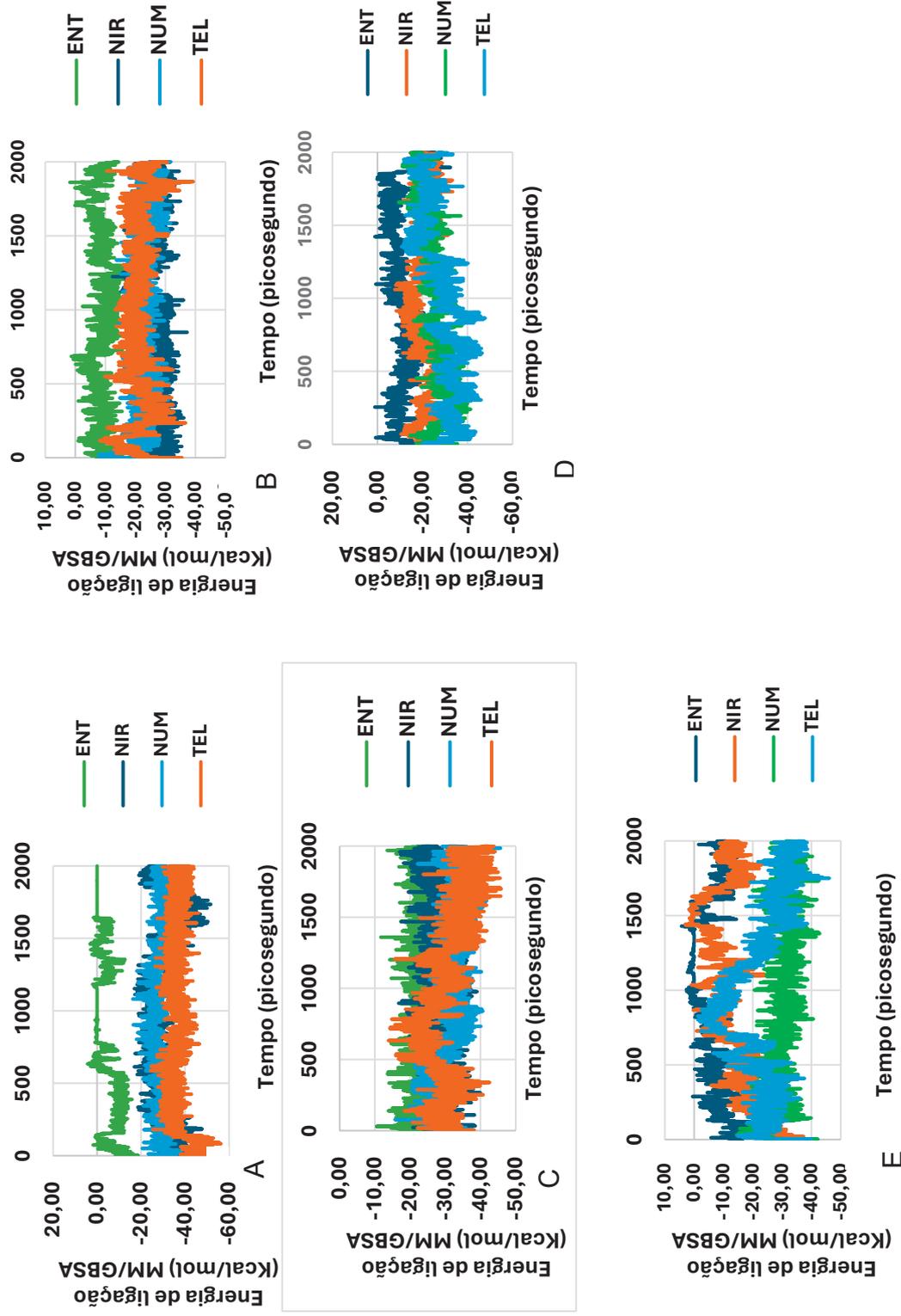


Figura 7.6. Gráfico de energia livre de ligação MM/GBSA dos quatro ligantes (nummularine B, telaprevir e Entecavir) com a proteína MPro do SARS-CoV-2 (Figura A), complex isomerase do HBV (Figura B), RNA polimerase de HCV (Figura C), a serina protease de NS3 da dengue (Figura D), e a transcriptase reversa de HIV (Figura E). **Fonte:** O Autor (2024).

7.5.5.1 Variação de energias livres de ligação MM/GBSA por cada resíduo dos aminoácidos das proteínas alvos de SARS-CoV-2, HBV, HCV, dengue e HIV

A Figura 7.7 ilustra as variações nas energias livres de ligação MM/GBSA para cada resíduo de aminoácido na protease principal (*Mpro*) do SARS-CoV-2. As interações entre a proteína *Mpro* do SARS-CoV-2 e os compostos nummularine B, telaprevir, nirmaltegravir (fármaco controle) e Entecavir são apresentadas nas subfiguras A, B, C e D, respectivamente. Os resíduos de aminoácidos da proteína *Mpro* que apresentaram interações favoráveis com os ligantes nummularine B, telaprevir, Entecavir e nirmaltegravir (fármaco controle) foram: Metionina 49, Asparagina 142, Serina 144, Serina 145, Metionina 165, Leucina 167, Prolina 168, Glicina 189, Glicina 192 e Glicina 197.

Para o vírus HBV, a variação das energias livres de ligação MM/GBSA de cada aminoácido da proteína complex isomerase são mostrados na Figura 7.8. Pode-se observar. Os resíduos de aminoácidos da proteína complex isomerase que apresentaram interações favoráveis com os ligantes nummularine B (A), telaprevir (B), Entecavir (C) e nirmaltegravir (D) foram: Leucina 501, Arginina 422, Metionina 423, Histidina 475, Serina 476, Tirosina 477, Isoleucina 482, Leucina 497, Triptofano 528.

A Figura 7.9 representa as variações nas energias livres de ligação MM/GBSA para cada resíduo de aminoácido na protease principal (*Mpro*) do SARS-CoV-2. As interações entre a proteína RNA polimerase do HCV e os compostos nummularine B, telaprevir, nirmaltegravir (fármaco controle) e Entecavir são exibidas nas subfiguras 8A, 8B, 8C e 8D, respectivamente. Os resíduos de aminoácidos da proteína *Mpro* que demonstraram interações favoráveis com os ligantes nummularine B, telaprevir, Entecavir e nirmaltegravir (fármaco controle) foram: Prolina 93, Fenilalanina 193, Glicina 194, Glicina 446, Isoleucina 447, Tirosina 448, Tirosina 555 e Serina 556.

A variação das energias livres de ligação MM/GBSA de cada aminoácido da enzima serina protease de NS3 da dengue são mostrados na Figura 7.10. Pode-se observar. Os resíduos de aminoácidos da enzima serina protease de NS3 que apresentaram interações favoráveis com os ligantes nummularine B (A), telaprevir (figura 9B), entecavir (figura 9C) e nirmaltegravir (figura 9D) foram: Leucina 97, Prolina 193, Tirosina 195, Valina 216, e Tirosina 222.

Finalmente, a variação das energias livres de ligação MM/GBSA de cada aminoácido da enzima transcriptase reversa do HIV são mostrados na Figura 7.11.

Pode-se observar. Os resíduos de aminoácidos da proteína complex isomerase que apresentaram interações favoráveis com os ligantes nummularine B (A), telaprevir (B), entecavir (C) e nirmaltegravir (D) foram: Glicina 91, Leucina 92, Prolina 95, Arginina 172, Tirosina 181, Isoleucina 382, e Valina 381.

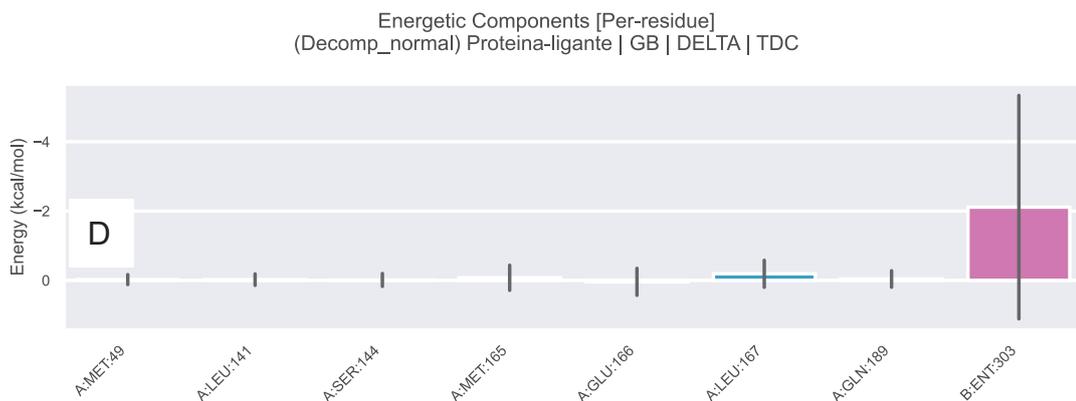
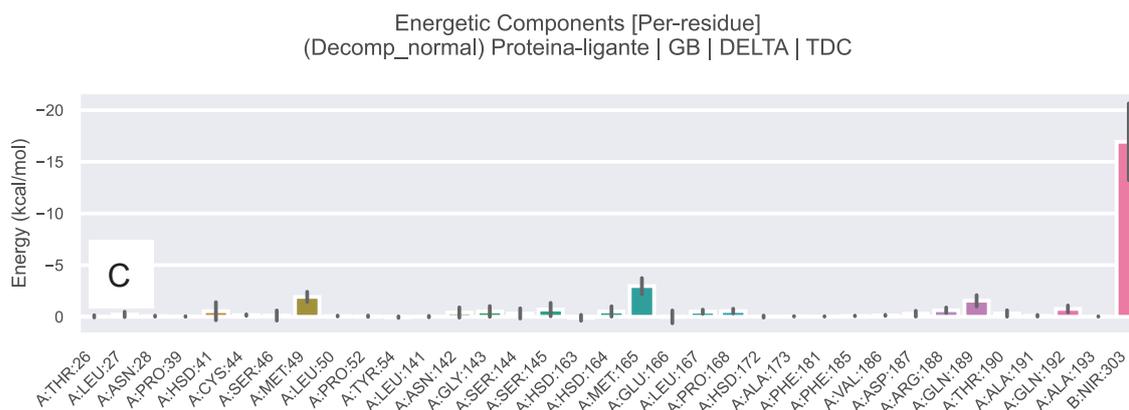
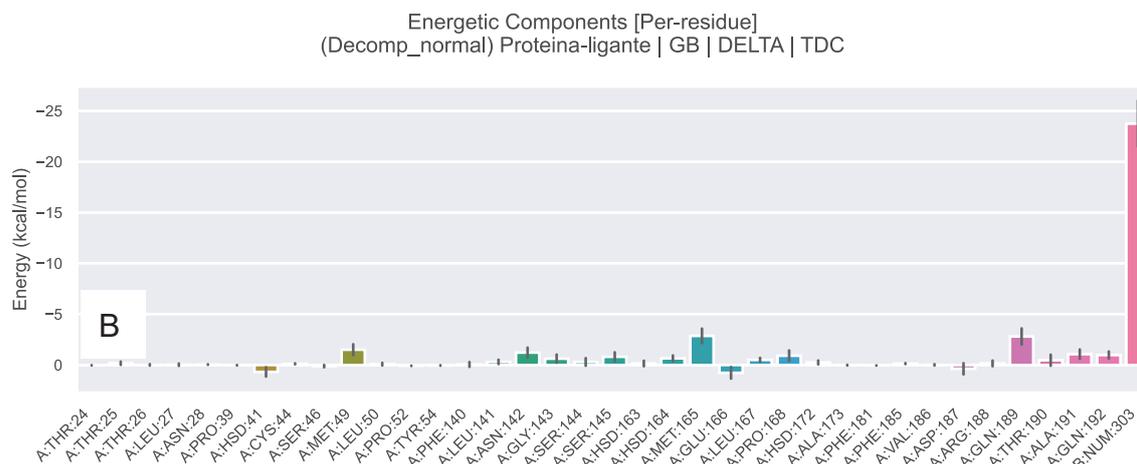
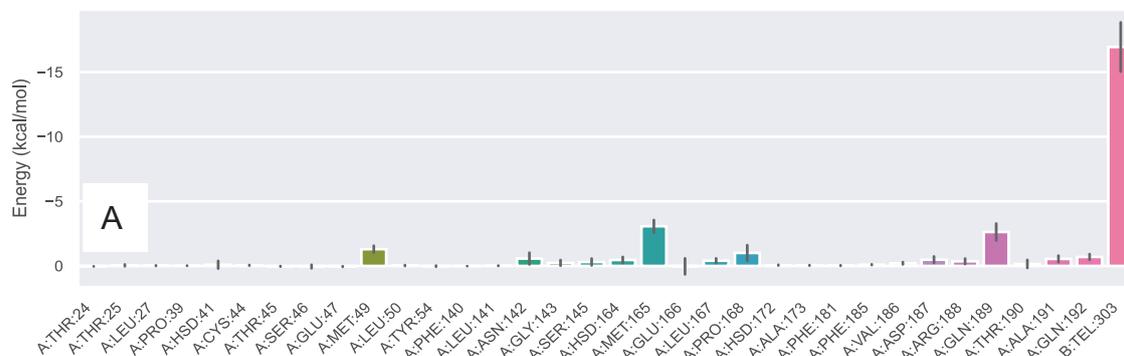


Figura 7.7. Variação das energias livres de ligação MM/GBSA de cada aminoácido da enzima Protease principal (PMro) do vírus SARS-CoV-2 com os ligantes nummularine B (A), telaprevir (B), Entecavir (C) e nirmaltegravir (D).

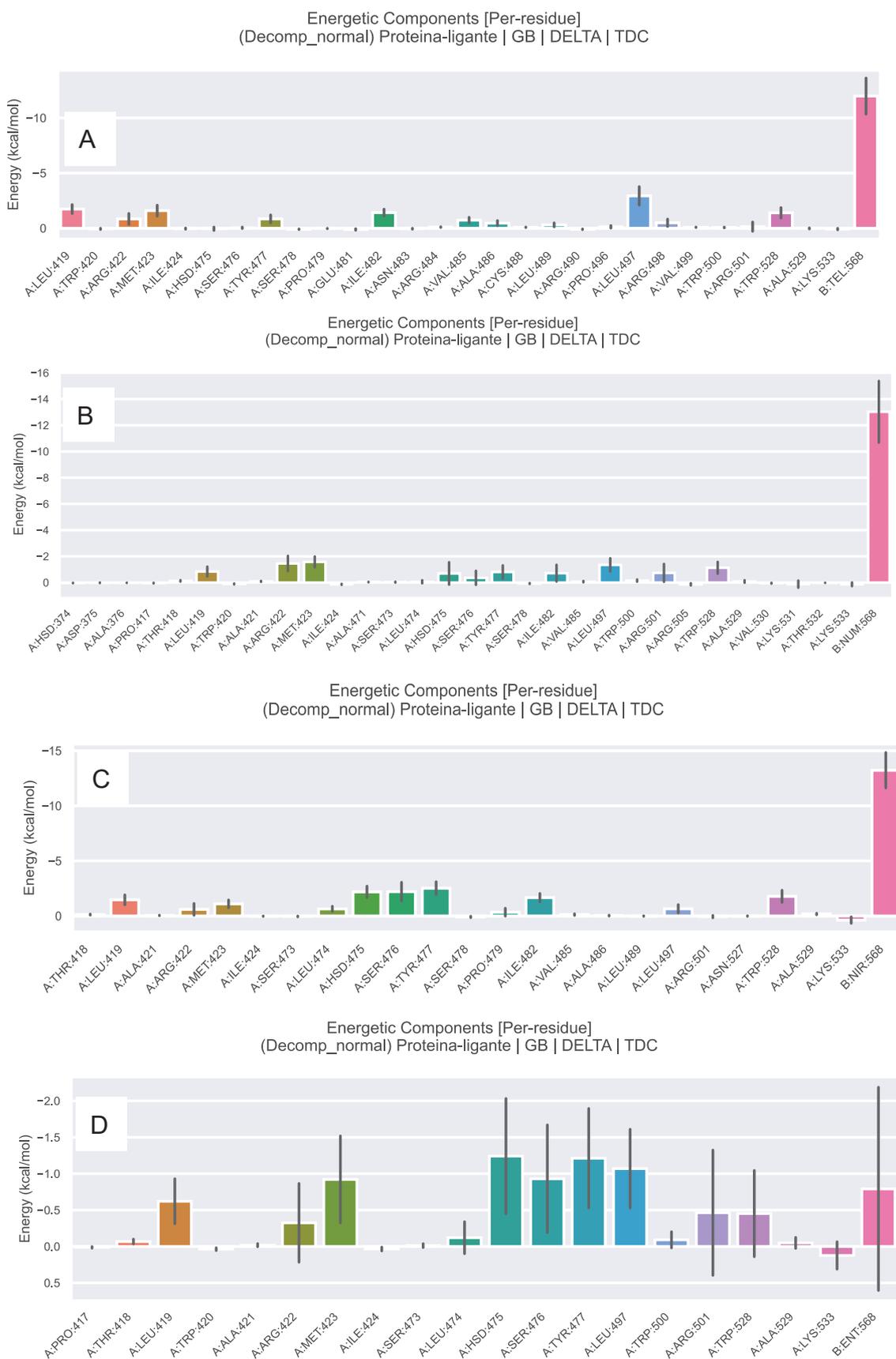


Figura 7.8. Variação das energias livres de ligação MM/GBSA de cada aminoácido da enzima complex isomerase do vírus da hepatite B com os ligantes nummularine B (A), telaprevir (B), Entecavir (C) e nirmaltegravir (D). **Fonte:** O Autor (2024).

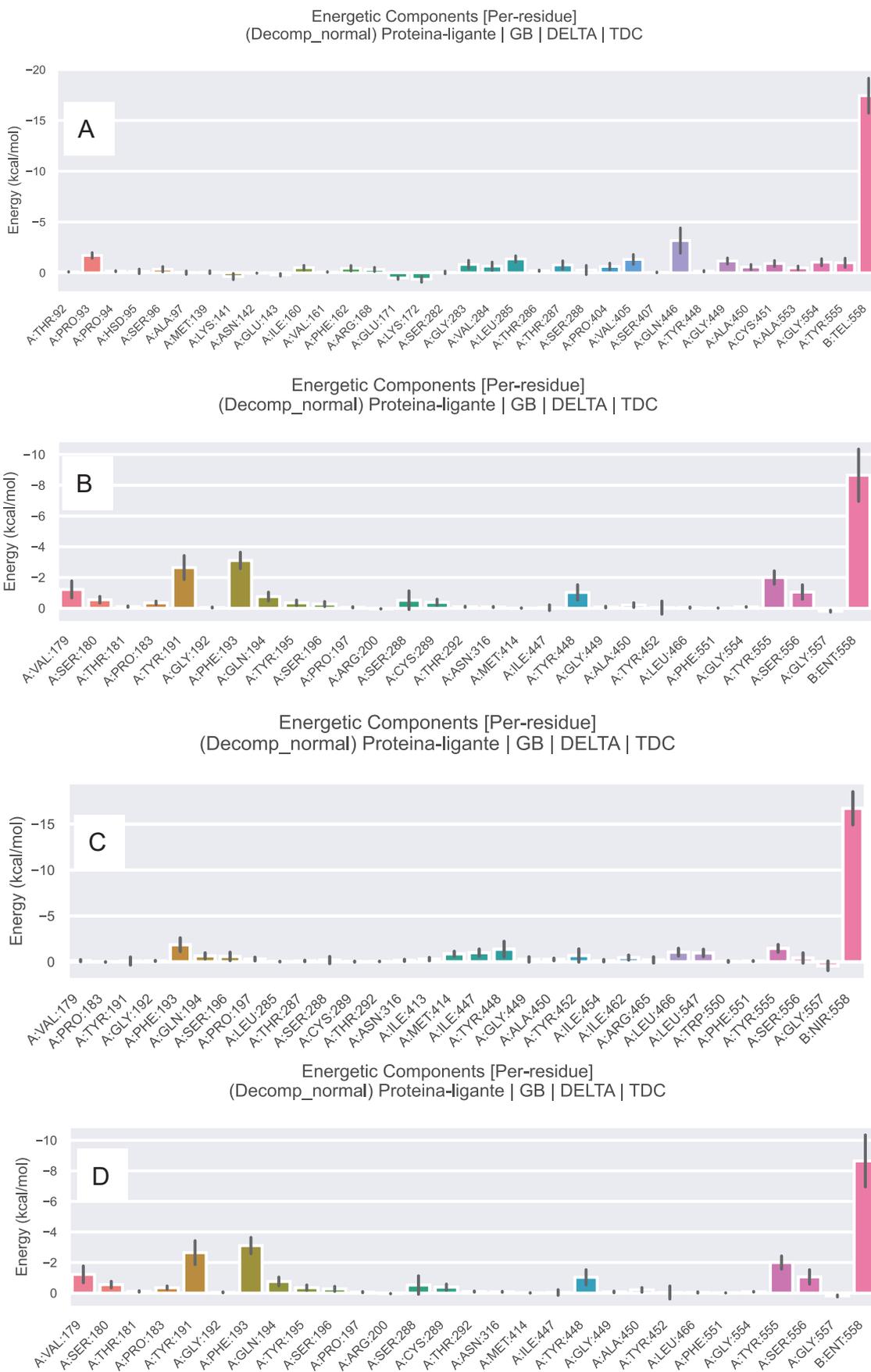


Figura 7.9. Variação das energias livres de ligação MM/GBSA de cada aminoácido da enzima RNA polimerase do vírus da hepatite C com os ligantes nummularine B (A), telaprevir (B), Entecavir (C) e nirmaltegravir (D). **Fonte:** O Autor (2024).

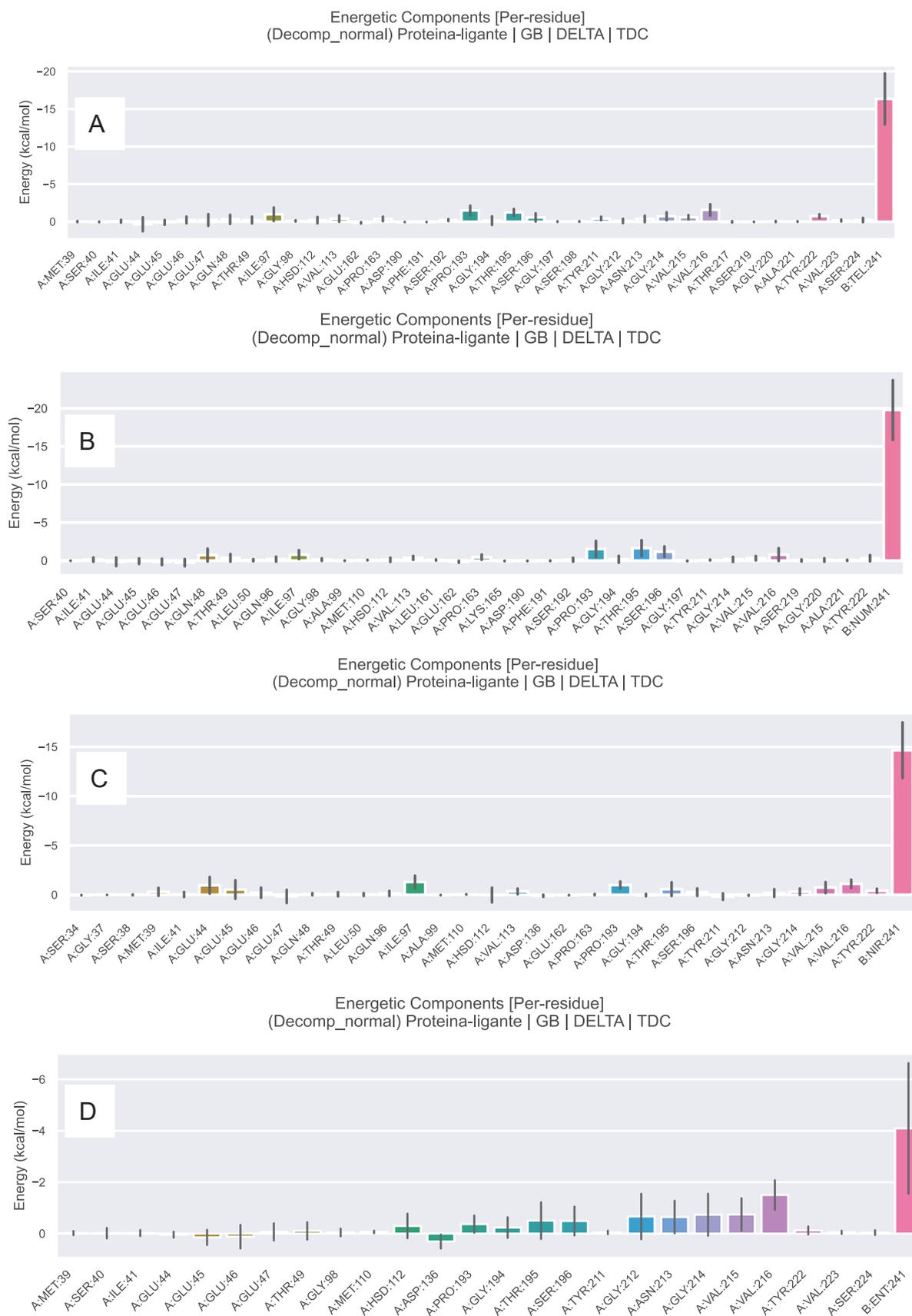


Figura 7.10. Variação das energias livres de ligação MM/GBSA de cada aminoácido da enzima serina protease de NS3 do vírus da dengue com os ligantes nummularine B (A), telaprevir (B), Entecavir (C) e nirmaltegravir (D). **Fonte:** O Autor (2024).

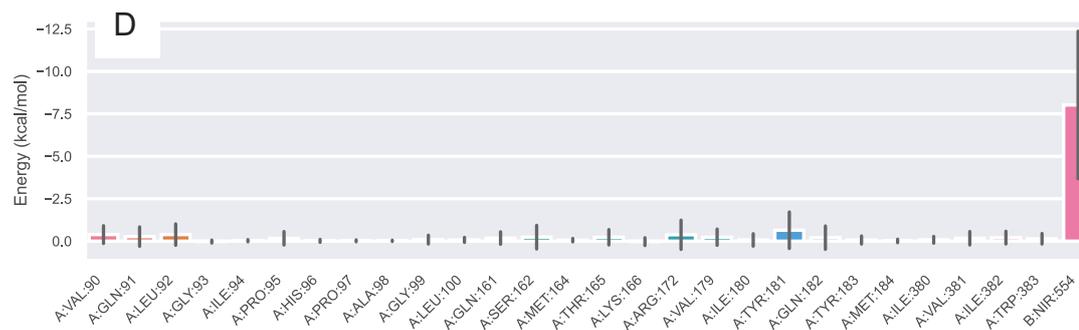
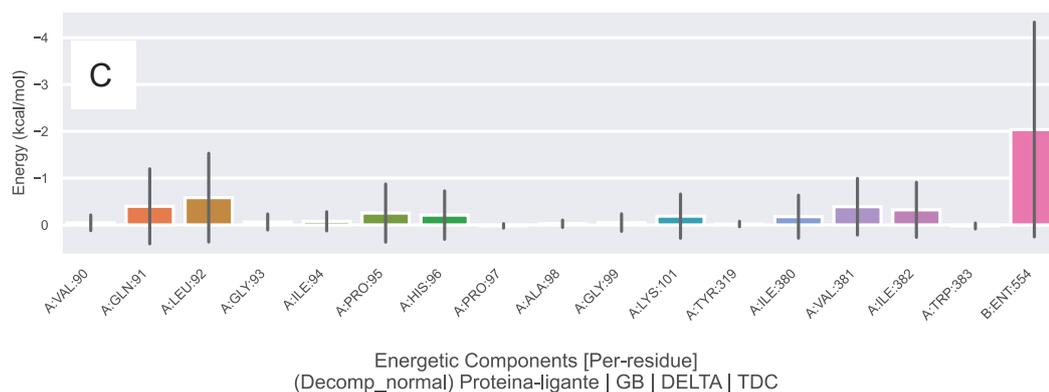
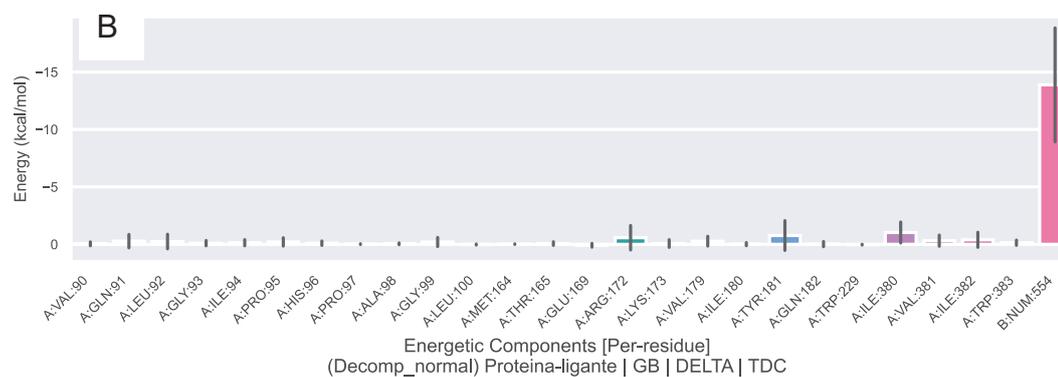
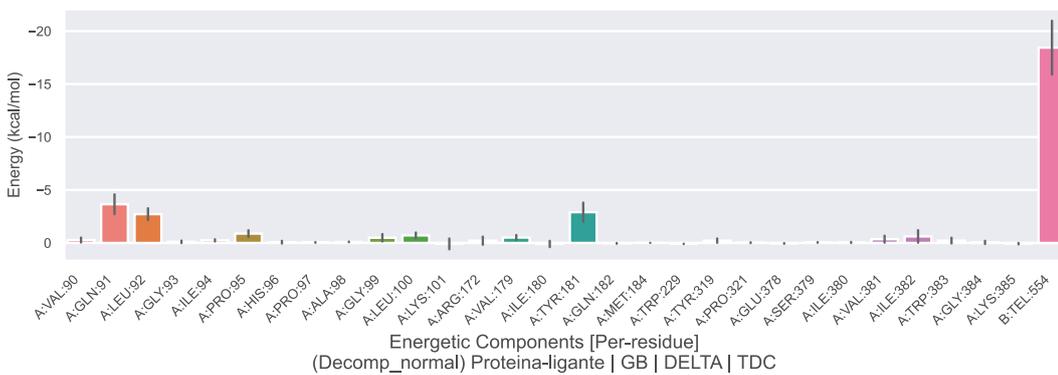


Figura 7.11. Variação das energias livres de ligação MM/GBSA de cada aminoacido da enzima transcriptase reversa do virus HIV-1 com os ligantes nummularine B (A), telaprevir (B), entecavir (C) e nirmaltegravir (D). **Fonte:** O Autor (2024).

7.6 DISCUSSÃO

Neste capítulo VII da tese, um total de modelos de machine learning baseados em QSAR foram treinados e validados usando dados experimentais da vida real visando prever novos candidatos a fármacos para o tratamento simultâneo de COVID-19, hepatites B e C, dengue e HIV/AIDS. Os algoritmos de *machine learning* *Extra Trees*, *RF*, *XGBoost*, *HBGB* e *LGBM* foram os que tiveram melhor desempenho preditivo. Esses top 5 melhores modelos foram usados para prever a bioatividade *multi-target* de um banco de dados externo (*Human metabolome Database*)⁵⁹⁶ contendo 222 mil ligantes, o que permitiu identificar três compostos que tiveram resultados promissores, nomeadamente: nummularine B ($IC_{50}=478.08$ nM), telaprevir (724.44 nM), e entecavir ($IC_{50} = 16.22$ nM). A bioatividade multi-target desses três compostos foi confirmada também por análises de docking molecular e simulações de dinâmica molecular. Dos três ligantes identificados a entecavir e a telaprevir já são fármacos aprovados pelo FDA e recomendados para o tratamento simultâneo da hepatite B crônica e hepatite C crônica, respectivamente. Além disso, desde janeiro de 2016, foi aprovado nos Estados Unidos o protocolo terapêutico que recomenda a entecavir para o tratamento de pacientes adultos e jovens co-infectados por HBV e HIV, devido a bioatividade simultânea de entecavir em esses dois vírus demonstrados nos ensaios clínicos randomizados^{597,598}, e esses resultados reforçam e consolidam a precisão dos nossos achados. No nosso para além de identificar o entecavir e a telaprevir para o tratamento de hepatite B e HIV, nós também propomos que esses medicamentos poderiam ser reposicionados para o tratamento simultâneo de HIV, SARS-CoV-2 e o dengue. Nós também identificamos pela primeira vez o fitoquímico nummularine B, que também demonstrou resultados promissores como potencial fármaco para o tratamento simultâneo de infecções SARS-CoV-2, HBV, HCV, dengue e HIV.

Um ponto que merece especial destaque é que todos os cinco principais modelos de machine learning que tiveram maior capacidade em prever potenciais fármacos com bioatividade multi-target são baseados em árvores de decisão (*Extra Trees*, *RF*, *XGBoost*, *HBGB* e *LGBM*), e há muitas publicações científicas mostrando o alto desempenho desses algoritmos, não apenas na área de descoberta de medicamentos [40–42], mas também em outras áreas, como em ciências ômicas [43,44]. Existem diversas razões que explicam o alto desempenho preditivo

apresentado por estes algoritmos baseados em árvores de decisão, e listamos algumas delas: (i) em primeiro lugar, estes modelos são versáteis e robustos, sendo capazes de lidar com diferentes tipos de dados, incluindo categóricos e numérico, sem necessidade de pré-processamento extenso, além de ser resistente a outliers e dados ruidosos. (ii) A capacidade inerente das árvores de decisão para capturar relações não lineares nos dados é crucial, especialmente em situações em que as relações entre as variáveis preditoras e a variável resposta são complexas e não podem ser facilmente modeladas por abordagens lineares. (iii) A seleção automática de características, atribuindo importância a cada variável de entrada, é outra vantagem proporcionada pelos modelos baseados em árvore. Isto é valioso para identificar quais características têm maior influência nas previsões. (iv) os modelos de *machine learning* podem ser suscetíveis ao overfitting. No entanto, técnicas como Random Forests e Gradient Boosting são projetadas para mitigar esse problema, melhorando a generalização do modelo. (v) A capacidade de capturar naturalmente interações entre variáveis é outra vantagem, permitindo ao modelo aprender relações complexas entre diferentes características. (vi) Por fim, a facilidade de ajuste de hiperparâmetros oferecida por esses modelos proporciona flexibilidade adicional, permitindo otimizar o desempenho do modelo de acordo com as características específicas do conjunto de dados. Essas características tornam os modelos baseados em árvore uma escolha popular em uma variedade de cenários de *machine learning*.

Dos três candidatos a fármacos identificados para o tratamento simultâneo das cinco doenças investigadas (COVID-19, hepatite B e C, dengue e HIV) dois deles já estão aprovados pelo FDA para o tratamento da hepatite B e C crônica^{599,600}. O primeiro deles é o primeiro O telaprevir, um fármaco que já está pelo FDA que pertence à classe dos inibidores de protease viral e atua inibindo a atividade da protease NS3/4A do vírus da hepatite C, impedindo assim a replicação viral⁵⁹⁹. O segundo, é o entecavir, um medicamento antiviral usado para o tratamento simultâneo de hepatite B crônica e HIV^{597,598}. Entecavir pertence à classe dos análogos de nucleosídeos e na HBV funciona inibindo a replicação do vírus ao agir como um inibidor da polimerase do HBV, uma enzima essencial para a replicação viral⁶⁰¹. Entecavir é considerado um medicamento de primeira linha para o tratamento da hepatite B crônica devido à sua eficácia e perfil de segurança^{602,603}. No nosso estudo, as análises de *machine learning*, docking molecular e simulações de dinâmica molecular, para além de identificar telaprevir e entecavir para o tratamento de hepatite,

também propomos que esses dois medicamentos sejam reposicionados para o tratamento simultâneo de SARS-CoV-2, dengue e HIV.

Nummularine B, um novo candidato a fármaco que se mostrou promissor em nosso estudo para o tratamento simultâneo de infecções por SARS-CoV-2, HBV, HCV, dengue e HIV, é um alcaloide ciclopeptídico isolado das raízes de *Ziziphus jujuba*⁶⁰⁴. Existem dados na literatura da sua bioatividade atividade antiviral contra o vírus da diarreia epidêmica suína (PEDV) e efeitos inibitórios potentes na replicação do PEDV⁶⁰⁴, e recentes estudos tem reportado possuir bioatividade contra o vírus SARS-CoV-2⁶⁰⁵. Portanto, os dados da literatura corroboram com os achados em nossos estudos, do potencial bioatividade multi-target contra os vírus SARS-CoV-2, HBV, HCV, dengue e HIV.

Pelo facto de todos os vírus (SARS-CoV-2, HBV, HCV, dengue e HIV) investigados serem vírus RNA e tendo um comum as enzimas (por exemplo proteases e RNA polimerase)⁶⁰⁶, nós propomos um possível mecanismo de ação da bioatividade *multi-target* desses do entecavir e telaprevir. Com base nos ensaios de simulações de dinamica molecular, a bioatividade anti SARS-CoV-2 de telaprevir (724.44 nM) e entecavir (12.22nM), se dá pela inibição da enzima proteína principal (MPro), especificamente nos seguintes resíduos de aminoácidos da MPro: Metionina 49, Asparagina 142, Serina 144, Serina 145, Metionina 165, Leucina 167, Prolina 168, Glicina 189, Glicina 192 e Glicina 197⁶⁰⁷. Já, a bioatividade anti-hepatite B dos três ligantes (nummularine B, entecavir e telaprevir) sobre o HBV pela via de inibição da enzima complex isomerase, especificamente, na sua interação dos seguintes aminoácidos: Leucina 501, Arginina 422, Metionina 423, Histidina 475, Serina 476, Tirosina 477, Isoleucina 482, Leucina 497, Triptofano 528⁶⁰⁸. No caso da bioatividade dos três compostos na hepatite C, os ensaios de dinâmica docking e simulações de dinâmica molecular sugerem que é pela inibição da enzima RNA Polimerase (NS5B), através de interação dos seguintes resíduos dos aminoácidos: Prolina 93, Fenilalanina 193, Glicina 194, Glicina 446, Isoleucina 447, Tirosina 448, Tirosina 555 e Serina 556⁶⁰⁹.

Para a bioatividade de nummularine B, entecavir e telaprevir contra o vírus da dengue, análises de docking molecular e as simulações de dinâmica molecular sugerem que se deve pela inibição da enzima serina protease de NS3 do vírus, tendo como foco na interação com os resíduos de Leucina 97, Prolina 193, Tirosina 195, Valina 216, e Tirosina 222. Por fim, a bioatividade anti-HIV de nummularine B,

entecavir e telaprevir foi sugerida pelos ensaios de docking molecular e simulações de dinâmica molecular pela inibição da enzima transcriptase reversa do vírus, especificamente nos resíduos dos aminoácidos: Glicina 91, Leucina 92, Prolina 95, Arginina 172, Tirosina 181, Isoleucina 382, e Valina 381⁵⁹⁷.

Uma limitação do nosso estudo é que apesar de identificarmos telaprevir e entecavir como candidatos promissores para o tratamento simultâneo de COVID-19, hepatites B e C, dengue e HIV/AIDS, a eficácia desses medicamentos em pacientes com infecções por esses vírus deve ser confirmada por meio de ensaios clínicos randomizados. Além disso, embora tenhamos proposto o reposicionamento desses medicamentos para tratar simultaneamente várias infecções virais, são necessários mais estudos para avaliar sua eficácia, segurança e possíveis interações medicamentosas em pacientes co-infectados. Por fim, a bioatividade do fitoquímico Nummularine B frente aos vírus SARS-CoV-2, HBV, HCV, dengue e HIV/AIDS, tanto in vitro quanto in vivo, também precisa ser confirmada.

7.7 CONCLUSÃO

No Capítulo VII da minha tese, destaco o uso inovador de modelos de *machine learning* baseados em QSAR como uma ferramenta crucial na descoberta de novos candidatos a fármacos para o tratamento simultâneo de várias doenças virais, incluindo COVID-19, hepatites B e C, dengue e HIV/AIDS.

Os resultados obtidos demonstram que os algoritmos de *machine learning*, especialmente *Extra Trees*, *Random Forest*, *XGBoost*, *HBGB* e *LGBM*, apresentaram excelente desempenho na predição da bioatividade *multi-target*. Esses modelos foram capazes de identificar três compostos promissores: nummularine B, telaprevir e entecavir, cuja eficácia foi confirmada por análises de docking molecular e simulações de dinâmica molecular.

É notável destacar que dois desses compostos, telaprevir e entecavir, já são aprovados pelo FDA para o tratamento das hepatites B e C, o que valida ainda mais a precisão dos nossos achados. Além disso, propomos o reposicionamento desses fármacos para o tratamento de outras doenças investigadas, como COVID-19 e dengue, ampliando seu potencial terapêutico.

Outro ponto relevante é a identificação do nummularine B como um novo candidato a fármaco com potencial para o tratamento simultâneo de várias infecções

virais. Essa descoberta, respaldada pela literatura científica, reforça a importância da abordagem *multi-target* na busca por terapias mais eficazes e abrangentes.

Além disso, destaco a relevância dos modelos baseados em árvores de decisão na predição de bioatividade *multi-target*, ressaltando suas vantagens, como versatilidade, robustez e capacidade de capturar relações complexas entre as variáveis.

Por fim, os mecanismos de ação propostos para os compostos identificados fornecem *insights* importantes sobre suas atividades antivirais, possibilitando uma compreensão mais profunda dos processos biológicos envolvidos e abrindo caminho para o desenvolvimento de novas estratégias terapêuticas.

Em suma, este estudo demonstra o potencial da integração de *machine learning* e inteligência artificial na descoberta de fármacos *multi-target*, oferecendo uma abordagem inovadora e promissora para o desenvolvimento de terapias mais eficazes contra doenças virais.

8 COMENTÁRIOS FINAIS

A presente tese de doutorado foi concebida como uma resposta emergencial à crise sanitária desencadeada pela pandemia da COVID-19. Iniciada no primeiro trimestre de 2020, graças a uma bolsa de estudo emergencial fornecida pela agência CAPES, esta pesquisa se propôs a abordar os desafios impostos pela pandemia ao longo do tempo. Organizada em sete capítulos, cada um desenvolvido em função dos desafios emergentes, esta investigação abrangeu desde os primeiros obstáculos de diagnóstico e identificação de fatores de risco até a busca por soluções terapêuticas e preventivas. A seguir, destacam-se os principais resultados e contribuições de cada capítulo:

Capítulo I: Seu destaque sobre a urgência de estratégias abrangentes para lidar com desafios pandêmicos, como a COVID-19, foi crucial. A ênfase na identificação dos fatores de risco, junto com a proposta de ações proativas, ofereceu uma perspectiva abrangente para enfrentar a pandemia.

Capítulo II: Sua abordagem de identificar alimentos que influenciam positivamente a recuperação da COVID-19 ofereceu *insights* valiosos para complementar abordagens terapêuticas. Essas descobertas ajudaram a informar políticas de saúde pública e diretrizes dietéticas.

Capítulo III: A aplicação da técnica FTIR mostrou ser uma ferramenta custo-benefício na triagem em larga escala para a detecção da COVID-19, especialmente em regiões com recursos limitados. As evidências mostram a viabilidade dessa abordagem e seu potencial impacto na prática clínica, especialmente em regiões com recursos limitados, tendo em vista o seu baixo custo e rapidez nas análises, e por não exigir profissionais altamente qualificados para sua utilização.

Capítulo IV: A utilização de modelos de *machine learning* para predição e diagnóstico da COVID-19, juntamente com a identificação de biomarcadores associados à gravidade da doença, foi uma abordagem inovadora e promissora. Alguns desses biomarcadores identificados neste estudo (por exemplo, a ferritina) foram amplamente aplicados na clínica em todo mundo como biomarcadores de referência para o diagnóstico precoce e o manejo clínico da doença.

Capítulo V: A aplicação de algoritmos de *machine learning* em combinação com as análises metabólica para prever o diagnóstico e a gravidade da COVID-19, juntamente com a identificação de biomarcadores associados, oferece uma

perspectiva valiosa para a gestão da doença. A identificação de novos biomarcadores, adiciona conhecimento significativo ao campo.

Capítulo VI: A investigação de compostos naturais para o tratamento da COVID-19, combinando abordagens de triagem, modelagem molecular, inteligência artificial e *machine learning*, é uma iniciativa promissora. O capítulo destacou a importância de utilização de novas tecnologias de inteligência artificial e Big Data para obtenção e candidatos a fármacos mais promissores para o tratamento da COVID-19 e outras doenças.

Capítulo VII: A utilização de modelos de *machine learning* para a descoberta de fármacos *multi-target* com vista ao tratamento simultâneo de múltiplas doenças (por exemplo SARS-CoV-2, HIV, HBV, HCV, dengue e HIV) é uma abordagem inovadora e relevante, especialmente para doenças virais. A identificação de compostos promissores (nummularine B) e a proposta de reposicionamento de fármacos existentes (telaprevir e entecavir) oferecem *insights* valiosos para o desenvolvimento de terapias mais eficazes.

No geral, a presente tese ofereceu uma contribuição significativa para o entendimento e o combate à COVID-19 durante os últimos três anos, abordando diversos aspectos da pandemia e explorando diferentes estratégias e abordagens. As descobertas tiveram significativo impacto na melhoria e otimização das políticas de saúde pública, orientação de práticas clínicas e impulsiona o desenvolvimento de novas terapias e tratamentos, não apenas para COVID-19, mas também para hepatite B, hepatite C, dengue, HIV/AIDS.

REFERÊNCIAS

1. Sun, J. *et al.* COVID-19: Epidemiology, Evolution, and Cross-Disciplinary Perspectives. *Trends in Molecular Medicine* vol. 26 483–495 Preprint at <https://doi.org/10.1016/j.molmed.2020.02.008> (2020).
2. Zhai, P. *et al.* The epidemiology, diagnosis and treatment of COVID-19. *Int J Antimicrob Agents* **55**, (2020).
3. Ulinici, M. *et al.* Screening, diagnostic and prognostic tests for COVID-19: A comprehensive review. *Life* vol. 11 Preprint at <https://doi.org/10.3390/life11060561> (2021).
4. Li, W. T. *et al.* Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis. *BMC Med Inform Decis Mak* **20**, (2020).
5. Alballa, N. & Al-Turaiki, I. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in Medicine Unlocked* vol. 24 Preprint at <https://doi.org/10.1016/j.imu.2021.100564> (2021).
6. Cobre, A. de F. *et al.* Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? *Comput Biol Med* **134**, (2021).
7. Alexandre De Fátima Cobre, G. *et al.* A multivariate analysis of risk factors associated with death by Covid-19 in the USA, Italy, Spain, and Germany. doi:10.1007/s10389-020-01397-7/Published.
8. Cobre, A. de F. *et al.* Risk factors associated with delay in diagnosis and mortality in patients with covid-19 in the city of rio de janeiro, brazil. *Ciencia e Saude Coletiva* **25**, 4131–4140 (2020).
9. Cobre, A. F. *et al.* Influence of foods and nutrients on COVID-19 recovery: A multivariate analysis of data from 170 countries using a generalized linear model. *Clinical Nutrition* **41**, 3077–3084 (2022).
10. de Fátima Cobre, A. *et al.* Diagnosis and prognosis of COVID-19 employing analysis of patients' plasma and serum via LC-MS and machine learning. *Comput Biol Med* **146**, (2022).
11. de Cobre, A. F. *et al.* MACHINE LEARNING-BASED VIRTUAL SCREENING, MOLECULAR DOCKING, DRUG-LIKENESS, PHARMACOKINETICS AND TOXICITY ANALYSES TO IDENTIFY NEW NATURAL INHIBITORS OF THE

- GLYCOPROTEIN SPIKE (S1) OF SARS-CoV-2. *Quim Nova* **46**, 450–459 (2023).
12. de Fátima Cobre, A. *et al.* Diagnosis and prognosis of COVID-19 employing analysis of patients' plasma and serum via LC-MS and machine learning. *Comput Biol Med* **146**, (2022).
 13. Cobre, A. *et al.* POSB422 Machine Learning-Based Virtual Screening, Molecular Docking and Drug-Likeness to Discover New Inhibitors of the Glycoprotein Spike (S1) of SARS-CoV-2. *Value in Health* **25**, S274 (2022).
 14. Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395**, 497–506 (2020).
 15. Lu, H., Stratton, C. W. & Tang, Y. W. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *Journal of Medical Virology* vol. 92 401–402 Preprint at <https://doi.org/10.1002/jmv.25678> (2020).
 16. Hui, D. S. *et al.* The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health — The latest 2019 novel coronavirus outbreak in Wuhan, China. *International Journal of Infectious Diseases* vol. 91 264–266 Preprint at <https://doi.org/10.1016/j.ijid.2020.01.009> (2020).
 17. World Health Organisation. Clinical management of severe acute respiratory infection when Novel coronavirus (nCoV) infection is suspected: interim guidance. [https://www.who.int/internalpublications-detail/clinical-management-of-severe-acute-respiratoryinfection-when-novel-coronavirus-\(ncov\)-infection-is-suspected](https://www.who.int/internalpublications-detail/clinical-management-of-severe-acute-respiratoryinfection-when-novel-coronavirus-(ncov)-infection-is-suspected) (2020).
 18. Chen, N. *et al.* Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The Lancet* **395**, 507–513 (2020).
 19. Rodriguez-Morales, A. J. *et al.* COVID-19 in Latin America: The implications of the first confirmed case in Brazil. *Travel Medicine and Infectious Disease* vol. 35 Preprint at <https://doi.org/10.1016/j.tmaid.2020.101613> (2020).
 20. Cucinotta, D. & Vanelli, M. WHO declares COVID-19 a pandemic. *Acta Biomedica* vol. 91 157–160 Preprint at <https://doi.org/10.23750/abm.v91i1.9397> (2020).
 21. World Health Organisation. WHO Director-General's statement on IHR Emergency Committee on Novel Coronavirus (2019-nCoV). <https://www.who.int/director-general/speeches/detail/who-director-general-s->

- statement-on-ihf-emergency-committee-on-novel-coronavirus-(2019-ncov)* (2020).
22. Wang, D. *et al.* Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China. *JAMA - Journal of the American Medical Association* **323**, 1061–1069 (2020).
 23. Zhang, J. jin *et al.* Clinical characteristics of 140 patients infected with SARS-CoV-2 in Wuhan, China. *Allergy: European Journal of Allergy and Clinical Immunology* **75**, 1730–1741 (2020).
 24. Hodgson, S. H. *et al.* What defines an efficacious COVID-19 vaccine? A review of the challenges assessing the clinical efficacy of vaccines against SARS-CoV-2. *The Lancet Infectious Diseases* vol. 21 e26–e35 Preprint at [https://doi.org/10.1016/S1473-3099\(20\)30773-8](https://doi.org/10.1016/S1473-3099(20)30773-8) (2021).
 25. World Health Organization. Number of COVID-19 cases reported to WHO (cumulative total). <https://data.who.int/dashboards/covid19/cases?n=c> (2024).
 26. World Health Organization. Number of COVID-19 deaths reported to WHO (cumulative total). <https://data.who.int/dashboards/covid19/deaths?n=c> (2024).
 27. Notarte, K. I. *et al.* Age, Sex and Previous Comorbidities as Risk Factors Not Associated with SARS-CoV-2 Infection for Long COVID-19: A Systematic Review and Meta-Analysis. *Journal of Clinical Medicine* vol. 11 Preprint at <https://doi.org/10.3390/jcm11247314> (2022).
 28. Starke, K. R. *et al.* The age-related risk of severe outcomes due to covid-19 infection: A rapid review, meta-analysis, and meta-regression. *Int J Environ Res Public Health* **17**, 1–24 (2020).
 29. Patel, U. *et al.* Age-Adjusted Risk Factors Associated with Mortality and Mechanical Ventilation Utilization Amongst COVID-19 Hospitalizations—a Systematic Review and Meta-Analysis. *SN Compr Clin Med* **2**, 1740–1749 (2020).
 30. Moazzami, B. *et al.* Metabolic risk factors and risk of Covid-19: A systematic review and meta-analysis. *PLoS ONE* vol. 15 Preprint at <https://doi.org/10.1371/journal.pone.0243600> (2020).
 31. Khanijahani, A., Iezadi, S., Gholipour, K., Azami-Aghdash, S. & Naghibi, D. A systematic review of racial/ethnic and socioeconomic disparities in COVID-19. *International Journal for Equity in Health* vol. 20 Preprint at <https://doi.org/10.1186/s12939-021-01582-4> (2021).

32. Magesh, S. *et al.* Disparities in COVID-19 Outcomes by Race, Ethnicity, and Socioeconomic Status: A Systematic-Review and Meta-analysis. *JAMA Network Open* vol. 4 Preprint at <https://doi.org/10.1001/jamanetworkopen.2021.34147> (2021).
33. Mackey, K. *et al.* Racial and ethnic disparities in covid-19-related infections, hospitalizations, and deaths a systematic review. *Annals of Internal Medicine* vol. 174 362–373 Preprint at <https://doi.org/10.7326/M20-6306> (2021).
34. Akbarialiabad, H. *et al.* Long COVID, a comprehensive systematic scoping review. *Infection* vol. 49 1163–1186 Preprint at <https://doi.org/10.1007/s15010-021-01666-x> (2021).
35. Chen, C. *et al.* Global Prevalence of Post-Coronavirus Disease 2019 (COVID-19) Condition or Long COVID: A Meta-Analysis and Systematic Review. *Journal of Infectious Diseases* **226**, 1593–1607 (2022).
36. De Miranda, D. A. P. *et al.* Long COVID-19 syndrome: A 14-months longitudinal study during the two first epidemic peaks in Southeast Brazil. *Trans R Soc Trop Med Hyg* **116**, 1007–1014 (2022).
37. Notarte, K. I. *et al.* Age, Sex and Previous Comorbidities as Risk Factors Not Associated with SARS-CoV-2 Infection for Long COVID-19: A Systematic Review and Meta-Analysis. *Journal of Clinical Medicine* vol. 11 Preprint at <https://doi.org/10.3390/jcm11247314> (2022).
38. Grandjean, L. *et al.* Long-Term Persistence of Spike Protein Antibody and Predictive Modeling of Antibody Dynamics after Infection with Severe Acute Respiratory Syndrome Coronavirus 2. *Clinical Infectious Diseases* **74**, 1220–1229 (2022).
39. Theoharides, T. C. Could SARS-CoV-2 Spike Protein Be Responsible for Long-COVID Syndrome? *Molecular Neurobiology* vol. 59 1850–1861 Preprint at <https://doi.org/10.1007/s12035-021-02696-0> (2022).
40. Raveendran, A. V., Jayadevan, R. & Sashidharan, S. Long COVID: An overview. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* vol. 15 869–875 Preprint at <https://doi.org/10.1016/j.dsx.2021.04.007> (2021).
41. Carobene, A., Milella, F., Famigliani, L. & Cabitza, F. How is test laboratory data used and characterised by machine learning models? A systematic review of diagnostic and prognostic models developed for COVID-19 patients using only

- laboratory data. *Clinical Chemistry and Laboratory Medicine* vol. 60 1887–1901 Preprint at <https://doi.org/10.1515/cclm-2022-0182> (2022).
42. Fan, Y., Liu, M. & Sun, G. An interpretable machine learning framework for diagnosis and prognosis of COVID-19. *PLoS One* **18**, (2023).
 43. Zhang, J. *et al.* Prediction of individual COVID-19 diagnosis using baseline demographics and lab data. *Sci Rep* **11**, (2021).
 44. Linardatos, P., Papastefanopoulos, V. & Kotsiantis, S. Explainable ai: A review of machine learning interpretability methods. *Entropy* vol. 23 1–45 Preprint at <https://doi.org/10.3390/e23010018> (2021).
 45. Li, W. T. *et al.* Using machine learning of clinical data to diagnose COVID-19: A systematic review and meta-analysis. *BMC Med Inform Decis Mak* **20**, (2020).
 46. Brinati, D. *et al.* Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. *J Med Syst* **44**, (2020).
 47. Weiss, P. & Murdoch, D. R. Clinical course and mortality risk of severe COVID-19. *The Lancet* vol. 395 1014–1015 Preprint at [https://doi.org/10.1016/S0140-6736\(20\)30633-4](https://doi.org/10.1016/S0140-6736(20)30633-4) (2020).
 48. Wu, J. *et al.* Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. doi:10.1101/2020.04.02.20051136.
 49. Banerjee, A. *et al.* Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. *Int Immunopharmacol* **86**, (2020).
 50. Kermali, M., Khalsa, R. K., Pillai, K., Ismail, Z. & Harky, A. The role of biomarkers in diagnosis of COVID-19 – A systematic review. *Life Sciences* vol. 254 Preprint at <https://doi.org/10.1016/j.lfs.2020.117788> (2020).
 51. Domingues, K. Z. A. *et al.* Systematic review and evidence gap mapping of biomarkers associated with neurological manifestations in patients with COVID-19. *Journal of Neurology* Preprint at <https://doi.org/10.1007/s00415-023-12090-6> (2023).
 52. Mulchandani, R., Lyngdoh, T. & Kakkar, A. K. Deciphering the COVID-19 cytokine storm: Systematic review and meta-analysis. *European Journal of Clinical Investigation* vol. 51 Preprint at <https://doi.org/10.1111/eci.13429> (2021).

53. Melo, A. K. G. *et al.* Biomarkers of cytokine storm as red flags for severe and fatal COVID-19 cases: A living systematic review and meta-analysis. *PLoS One* **16**, (2021).
54. Struck-Lewicka, W. *et al.* Urine metabolic fingerprinting using LC-MS and GC-MS reveals metabolite changes in prostate cancer: A pilot study. *J Pharm Biomed Anal* **111**, 351–361 (2015).
55. Liu, X. *et al.* LC-MS-Based Plasma Metabolomics and Lipidomics Analyses for Differential Diagnosis of Bladder Cancer and Renal Cell Carcinoma. *Front Oncol* **10**, (2020).
56. Lubes, G. & Goodarzi, M. GC-MS based metabolomics used for the identification of cancer volatile organic compounds as biomarkers. *Journal of Pharmaceutical and Biomedical Analysis* vol. 147 313–322 Preprint at <https://doi.org/10.1016/j.jpba.2017.07.013> (2018).
57. Emwas, A. H. M., Salek, R. M., Griffin, J. L. & Merzaban, J. NMR-based metabolomics in human disease diagnosis: Applications, limitations, and recommendations. *Metabolomics* vol. 9 1048–1072 Preprint at <https://doi.org/10.1007/s11306-013-0524-y> (2013).
58. Spick, M. *et al.* Systematic review with meta-analysis of diagnostic test accuracy for COVID-19 by mass spectrometry. *Metabolism* **126**, (2022).
59. Pang, Z., Zhou, G., Chong, J. & Xia, J. Comprehensive meta-analysis of covid-19 global metabolomics datasets. *Metabolites* **11**, 1–14 (2021).
60. Ruszkiewicz, D. M. *et al.* Diagnosis of COVID-19 by analysis of breath with gas chromatography-ion mobility spectrometry - a feasibility study. *EClinicalMedicine* **29–30**, (2020).
61. Steel, J. J. *et al.* Empowering academic labs and scientists to test for COVID-19. *BioTechniques* vol. 69 245–248 Preprint at <https://doi.org/10.2144/BTN-2020-0079> (2020).
62. Pokhrel, P., Hu, C. & Mao, H. Detecting the coronavirus (CoVID-19). *ACS Sensors* vol. 5 2283–2297 Preprint at <https://doi.org/10.1021/ACSSENSORS.0C01153> (2020).
63. Zhou, L. *et al.* Programmable Low-Cost DNA-Based Platform for Viral RNA Detection. *Sci. Adv* vol. 6 <https://www.science.org> (2020).
64. Morel, C. M., Lindahl, O. M. & Ozenci, V. Lessons from COVID-19 on the role of the state and the market in providing early testing. *J Glob Health* **10**, 1–6 (2020).

65. Scohy, A. *et al.* Low performance of rapid antigen detection test as frontline testing for COVID-19 diagnosis. *Journal of Clinical Virology* **129**, (2020).
66. Gremmels, H. *et al.* Real-life validation of the Panbio™ COVID-19 antigen rapid test (Abbott) in community-dwelling subjects with symptoms of potential SARS-CoV-2 infection. *EClinicalMedicine* **31**, (2021).
67. Fitzpatrick, M. C., Pandey, A., Wells, C. R., Sah, P. & Galvani, A. P. Buyer beware: inflated claims of sensitivity for rapid COVID-19 tests. *The Lancet* vol. 397 24–25 Preprint at [https://doi.org/10.1016/S0140-6736\(20\)32635-0](https://doi.org/10.1016/S0140-6736(20)32635-0) (2021).
68. Fitzpatrick, M. C., Pandey, A., Wells, C. R., Sah, P. & Galvani, A. P. Buyer beware: inflated claims of sensitivity for rapid COVID-19 tests. *The Lancet* vol. 397 24–25 Preprint at [https://doi.org/10.1016/S0140-6736\(20\)32635-0](https://doi.org/10.1016/S0140-6736(20)32635-0) (2021).
69. Baker, M. J. *et al.* Using Fourier transform IR spectroscopy to analyze biological materials. *Nat Protoc* **9**, 1771–1791 (2014).
70. Mitchell, A. L., Gajjar, K. B., Theophilou, G., Martin, F. L. & Martin-Hirsch, P. L. Vibrational spectroscopy of biofluids for disease screening or diagnosis: Translation from the laboratory to a clinical setting. *J Biophotonics* **7**, 153–165 (2014).
71. Movasaghi, Z., Rehman, S. & Rehman, I. U. Fourier transform infrared (FTIR) spectroscopy of biological tissues. *Applied Spectroscopy Reviews* vol. 43 134–179 Preprint at <https://doi.org/10.1080/05704920701829043> (2008).
72. Martin, F. L. *et al.* Distinguishing cell types or populations based on the computational analysis of their infrared spectra. *Nat Protoc* **5**, 1748–1760 (2010).
73. Maitra, I. *et al.* Attenuated total reflection Fourier-transform infrared spectral discrimination in human bodily fluids of oesophageal transformation to adenocarcinoma. *Analyst* **144**, 7447–7456 (2019).
74. Pinals, R. L. *et al.* Rapid SARS-CoV-2 Spike Protein Detection by Carbon Nanotube-Based Near-Infrared Nanosensors. *Nano Lett* **21**, 2272–2280 (2021).
75. Guleken, Z. *et al.* Development of novel spectroscopic and machine learning methods for the measurement of periodic changes in COVID-19 antibody level. *Measurement (Lond)* **196**, (2022).
76. Khan, R. S. & Rehman, I. U. Spectroscopy as a tool for detection and monitoring of Coronavirus (COVID-19). *Expert Review of Molecular Diagnostics* vol. 20 647–649 Preprint at <https://doi.org/10.1080/14737159.2020.1766968> (2020).

77. Cawood, A. L., Walters, E. R., Smith, T. R., Sipaul, R. H. & Stratton, R. J. A review of nutrition support guidelines for individuals with or recovering from COVID-19 in the community. *Nutrients* vol. 12 1–13 Preprint at <https://doi.org/10.3390/nu12113230> (2020).
78. Moisey, L. L., Merriweather, J. L. & Drover, J. W. The role of nutrition rehabilitation in the recovery of survivors of critical illness: underrecognized and underappreciated. *Critical Care* vol. 26 Preprint at <https://doi.org/10.1186/s13054-022-04143-5> (2022).
79. Mechanick, J. I. *et al.* Clinical Nutrition Research and the COVID-19 Pandemic: A Scoping Review of the ASPEN COVID-19 Task Force on Nutrition Research. *Journal of Parenteral and Enteral Nutrition* vol. 45 13–31 Preprint at <https://doi.org/10.1002/jpen.2036> (2021).
80. CLINICAL NUTRITION JOURNAL. Call for papers: Nutritional status and nutritional care in COVID-19 patients. <https://www.clinicalnutritionjournal.com/call-for-papers-covid> (2021).
81. Latif, J. *et al.* Strategies to ensure continuity of nutritional care in patients with COVID-19 infection on discharge from hospital: A rapid review. *Clin Nutr ESPEN* **47**, 106–116 (2022).
82. Wang, Y., Wang, Y., Chen, Y. & Qin, Q. Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. *Journal of Medical Virology* vol. 92 568–576 Preprint at <https://doi.org/10.1002/jmv.25748> (2020).
83. Kluge, H. H. P. *et al.* Prevention and control of non-communicable diseases in the COVID-19 response. *The Lancet* vol. 395 1678–1680 Preprint at [https://doi.org/10.1016/S0140-6736\(20\)31067-9](https://doi.org/10.1016/S0140-6736(20)31067-9) (2020).
84. Belanger, M. J. *et al.* Covid-19 and Disparities in Nutrition and Obesity. *New England Journal of Medicine* **383**, e69 (2020).
85. Abdalazim, A. & Albashir, D. The potential impacts of obesity on COVID-19. *Clinical Medicine* **20**, (2020).
86. Price-Haywood, E. G., Burton, J., Fort, D. & Seoane, L. Hospitalization and Mortality among Black Patients and White Patients with Covid-19. *New England Journal of Medicine* **382**, 2534–2543 (2020).

87. Butler, M. J. & Barrientos, R. M. The impact of nutrition on COVID-19 susceptibility and long-term consequences. *Brain, Behavior, and Immunity* vol. 87 53–54 Preprint at <https://doi.org/10.1016/j.bbi.2020.04.040> (2020).
88. Balbi, M. E. *et al.* Antioxidant effects of vitamins in type 2 diabetes: A meta-analysis of randomized controlled trials. *Diabetol Metab Syndr* **10**, (2018).
89. Tonin, F. S. *et al.* Impact of natural juice consumption on plasma antioxidant status: A systematic review and meta-analysis. *Molecules* **20**, 22146–22156 (2015).
90. Ligon, B. L. Penicillin: Its Discovery and Early Development. *Semin Pediatr Infect Dis* **15**, 52–57 (2004).
91. *Drug Discovery Today: HTS Supplement | Reviews*. (2000).
92. Vemula, D., Jayasurya, P., Sushmitha, V., Kumar, Y. N. & Bhandari, V. CADD, AI and ML in drug discovery: A comprehensive review. *European Journal of Pharmaceutical Sciences* vol. 181 Preprint at <https://doi.org/10.1016/j.ejps.2022.106324> (2023).
93. Jones, A. W. Early drug discovery and the rise of pharmaceutical chemistry. *Drug Test Anal* **3**, 337–344 (2011).
94. Chaudhary, K. K. & Mishra, N. Central Bringing Excellence in Open Access OPEN ACCESS A Review on Molecular Docking: Novel Tool for Drug Discovery. *A Review on Molecular Docking: Novel Tool for Drug Discovery. JSM Chem* **4**, 1029 (2016).
95. Muhammad, U., Uzairu, A. & Arthur, D. E. REVIEW ON: QUANTITATIVE STRUCTURE ACTIVITY RELATIONSHIP (QSAR) MODELING. *International Journal of Advanced Academic Research | Sciences* vol. 4 (2018).
96. Hospital, A., Goñi, J. R., Orozco, M. & Gelpí, J. L. Molecular dynamics simulations: Advances and applications. *Advances and Applications in Bioinformatics and Chemistry* vol. 8 37–47 Preprint at <https://doi.org/10.2147/AABC.S70333> (2015).
97. Moroy, G., Martiny, V. Y., Vayer, P., Villoutreix, B. O. & Miteva, M. A. Toward in silico structure-based ADMET prediction in drug discovery. *Drug Discovery Today* vol. 17 44–55 Preprint at <https://doi.org/10.1016/j.drudis.2011.10.023> (2012).
98. Singh, M. B. *et al.* An understanding of coronavirus and exploring the molecular dynamics simulations to find promising candidates against the Mpro of nCoV to

- combat the COVID-19: A systematic review. *Journal of Infection and Public Health* vol. 15 1326–1349 Preprint at <https://doi.org/10.1016/j.jiph.2022.10.013> (2022).
99. Mohamed, K., Yazdanpanah, N., Saghazadeh, A. & Rezaei, N. Computational drug discovery and repurposing for the treatment of COVID-19: A systematic review. *Bioorganic Chemistry* vol. 106 Preprint at <https://doi.org/10.1016/j.bioorg.2020.104490> (2021).
100. Fadlalla, M., Ahmed, M., Ali, M., Elshiekh, A. A. & Yousef, B. A. Molecular Docking as a Potential Approach in Repurposing Drugs Against COVID-19: a Systematic Review and Novel Pharmacophore Models. *Current Pharmacology Reports* vol. 8 212–226 Preprint at <https://doi.org/10.1007/s40495-022-00285-w> (2022).
101. Prajapat, M. *et al.* Virtual screening and molecular dynamics study of approved drugs as inhibitors of spike protein S1 domain and ACE2 interaction in SARS-CoV-2. *J Mol Graph Model* **101**, (2020).
102. Ramesh, A. N., Kambhampati, C., Monson, J. R. T. & Drew, P. J. Artificial intelligence in medicine. *Annals of the Royal College of Surgeons of England* vol. 86 334–338 Preprint at <https://doi.org/10.1308/147870804290> (2004).
103. Miles, J. C. & Walker, A. J. The potential application of artificial intelligence in transport. *IEE Proceedings: Intelligent Transport Systems* **153**, (2006).
104. Wirtz, B. W., Weyerer, J. C. & Geyer, C. Artificial Intelligence and the Public Sector—Applications and Challenges. *International Journal of Public Administration* **42**, 596–615 (2019).
105. Siau, K. *A Qualitative Research on Marketing and Sales in the Artificial Intelligence Age*. <http://aisel.aisnet.org/mwais2018/41> (2018).
106. Lamberti, M. J. *et al.* A Study on the Application and Use of Artificial Intelligence to Support Drug Development. *Clin Ther* **41**, 1414–1426 (2019).
107. Smith, R. G. & Farquhar, A. *The Road Ahead for Knowledge Management: An AI Perspective*. (2000).
108. Mak, K. K. & Pichika, M. R. Artificial intelligence in drug development: present status and future prospects. *Drug Discovery Today* vol. 24 773–780 Preprint at <https://doi.org/10.1016/j.drudis.2018.11.014> (2019).

109. Sellwood, M. A., Ahmed, M., Segler, M. H. S. & Brown, N. Artificial intelligence in drug discovery. *Future Medicinal Chemistry* vol. 10 2025–2028 Preprint at <https://doi.org/10.4155/fmc-2018-0212> (2018).
110. Paul, D. *et al.* Artificial intelligence in drug discovery and development. *Drug Discovery Today* vol. 26 80–93 Preprint at <https://doi.org/10.1016/j.drudis.2020.10.010> (2021).
111. Zhu, H. Big Data and Artificial Intelligence Modeling for Drug Discovery. *Annu Rev Pharmacol Toxicol* (2019) doi:10.1146/annurev-pharmtox-010919.
112. Ciallella, H. L. & Zhu, H. Advancing Computational Toxicology in the Big Data Era by Artificial Intelligence: Data-Driven and Mechanism-Driven Modeling for Chemical Toxicity. *Chem Res Toxicol* **32**, 536–547 (2019).
113. Chan, H. C. S., Shan, H., Dahoun, T., Vogel, H. & Yuan, S. Advancing Drug Discovery via Artificial Intelligence. *Trends in Pharmacological Sciences* vol. 40 592–604 Preprint at <https://doi.org/10.1016/j.tips.2019.06.004> (2019).
114. Sellwood, M. A., Ahmed, M., Segler, M. H. S. & Brown, N. Artificial intelligence in drug discovery. *Future Medicinal Chemistry* vol. 10 2025–2028 Preprint at <https://doi.org/10.4155/fmc-2018-0212> (2018).
115. Pereira, J. C., Caffarena, E. R. & Dos Santos, C. N. Boosting Docking-Based Virtual Screening with Deep Learning. *J Chem Inf Model* **56**, 2495–2506 (2016).
116. Firth, N. C., Atrash, B., Brown, N. & Blagg, J. MOARF, an Integrated Workflow for Multiobjective Optimization: Implementation, Synthesis, and Biological Evaluation. *J Chem Inf Model* **55**, 1169–1180 (2015).
117. Zhang, L., Tan, J., Han, D. & Zhu, H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discovery Today* vol. 22 1680–1685 Preprint at <https://doi.org/10.1016/j.drudis.2017.08.010> (2017).
118. Wang, Y. *et al.* A comparative study of family-specific protein-ligand complex affinity prediction based on random forest approach. *J Comput Aided Mol Des* **29**, 349–360 (2015).
119. King, R. D., Hirst, J. D. & Sternberg, M. J. E. Comparison of artificial intelligence methods for modeling pharmaceutical QSARS. *Applied Artificial Intelligence* **9**, 213–233 (1995).
120. Álvarez-Machancoses, Ó. & Fernández-Martínez, J. L. Using artificial intelligence methods to speed up drug discovery. *Expert Opinion on Drug*

- Discovery* vol. 14 769–777 Preprint at <https://doi.org/10.1080/17460441.2019.1621284> (2019).
121. Dana, D. *et al.* Deep learning in drug discovery and medicine; scratching the surface. *Molecules* vol. 23 Preprint at <https://doi.org/10.3390/molecules23092384> (2018).
122. Mak, K. K. & Pichika, M. R. Artificial intelligence in drug development: present status and future prospects. *Drug Discovery Today* vol. 24 773–780 Preprint at <https://doi.org/10.1016/j.drudis.2018.11.014> (2019).
123. Zang, Q. *et al.* *In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning.* <http://pubs.acs.org> (2016).
124. Hessler, G. & Baringhaus, K. H. Artificial intelligence in drug design. *Molecules* vol. 23 Preprint at <https://doi.org/10.3390/molecules23102520> (2018).
125. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chemical Reviews* vol. 119 10520–10594 Preprint at <https://doi.org/10.1021/acs.chemrev.8b00728> (2019).
126. Kumar, R., Sharma, A., Siddiqui, M. H. & Tiwari, R. K. Prediction of Human Intestinal Absorption of Compounds Using Artificial Intelligence Techniques. *Curr Drug Discov Technol* **14**, (2017).
127. Chai, S. *et al.* A grand product design model for crystallization solvent design. *Comput Chem Eng* **135**, (2020).
128. Thafar, M., Raies, A. Bin, Albaradei, S., Essack, M. & Bajic, V. B. Comparison Study of Computational Prediction Tools for Drug-Target Binding Affinities. *Frontiers in Chemistry* vol. 7 Preprint at <https://doi.org/10.3389/fchem.2019.00782> (2019).
129. Öztürk, H., Özgür, A. & Ozkirimli, E. DeepDTA: Deep drug-target binding affinity prediction. in *Bioinformatics* vol. 34 i821–i829 (Oxford University Press, 2018).
130. Lounkine, E. *et al.* Large-scale prediction and testing of drug activity on side-effect targets. *Nature* **486**, 361–367 (2012).
131. Karimi, M., Wu, D., Wang, Z. & Shen, Y. DeepAffinity: Interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* **35**, 3329–3338 (2019).
132. Feng, Q., Dueva, E., Cherkasov, A. & Ester, M. PADME: A Deep Learning-based Framework for Drug-Target Interaction Prediction. (2018).

133. Pu, L. *et al.* EToxPred: A machine learning-based approach to estimate the toxicity of drug candidates 03 Chemical Sciences 0305 Organic Chemistry 03 Chemical Sciences 0304 Medicinal and Biomolecular Chemistry. *BMC Pharmacol Toxicol* **20**, (2019).
134. Yang, X., Wang, Y., Byrne, R., Schneider, G. & Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chemical Reviews* vol. 119 10520–10594 Preprint at <https://doi.org/10.1021/acs.chemrev.8b00728> (2019).
135. Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. DeepTox: Toxicity prediction using deep learning. *Front Environ Sci* **3**, (2016).
136. Lysenko, A., Sharma, A., Boroevich, K. A. & Tsunoda, T. An integrative machine learning approach for prediction of toxicity-related drug safety. *Life Sci Alliance* **1**, (2018).
137. Basile, A. O., Yah, A. & Tatonetti, N. P. Artificial Intelligence for Drug Toxicity and Safety. *Trends in Pharmacological Sciences* vol. 40 624–635 Preprint at <https://doi.org/10.1016/j.tips.2019.07.005> (2019).
138. Gayvert, K. M., Madhukar, N. S. & Elemento, O. A Data-Driven Approach to Predicting Successes and Failures of Clinical Trials. *Cell Chem Biol* **23**, 1294–1301 (2016).
139. Jimenez-Carretero, D. *et al.* Tox_(R)CNN: Deep learning-based nuclei profiling tool for drug toxicity screening. *PLoS Comput Biol* **14**, (2018).
140. Park, M., Cook, A. R., Lim, J. T., Sun, Y. & Dickens, B. L. A systematic review of covid-19 epidemiology based on current evidence. *Journal of Clinical Medicine* vol. 9 Preprint at <https://doi.org/10.3390/jcm9040967> (2020).
141. Zeng, N. *et al.* A systematic review and meta-analysis of long term physical and mental sequelae of COVID-19 pandemic: call for research priority and action. *Molecular Psychiatry* vol. 28 423–433 Preprint at <https://doi.org/10.1038/s41380-022-01614-7> (2023).
142. Zhang, J. jin, Dong, X., Liu, G. hui & Gao, Y. dong. Risk and Protective Factors for COVID-19 Morbidity, Severity, and Mortality. *Clinical Reviews in Allergy and Immunology* vol. 64 90–107 Preprint at <https://doi.org/10.1007/s12016-022-08921-5> (2023).
143. Kaplan, E. L. & Meier, P. *Nonparametric Estimation from Incomplete Observations* NONPARAMETRIC ESTIMATION FROM INCOMPLETE

- OBSERVATIONS*. Source: *Journal of the American Statistical Association* vol. 53 (1958).
144. Cox, D. R. *Regression Models and Life-Tables. Research Section, on Wednesday* (1972).
 145. Alfaro, M. & Huelsenbeck, J. Comparative performance of Bayesian and AIC-based measures of phylogenetic model uncertainty. *Syst Biol* **55**, 89–96 (2006).
 146. COVID-19, Australia: Epidemiology Report 16 (Reporting week to 23:59 AEST 17 May 2020). *Commun Dis Intell* (2018) **44**, (2020).
 147. Gao, Q. *et al.* The epidemiological characteristics of 2019 novel coronavirus diseases (COVID-19) in Jingmen, Hubei, China. *Medicine (United States)* **99**, (2020).
 148. Wei, X. *et al.* *Sex Differences in Severity and Mortality Among Patients With COVID-19: Evidence from Pooled Literature Analysis and Insights from Integrated Bioinformatic Analysis.*
 149. Cai, H. Sex difference and smoking predisposition in patients with COVID-19. *The Lancet Respiratory Medicine* vol. 8 e20 Preprint at [https://doi.org/10.1016/S2213-2600\(20\)30117-X](https://doi.org/10.1016/S2213-2600(20)30117-X) (2020).
 150. Wei, X. *et al.* *Sex Differences in Severity and Mortality Among Patients With COVID-19: Evidence from Pooled Literature Analysis and Insights from Integrated Bioinformatic Analysis.*
 151. Sharma, G., Volgman, A. S. & Michos, E. D. Sex Differences in Mortality From COVID-19 Pandemic. *JACC Case Rep* **2**, 1407–1410 (2020).
 152. Single-Cell RNA Expression Profiling of ACE2, the Receptor of SARS-CoV-2. doi:10.1101/2020.01.26.919985.
 153. Laurenti, R., Helena Prado de Mello Jorge, M. & Léa Davidson Gotlieb, S. *Perfil Epidemiológico Da Morbi-Mortalidade Masculina Epidemiological Profile of Men: Morbidity and Mortality.*
 154. Wei, X. *et al.* *Sex Differences in Severity and Mortality Among Patients With COVID-19: Evidence from Pooled Literature Analysis and Insights from Integrated Bioinformatic Analysis.*
 155. Coletiva, S., Basílio da Gama, R. & Paim, J. The Brazilian health system: history, advances, and challenges. *Lancet* **377**, 1778–97 (2011).

156. Mendes, W., Martins, M., Rozenfeld, S. & Travassos, C. The assessment of adverse events in hospitals in Brazil. *International Journal for Quality in Health Care* **21**, 279–284 (2009).
157. Dos Santos, J. P. C., Siqueira, A. S. P., Praça, H. L. F. & Albuquerque, H. G. Vulnerability to severe forms of COVID-19: An intra-municipal analysis in the city of Rio de Janeiro, Brazil. *Cad Saude Publica* **36**, (2020).
158. Corburn, J. *et al.* Slum Health: Arresting COVID-19 and Improving Well-Being in Urban Informal Settlements. *Journal of Urban Health* **97**, 348–357 (2020).
159. Riley, L. W., Ko, A. I., Unger, A. & Reis, M. G. Slum health: Diseases of neglected populations. *BMC International Health and Human Rights* vol. 7 Preprint at <https://doi.org/10.1186/1472-698X-7-2> (2007).
160. Chen, Y. *et al.* A sensitive and specific antigen detection assay for Middle East respiratory syndrome coronavirus. *Emerg Microbes Infect* **4**, 1–5 (2015).
161. Reis, R. F. *et al.* Characterization of the COVID-19 pandemic and the impact of uncertainties, mitigation strategies, and underreporting of cases in South Korea, Italy, and Brazil. *Chaos Solitons Fractals* **136**, (2020).
162. Ribeiro, L. C. & Bernardes, A. T. *Estimate of Underreporting of COVID-19 in Brazil by Acute Respiratory Syndrome Hospitalization Reports.* https://cmmid.github.io/topics/covid19/severity/global_cfr_estimates.html.
163. Phua, J. *et al.* Intensive care management of coronavirus disease 2019 (COVID-19): challenges and recommendations. *The Lancet Respiratory Medicine* vol. 8 506–517 Preprint at [https://doi.org/10.1016/S2213-2600\(20\)30161-2](https://doi.org/10.1016/S2213-2600(20)30161-2) (2020).
164. Remuzzi, A. & Remuzzi, G. COVID-19 and Italy: what next? *The Lancet* vol. 395 1225–1228 Preprint at [https://doi.org/10.1016/S0140-6736\(20\)30627-9](https://doi.org/10.1016/S0140-6736(20)30627-9) (2020).
165. Di Lorenzo, G. & Di Trolio, R. Coronavirus Disease (COVID-19) in Italy: Analysis of Risk Factors and Proposed Remedial Measures. *Front Med (Lausanne)* **7**, (2020).
166. Verelst, F., Kuylen, E. & Beutels, P. Indications for healthcare surge capacity in European countries facing an exponential increase in coronavirus disease (COVID-19) cases, March 2020. *Eurosurveillance* **25**, (2020).
167. Rhodes, A. *et al.* The variability of critical care bed numbers in Europe. *Intensive Care Med* **38**, 1647–1653 (2012).

168. Akkız, H. The Biological Functions and Clinical Significance of SARS-CoV-2 Variants of Concern. *Frontiers in Medicine* vol. 9 Preprint at <https://doi.org/10.3389/fmed.2022.849217> (2022).
169. Meng, L. *et al.* Intubation and Ventilation amid the COVID-19 Outbreak: Wuhan's Experience. *Anesthesiology* 1317–1332 (2020) doi:10.1097/ALN.0000000000003296.
170. Whittle, J. S., Pavlov, I., Sacchetti, A. D., Atwood, C. & Rosenberg, M. S. Respiratory support for adult patients with COVID-19. *JACEP Open* 1, 95–101 (2020).
171. Rosenbaum, L. Harnessing Our Humanity — How Washington's Health Care Workers Have Risen to the Pandemic Challenge. *New England Journal of Medicine* **382**, 2069–2071 (2020).
172. Arabi, Y. M., Murthy, S. & Webb, S. COVID-19: a novel coronavirus and a novel challenge for critical care. *Intensive Care Med* **46**, 833–836 (2020).
173. Mattos, A. M. de *et al.* Fake News em tempos de COVID-19 e seu tratamento jurídico no ordenamento brasileiro. *Escola Anna Nery* **25**, (2021).
174. Mendelson, M. *et al.* The political theatre of the UK's travel ban on South Africa. *The Lancet* **398**, 2211–2213 (2021).
175. Uddin, M. N. & Roni, M. A. Challenges of storage and stability of mrna-based covid-19 vaccines. *Vaccines* vol. 9 Preprint at <https://doi.org/10.3390/vaccines9091033> (2021).
176. Khubchandani, J. *et al.* COVID-19 Vaccination Hesitancy in the United States: A Rapid National Assessment. *J Community Health* **46**, 270–277 (2021).
177. Nachege, J. B. *et al.* Addressing challenges to rolling out COVID-19 vaccines in African countries. *The Lancet Global Health* vol. 9 e746–e748 Preprint at [https://doi.org/10.1016/S2214-109X\(21\)00097-8](https://doi.org/10.1016/S2214-109X(21)00097-8) (2021).
178. Mechanick, J. I. *et al.* Clinical Nutrition Research and the COVID-19 Pandemic: A Scoping Review of the ASPEN COVID-19 Task Force on Nutrition Research. *Journal of Parenteral and Enteral Nutrition* vol. 45 13–31 Preprint at <https://doi.org/10.1002/jpen.2036> (2021).
179. Maria Ren. COVID-19 Healthy Diet Dataset. <https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset> (2020).
180. Food and Agriculture Organization of the United Nations. FAOSTAT. <https://www.fao.org/faostat/en/> (2021).

181. PRB. Disasters raise risk of a homeless undercount in 2020 Census. <https://www.prb.org/> (2021).
182. Johns Hopkins University & Medicine. Coronavirus Resource Center. <https://coronavirus.jhu.edu/map.html>.
183. World Bank. World Development Indicators - World Bank Collection. <https://datacatalog.worldbank.org/search/dataset/0037712> (2021).
184. Saumard, A. & Navarro, F. Finite sample improvement of Akaike's Information Criterion. (2018).
185. Bozdogan, H. *THE GENERAL THEORY AND ITS ANALYTICAL EXTENSIONS. PSYCHOMETRIKA* vol. 52 (1987).
186. Pan, W. *Akaike's Information Criterion in Generalized Estimating Equations*. (2001).
187. Arnold, T. W. Uninformative Parameters and Model Selection Using Akaike's Information Criterion. *Journal of Wildlife Management* **74**, 1175–1178 (2010).
188. Zabetakis, I., Lordan, R., Norton, C. & Tsoupras, A. Covid-19: The inflammation link and the role of nutrition in potential mitigation. *Nutrients* vol. 12 Preprint at <https://doi.org/10.3390/nu12051466> (2020).
189. Booth, S. L., Johns, T. & Kuhnlein, H. V. *Natural Food Sources of Vitamin A and Provitamin A Overview of Natural Sources. Food and Nutrition Bulletin* vol. 14 (1992).
190. Kieliszek, M. Selenium—fascinating microelement, properties and sources in food. *Molecules* vol. 24 Preprint at <https://doi.org/10.3390/molecules24071298> (2019).
191. Fessler, M. B. Regulation of Adaptive Immunity in Health and Disease by Cholesterol Metabolism. *Current Allergy and Asthma Reports* vol. 15 1–16 Preprint at <https://doi.org/10.1007/s11882-015-0548-7> (2015).
192. *Dietary Sources of Zinc and Factors Affecting Its Bioavailability*. (2001).
193. Itkonen, S. T., Erkkola, M. & Lamberg-Allardt, C. J. E. Vitamin D fortification of fluid milk products and their contribution to vitamin D intake and vitamin D status in observational studies—a review. *Nutrients* vol. 10 Preprint at <https://doi.org/10.3390/nu10081054> (2018).
194. Lund, E. K. Health benefits of seafood; Is it just the fatty acids? in *Food Chemistry* vol. 140 413–420 (2013).

195. Ros, E. Health benefits of nut consumption. *Nutrients* vol. 2 652–682 Preprint at <https://doi.org/10.3390/nu2070652> (2010).
196. El Gharras, H. Polyphenols: Food sources, properties and applications - A review. *Int J Food Sci Technol* **44**, 2512–2518 (2009).
197. Sun, J., Chu, Y. F., Wu, X. & Liu, R. H. Antioxidant and antiproliferative activities of common fruits. *J Agric Food Chem* **50**, 7449–7454 (2002).
198. Polonikov, A. Endogenous Deficiency of Glutathione as the Most Likely Cause of Serious Manifestations and Death in COVID-19 Patients. *ACS Infectious Diseases* vol. 6 1558–1562 Preprint at <https://doi.org/10.1021/acsinfecdis.0c00288> (2020).
199. Jayawardena, R., Sooriyaarachchi, P., Chourdakis, M., Jeewandara, C. & Ranasinghe, P. Enhancing immunity in viral infections, with special emphasis on COVID-19: A review. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* **14**, 367–382 (2020).
200. Orsavova, J., Misurcova, L., Vavra Ambrozova, J., Vicha, R. & Mlcek, J. Fatty acids composition of vegetable oils and its contribution to dietary energy intake and dependence of cardiovascular mortality on dietary intake of fatty acids. *Int J Mol Sci* **16**, 12871–12890 (2015).
201. Sarkar, D., Katherine Jung, M., Joe Wang, H. & Sarkar, D. K. *Alcohol and the Immune System*.
202. Iddir, M. *et al.* Strengthening the immune system and reducing inflammation and oxidative stress through diet and nutrition: Considerations during the covid-19 crisis. *Nutrients* vol. 12 Preprint at <https://doi.org/10.3390/nu12061562> (2020).
203. Wei, X. *et al.* Hypolipidemia is associated with the severity of COVID-19. *J Clin Lipidol* **14**, 297–304 (2020).
204. Ravnskov, U. High cholesterol may protect against infections and atherosclerosis. *QJM* **96**, 927–934 (2003).
205. Steinbrenner, H., Al-Quraishy, S., Dkhil, M. A., Wunderlich, F. & Sies, H. Dietary selenium in adjuvant therapy of viral and bacterial infections. *Advances in Nutrition* vol. 6 73–82 Preprint at <https://doi.org/10.3945/an.114.007575> (2015).
206. Han, J. E. *et al.* High dose Vitamin D administration in ventilated intensive care unit patients: A pilot double blind randomized controlled trial. *J Clin Transl Endocrinol* **4**, 59–65 (2016).

207. Grant, W. B. *et al.* Evidence that vitamin d supplementation could reduce risk of influenza and covid-19 infections and deaths. *Nutrients* vol. 12 Preprint at <https://doi.org/10.3390/nu12040988> (2020).
208. Jayawardena, R., Sooriyaarachchi, P., Chourdakis, M., Jeewandara, C. & Ranasinghe, P. Enhancing immunity in viral infections, with special emphasis on COVID-19: A review. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews* **14**, 367–382 (2020).
209. Carr, A. C. A new clinical trial to test high-dose vitamin C in patients with COVID-19. *Critical Care* vol. 24 Preprint at <https://doi.org/10.1186/s13054-020-02851-4> (2020).
210. Fowler, A. A. *et al.* Effect of Vitamin C Infusion on Organ Failure and Biomarkers of Inflammation and Vascular Injury in Patients with Sepsis and Severe Acute Respiratory Failure: The CITRIS-ALI Randomized Clinical Trial. in *JAMA - Journal of the American Medical Association* vol. 322 1261–1270 (American Medical Association, 2019).
211. Sorokin, A. V. *et al.* COVID-19—Associated dyslipidemia: Implications for mechanism of impaired resolution and novel therapeutic approaches. *FASEB Journal* **34**, 9843–9853 (2020).
212. Zhu, X. *et al.* Alpha-linolenic acid protects against lipopolysaccharide-induced acute lung injury through anti-inflammatory and anti-oxidative pathways. *Microb Pathog* **142**, (2020).
213. Das, J. K., Salam, R. A., Saeed, M., Kazmi, F. A. & Bhutta, Z. A. Effectiveness of interventions for managing acute malnutrition in children under five years of age in low-income and middle-income countries: A systematic review and meta-analysis. *Nutrients* vol. 12 Preprint at <https://doi.org/10.3390/nu12010116> (2020).
214. Jaacks, L. M. & Bellows, A. L. Let food be thy medicine: Linking local food and health systems to address the full spectrum of malnutrition in low-income and middle-income countries. *BMJ Global Health* vol. 2 Preprint at <https://doi.org/10.1136/bmjgh-2017-000564> (2017).
215. Rogers, J. P. *et al.* Psychiatric and neuropsychiatric presentations associated with severe coronavirus infections: a systematic review and meta-analysis with comparison to the COVID-19 pandemic. *Lancet Psychiatry* **7**, 611–627 (2020).

216. Rooney, S., Webster, A. & Paul, L. *Systematic Review of Changes and Recovery in Physical Function and Fitness After Severe Acute Respiratory Syndrome-Related Coronavirus Infection: Implications for COVID-19 Rehabilitation*. *Phys Ther* vol. 100 <https://academic.oup.com/ptj> (2020).
217. Cederholm, T. *et al.* GLIM criteria for the diagnosis of malnutrition – A consensus report from the global clinical nutrition community. *J Cachexia Sarcopenia Muscle* **10**, 207–217 (2019).
218. Gracia-Iguacel, C., González-Parra, E., Mahillo, I. & Ortiz, A. Criteria for classification of protein-energy wasting in dialysis patients: Impact on prevalence. *British Journal of Nutrition* **121**, 1271–1278 (2019).
219. Filchakova, O. *et al.* Review of COVID-19 testing and diagnostic methods. *Talanta* vol. 244 Preprint at <https://doi.org/10.1016/j.talanta.2022.123409> (2022).
220. Maia, R. *et al.* Diagnosis Methods for COVID-19: A Systematic Review. *Micromachines* vol. 13 Preprint at <https://doi.org/10.3390/mi13081349> (2022).
221. Peeling, R. W., Heymann, D. L., Teo, Y. Y. & Garcia, P. J. Diagnostics for COVID-19: moving from pandemic response to control. *The Lancet* vol. 399 757–768 Preprint at [https://doi.org/10.1016/S0140-6736\(21\)02346-1](https://doi.org/10.1016/S0140-6736(21)02346-1) (2022).
222. De Bruyne, S., Speeckaert, M. M. & Delanghe, J. R. Applications of mid-infrared spectroscopy in the clinical laboratory setting. *Critical Reviews in Clinical Laboratory Sciences* vol. 55 1–20 Preprint at <https://doi.org/10.1080/10408363.2017.1414142> (2018).
223. Baker, M. J. *et al.* Clinical applications of infrared and Raman spectroscopy: State of play and future challenges. *Analyst* vol. 143 1735–1757 Preprint at <https://doi.org/10.1039/c7an01871a> (2018).
224. McInnes, M. D. F. *et al.* Preferred Reporting Items for a Systematic Review and Meta-analysis of Diagnostic Test Accuracy Studies The PRISMA-DTA Statement. *JAMA - Journal of the American Medical Association* **319**, 388–396 (2018).
225. Whiting, P. F. *et al.* QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies. www.annals.org (2011).
226. Bandeira, C. C. S. *et al.* Micro-Fourier-transform infrared reflectance spectroscopy as tool for probing IgG glycosylation in COVID-19 patients. *Sci Rep* **12**, (2022).

227. Barauna, V. G. *et al.* Ultrarapid On-Site Detection of SARS-CoV-2 Infection Using Simple ATR-FTIR Spectroscopy and an Analysis Algorithm: High Sensitivity and Specificity. *Anal Chem* **93**, 2950–2958 (2021).
228. Nogueira, M. S. *et al.* Rapid diagnosis of COVID-19 using FT-IR ATR spectroscopy and machine learning. *Sci Rep* **11**, (2021).
229. Nascimento, M. H. C. *et al.* Noninvasive Diagnostic for COVID-19 from Saliva Biofluid via FTIR Spectroscopy and Multivariate Analysis. *Anal Chem* **94**, 2425–2433 (2022).
230. Wood, B. R. *et al.* Infrared Based Saliva Screening Test for COVID-19. *Angewandte Chemie - International Edition* **60**, 17102–17107 (2021).
231. Heino, H. *et al.* Diagnostic performance of attenuated total reflection Fourier-transform infrared spectroscopy for detecting COVID-19 from routine nasopharyngeal swab samples. *Sci Rep* **12**, (2022).
232. Kazmer, S. T. *et al.* Pathophysiological Response to SARS-CoV-2 Infection Detected by Infrared Spectroscopy Enables Rapid and Robust Saliva Screening for COVID-19. *Biomedicines* **10**, (2022).
233. Calvo-Gomez, O. *et al.* Potential of ATR-FTIR-Chemometrics in Covid-19: Disease Recognition. *ACS Omega* **7**, 30756–30767 (2022).
234. Martinez-Cuazitl, A. *et al.* ATR-FTIR spectrum analysis of saliva samples from COVID-19 positive patients. *Sci Rep* **11**, (2021).
235. Karas, B. Y. *et al.* ATR-FTIR spectrum analysis of plasma samples for rapid identification of recovered COVID-19 individuals. *J Biophotonics* **16**, (2023).
236. Guleken, Z. *et al.* Characterization of Covid-19 infected pregnant women sera using laboratory indexes, vibrational spectroscopy, and machine learning classifications. *Talanta* **237**, (2022).
237. Kitane, D. L. *et al.* A simple and fast spectroscopy-based technique for Covid-19 diagnosis. *Sci Rep* **11**, (2021).
238. Laird, S. *et al.* Breath Analysis of COVID-19 Patients in a Tertiary UK Hospital by Optical Spectrometry: The E-Nose CoVal Study. *Biosensors (Basel)* **13**, (2023).
239. Shlomo, I. Ben, Frankenthal, H., Laor, A. & Kobo Greenhut, A. Detection of SARS-CoV-2 infection by exhaled breath spectral analysis: Introducing a ready-to-use point-of-care mass screening method. doi:10.1016/j.

240. Zhang, L. *et al.* Fast Screening and Primary Diagnosis of COVID-19 by ATR-FT-IR. *Anal Chem* **93**, 2191–2199 (2021).
241. Zhao, B., Zhai, H., Shao, H., Bi, K. & Zhu, L. Potential of vibrational spectroscopy coupled with machine learning as a non-invasive diagnostic method for COVID-19. *Comput Methods Programs Biomed* **229**, (2023).
242. Karthikeyan, S. *et al.* Dynamic response antibodies SARS-CoV-2 human saliva studied using two-dimensional correlation (2DCOS) infrared spectral analysis coupled with receiver operation characteristics analysis. *Biochim Biophys Acta Mol Basis Dis* **1869**, (2023).
243. Furman, G. *et al.* Remote Analysis of Respiratory Sounds in Patients with COVID-19: Development of Fast Fourier Transform–Based Computer-Assisted Diagnostic Methods. *JMIR Form Res* **6**, (2022).
244. Yang, X. *et al.* Diagnosis of Lung Cancer by ATR-FTIR Spectroscopy and Chemometrics. *Front Oncol* **11**, (2021).
245. Fan, M. *et al.* Near-infrared spectroscopy and chemometric modelling for rapid diagnosis of kidney disease. *Sci China Chem* **60**, 299–304 (2017).
246. Gromski, P. S. *et al.* A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta* vol. 879 10–23 Preprint at <https://doi.org/10.1016/j.aca.2015.02.012> (2015).
247. Mendez, K. M., Reinke, S. N. & Broadhurst, D. I. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* **15**, (2019).
248. Mendez, K. M., Broadhurst, D. I. & Reinke, S. N. The application of artificial neural networks in metabolomics: a historical perspective. *Metabolomics* vol. 15 Preprint at <https://doi.org/10.1007/s11306-019-1608-0> (2019).
249. Mendez, K. M., Reinke, S. N. & Broadhurst, D. I. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* **15**, 150 (2019).
250. Dunn, W. B., Broadhurst, D. I., Atherton, H. J., Goodacre, R. & Griffin, J. L. Systems level studies of mammalian metabolomes: the roles of mass

- spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* **40**, 387–426 (2011).
251. Réda, C., Kaufmann, E. & Delahaye-Duriez, A. Machine learning applications in drug development. *Comput Struct Biotechnol J* **18**, 241–252 (2020).
252. Alwosheel, A., van Cranenburgh, S. & Chorus, C. G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling* **28**, 167–182 (2018).
253. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
254. Qin, S. J. & Chiang, L. H. Advances and opportunities in machine learning for process data analytics. *Comput Chem Eng* **126**, 465–473 (2019).
255. Böger, B. *et al.* Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. *Am J Infect Control* **49**, 21–29 (2021).
256. Bag Soytas, R. *et al.* Antibody responses to COVID-19 vaccines in older adults. *J Med Virol* **94**, 1650–1654 (2022).
257. Moeller, M. E. *et al.* Rapid Quantitative Point-Of-Care Diagnostic Test for Post COVID-19 Vaccination Antibody Monitoring. *Microbiol Spectr* **10**, (2022).
258. Bhuiyan, M. U. *et al.* Epidemiology of COVID-19 infection in young children under five years: A systematic review and meta-analysis. *Vaccine* vol. 39 667–677 Preprint at <https://doi.org/10.1016/j.vaccine.2020.11.078> (2021).
259. Narain*, J. P., Raviglione, M. C. & Kochi, A. *Tuberculosis Programme and FGlobal Programme on AIDS, World Health Organization.*
260. Zhai, P. *et al.* The epidemiology, diagnosis and treatment of COVID-19. *Int J Antimicrob Agents* **55**, (2020).
261. Wang, L. *et al.* Artificial Intelligence for COVID-19: A Systematic Review. *Frontiers in Medicine* vol. 8 Preprint at <https://doi.org/10.3389/fmed.2021.704256> (2021).
262. Marcus, J. L., Sewell, W. C., Balzer, L. B. & Krakower, D. S. Artificial Intelligence and Machine Learning for HIV Prevention: Emerging Approaches to Ending the Epidemic. *Current HIV/AIDS Reports* vol. 17 171–179 Preprint at <https://doi.org/10.1007/s11904-020-00490-6> (2020).
263. Kulkarni, S. & Jha, S. Artificial Intelligence, Radiology, and Tuberculosis: A Review. *Academic Radiology* vol. 27 71–75 Preprint at <https://doi.org/10.1016/j.acra.2019.10.003> (2020).

264. Blagojević, A. *et al.* Artificial intelligence approach towards assessment of condition of COVID-19 patients - Identification of predictive biomarkers associated with severity of clinical condition and disease progression. *Comput Biol Med* **138**, (2021).
265. Kermali, M., Khalsa, R. K., Pillai, K., Ismail, Z. & Harky, A. The role of biomarkers in diagnosis of COVID-19 – A systematic review. *Life Sciences* vol. 254 Preprint at <https://doi.org/10.1016/j.lfs.2020.117788> (2020).
266. Hospital Israelita Albert Einstein. COVID-19 Open Research Dataset Challenge (CORD-19):HELP Diagnosis of COVID-19 and its clinical spectrum. <https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/139347> (2020).
267. Burdack, J. *et al.* Systematic Comparison of the Influence of Different Data Preprocessing Methods on the Performance of Gait Classifications Using Machine Learning. *Front Bioeng Biotechnol* **8**, (2020).
268. Astorino, A., Gorgone, E., Gaudioso, M. & Pallaschke, D. Data preprocessing in semi-supervised SVM classification. *Optimization* **60**, 143–151 (2011).
269. Ballabio, D. & Consonni, V. Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods* vol. 5 3790–3798 Preprint at <https://doi.org/10.1039/c3ay40582f> (2013).
270. Walczak, B. & Massart, D. L. *Multiple Outlier Detection Revisited*. *Chemometrics and Intelligent Laboratory Systems* vol. 41.
271. Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K. & Narasimhan, G. So you think you can PLS-DA? *BMC Bioinformatics* **21**, (2020).
272. Matthews, B. W. *BBA 37170 COMPARISON OF THE PREDICTED AND OBSERVED SECONDARY STRUCTURE OF T4 PHAGE LYSOZYME*. *Biochimica et Biophysica Acta* (1975).
273. Walczak, B. & Massart, D. L. *Multiple Outlier Detection Revisited*. *Chemometrics and Intelligent Laboratory Systems* vol. 41.
274. Mello, L. E. ; S. A. ; M. C. B. ; P. C. A. ; R. E. G. ; N. F. L. S. ; B. G. F. ; F. J. E. ; S. J. ; R. L. F. L. ; R. L. V. ; S. L. ; de L. R. ; M. R. M. de B. ; C.-J. R. M. ; C. R. *Opening Brazilian COVID-19 Patient Data to Support World Research on Pandemics*. <https://orcid.org/0000-0002-6969-1108>.
275. World Health Organization. *World Health Organization. Mozambique – A Comprehensive Community-Based Service Delivery Intervention for TB*.

- <https://www.who.int/publications/m/item/mozambique-a-comprehensive-community-based-service-delivery-intervention-for-tb> (2023).
276. Ministério da Saúde, I. N. de E. *Inquérito de Indicadores de Imunização, Malária e HIV/SIDA Em Moçambique (IMASIDA):2015-2018*. (2023).
277. Pougá, L. *et al.* New resistance mutations to nucleoside reverse transcriptase inhibitors at codon 184 of HIV-1 reverse transcriptase (M184L and M184T). *Chem Biol Drug Des* **93**, 50–59 (2019).
278. Van Stralen, K. J. *et al.* Diagnostic methods I: Sensitivity, specificity, and other measures of accuracy. *Kidney Int* **75**, 1257–1263 (2009).
279. Burdack, J. *et al.* Systematic Comparison of the Influence of Different Data Preprocessing Methods on the Performance of Gait Classifications Using Machine Learning. *Front Bioeng Biotechnol* **8**, (2020).
280. Astorino, A., Gorgone, E., Gaudio, M. & Pallaschke, D. Data preprocessing in semi-supervised SVM classification. *Optimization* **60**, 143–151 (2011).
281. Fu, J. *et al.* Pharmacometabonomics: Data processing and statistical analysis. *Brief Bioinform* **22**, (2021).
282. Tang, J., Mou, M., Wang, Y., Luo, Y. & Zhu, F. MetaFS: Performance assessment of biomarker discovery in metaproteomics. *Brief Bioinform* **22**, (2021).
283. Ballabio, D. & Consonni, V. Classification tools in chemistry. Part 1: Linear models. PLS-DA. *Analytical Methods* vol. 5 3790–3798 Preprint at <https://doi.org/10.1039/c3ay40582f> (2013).
284. Baumann, K. Cross-validation as the objective function for variable-selection techniques. *TrAC - Trends in Analytical Chemistry* vol. 22 395–406 Preprint at [https://doi.org/10.1016/S0165-9936\(03\)00607-1](https://doi.org/10.1016/S0165-9936(03)00607-1) (2003).
285. Voloch, C. M. *et al.* Genomic Characterization of a Novel SARS-CoV-2 Lineage from Rio de Janeiro, Brazil. *J Virol* **95**, (2021).
286. Nonaka, C. K. V. *et al.* Genomic evidence of SARS-CoV-2 reinfection involving E484K spike mutation, Brazil. *Emerg Infect Dis* **27**, 1522–1524 (2021).
287. Arthur, D. & Vassilvitskii, S. *K-Means++: The Advantages of Careful Seeding*.
288. Qin, S. J. & Chiang, L. H. Advances and opportunities in machine learning for process data analytics. *Comput Chem Eng* **126**, 465–473 (2019).

289. Réda, C., Kaufmann, E. & Delahaye-Duriez, A. Machine learning applications in drug development. *Computational and Structural Biotechnology Journal* vol. 18 241–252 Preprint at <https://doi.org/10.1016/j.csbj.2019.12.006> (2020).
290. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* vol. 559 547–555 Preprint at <https://doi.org/10.1038/s41586-018-0337-2> (2018).
291. Shallue, C. J. *et al.* *Measuring the Effects of Data Parallelism on Neural Network Training.* *Journal of Machine Learning Research* vol. 20 <https://www.blog.google/products/google-cloud/google-cloud-offer-tpus-machine-learning/>. (2019).
292. Alwosheel, A., van Cranenburgh, S. & Chorus, C. G. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling* **28**, 167–182 (2018).
293. Liu, B., Wei, Y., Zhang, Y., Yang, Q. & Kong, H. *Deep Neural Networks for High Dimension, Low Sample Size Data.* (2017).
294. Dai, W. *et al.* Establishing classifiers with clinical laboratory indicators to distinguish COVID-19 from community-acquired pneumonia: Retrospective cohort study. *J Med Internet Res* **23**, (2021).
295. Wu, P. *et al.* An effective machine learning approach for identifying non-severe and severe coronavirus disease 2019 patients in a rural Chinese population: The wenzhou retrospective study. *IEEE Access* **9**, 45486–45503 (2021).
296. Laing, A. G. *et al.* A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nat Med* **26**, 1623–1635 (2020).
297. Banerjee, A. *et al.* Use of Machine Learning and Artificial Intelligence to predict SARS-CoV-2 infection from Full Blood Counts in a population. *Int Immunopharmacol* **86**, (2020).
298. Joshi, R. P. *et al.* A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results. *Journal of Clinical Virology* **129**, (2020).
299. Formica, V. *et al.* Complete blood count might help to identify subjects with high probability of testing positive to SARS-CoV-2. *Clinical Medicine, Journal of the Royal College of Physicians of London* **20**, (2020).

300. Cabitza, F. *et al.* Development, evaluation, and validation of machine learning models for COVID-19 detection based on routine blood tests. *Clin Chem Lab Med* **59**, 421–431 (2021).
301. Zhou, X. *et al.* Machine learning-based decision model to distinguish between covid-19 and influenza: A retrospective, two-centered, diagnostic study. *Risk Manag Healthc Policy* **14**, 595–604 (2021).
302. Wu, P. *et al.* An effective machine learning approach for identifying non-severe and severe coronavirus disease 2019 patients in a rural Chinese population: The wenzhou retrospective study. *IEEE Access* **9**, 45486–45503 (2021).
303. Danzi, G. B., Loffi, M., Galeazzi, G. & Gherbesi, E. Acute pulmonary embolism and COVID-19 pneumonia: A random association? *European Heart Journal* vol. 41 1858 Preprint at <https://doi.org/10.1093/eurheartj/ehaa254> (2020).
304. Taneri, P. E. *et al.* Anemia and iron metabolism in COVID-19: a systematic review and meta-analysis. *Eur J Epidemiol* **35**, 763–773 (2020).
305. Fox, S. E. *et al.* Pulmonary and cardiac pathology in African American patients with COVID-19: an autopsy series from New Orleans. *Lancet Respir Med* **8**, 681–686 (2020).
306. Liu, Z., Sun, R., Li, J., Cheng, W. & Li, L. Relations of Anemia With the All-Cause Mortality and Cardiovascular Mortality in General Population: A Meta-Analysis. *American Journal of the Medical Sciences* **358**, 191–199 (2019).
307. Bennett, T. D. *et al.* Very high serum ferritin levels are associated with increased mortality and critical care in pediatric patients. *Pediatric Critical Care Medicine* **12**, (2011).
308. Rasmussen, L., Christensen, S., Lenler-Petersen, P. & Johnsen, S. P. Anemia and 90-day mortality in COPD patients requiring invasive mechanical ventilation. *Clin Epidemiol* **3**, 1–5 (2011).
309. Lomholt, F. K., Laulund, A. S., Bjarnason, N. H., Jørgensen, H. L. & Godtfredsen, N. S. Meta-analysis of routine blood tests as predictors of mortality in COPD. *Eur Clin Respir J* **1**, 24110 (2014).
310. Abbaspour, N., Hurrell, R. & Kelishadi, R. *Review on Iron and Its Importance for Human Health. Journal of Research in Medical Sciences* (2014).
311. Vargas-Vargas, M. & Cortés-Rojo, C. Ferritin levels and COVID-19. *Revista Panamericana de Salud Publica/Pan American Journal of Public Health* vol. 44 Preprint at <https://doi.org/10.26633/RPSP.2020.72> (2020).

312. Mehta, P. *et al.* COVID-19: consider cytokine storm syndromes and immunosuppression. *The Lancet* vol. 395 1033–1034 Preprint at [https://doi.org/10.1016/S0140-6736\(20\)30628-0](https://doi.org/10.1016/S0140-6736(20)30628-0) (2020).
313. Alper, S. L. Genetic diseases of acid-base transporters. *Annual Review of Physiology* vol. 64 899–923 Preprint at <https://doi.org/10.1146/annurev.physiol.64.092801.141759> (2002).
314. Bruno, C. M. & Valenti, M. Acid-base disorders in patients with chronic obstructive pulmonary disease: A pathophysiological review. *Journal of Biomedicine and Biotechnology* vol. 2012 Preprint at <https://doi.org/10.1155/2012/915150> (2012).
315. Goraya, N. & Wesson, D. E. Acid-base status and progression of chronic kidney disease. *Current Opinion in Nephrology and Hypertension* vol. 21 552–556 Preprint at <https://doi.org/10.1097/MNH.0b013e328356233b> (2012).
316. Vincent, J. *INTENSIVE CARE MEDICINE 1001 ANNUAL UPDATE*.
317. Lavenus, S., Rozé, J., Louarn, G. & Layrolle, P. Impact of Nanotechnology on Dental Implants. in *Nanobiomaterials in Clinical Dentistry* 323–336 (Elsevier Inc., 2012). doi:10.1016/B978-1-4557-3127-5.00016-7.
318. Gunnerson, K. J. Clinical review: The meaning of acid-base abnormalities in the intensive care unit - Epidemiology. *Critical Care* vol. 9 508–516 Preprint at <https://doi.org/10.1186/cc3796> (2005).
319. De Backer, D. Lactic acidosis. *Intensive Care Med* **29**, 699–702 (2003).
320. Li, J. *et al.* COVID-19 infection may cause ketosis and ketoacidosis. *Diabetes Obes Metab* **22**, 1935–1941 (2020).
321. Kamel, K. S., Oh, M. S. & Halperin, M. L. L-lactic acidosis: pathophysiology, classification, and causes; emphasis on biochemical and metabolic basis. *Kidney International* vol. 97 75–88 Preprint at <https://doi.org/10.1016/j.kint.2019.08.023> (2020).
322. Kogelmann, K., Jarczак, D., Scheller, M. & Drüner, M. Hemoadsorption by CytoSorb in septic patients: A case series. *Crit Care* **21**, (2017).
323. Li, J. *et al.* COVID-19 infection may cause ketosis and ketoacidosis. *Diabetes Obes Metab* **22**, 1935–1941 (2020).
324. Venkatalaxmi, A., Padmavathi, B. S. & Amaranath, T. A general solution of unsteady Stokes equations. *Dynamics Research* **35**, 229–236 (2004).

325. Goldman, N., Fink, D., Cai, J., Lee, Y. N. & Davies, Z. High prevalence of COVID-19-associated diabetic ketoacidosis in UK secondary care. *Diabetes Res Clin Pract* **166**, (2020).
326. Chan, K. H. *et al.* Clinical characteristics and outcome in patients with combined diabetic ketoacidosis and hyperosmolar hyperglycemic state associated with COVID-19: A retrospective, hospital-based observational case series. *Diabetes Res Clin Pract* **166**, (2020).
327. Cheng, X. *et al.* Metformin Is Associated with Higher Incidence of Acidosis, but Not Mortality, in Individuals with COVID-19 and Pre-existing Type 2 Diabetes. *Cell Metab* **32**, 537-547.e3 (2020).
328. Parohan, M., Yaghoubi, S. & Seraji, A. Liver injury is associated with severe coronavirus disease 2019 (COVID-19) infection: A systematic review and meta-analysis of retrospective studies. *Hepatology Research* vol. 50 924–935 Preprint at <https://doi.org/10.1111/hepr.13510> (2020).
329. Fox, S. E. *et al.* Pulmonary and cardiac pathology in African American patients with COVID-19: an autopsy series from New Orleans. *Lancet Respir Med* **8**, 681–686 (2020).
330. Thiele, J. R. *et al.* Dissociation of pentameric to monomeric C-reactive protein localizes and aggravates inflammation: In vivo proof of a powerful proinflammatory mechanism and a new anti-inflammatory strategy. *Circulation* **130**, 35–50 (2014).
331. Huang, I., Pranata, R., Lim, M. A., Oehadian, A. & Alisjahbana, B. C-reactive protein, procalcitonin, D-dimer, and ferritin in severe coronavirus disease-2019: a meta-analysis. *Ther Adv Respir Dis* **14**, (2020).
332. Chung, J. W., Ryu, W. S., Kim, B. J. & Yoon, B. W. Elevated calcium after acute ischemic stroke: Association with a poor short-term outcome and long-term mortality. *J Stroke* **17**, 54–59 (2015).
333. Appel, S. A. *et al.* Serum calcium levels and long-term mortality in patients with acute stroke. *Cerebrovascular Diseases* **31**, 93–99 (2010).
334. Yan, S. Di *et al.* Admission Serum Calcium Levels Improve the GRACE Risk Score Prediction of Hospital Mortality in Patients With Acute Coronary Syndrome. *Clin Cardiol* **39**, 516–523 (2016).
335. Holowaychuk, M. K. & Martin, L. G. Review of hypocalcemia in septic patients: State-of-the-Art Review. *Journal of Veterinary Emergency and Critical Care* vol.

- 17 348–358 Preprint at <https://doi.org/10.1111/j.1476-4431.2007.00246.x> (2007).
336. Sankaran, R. T. *et al.* Laboratory abnormalities in patients with bacterial pneumonia. *Chest* **111**, 595–600 (1997).
337. Venkatalaxmi, A., Padmavathi, B. S. & Amaranath, T. A general solution of unsteady Stokes equations. *Dynamics Research* **35**, 229–236 (2004).
338. di Filippo, L. *et al.* Hypocalcemia is a distinctive biochemical feature of hospitalized COVID-19 patients. *Endocrine* vol. 71 9–13 Preprint at <https://doi.org/10.1007/s12020-020-02541-9> (2021).
339. Liu, J., Han, P., Wu, J., Gong, J. & Tian, D. Prevalence and predictive value of hypocalcemia in severe COVID-19 patients. *J Infect Public Health* **13**, 1224–1228 (2020).
340. Nonaka, C. K. V. *et al.* Genomic evidence of SARS-CoV-2 reinfection involving E484K spike mutation, Brazil. *Emerg Infect Dis* **27**, 1522–1524 (2021).
341. Wang, R., Hozumi, Y., Yin, C. & Wei, G. W. Mutations on COVID-19 diagnostic targets. *Genomics* **112**, 5204–5213 (2020).
342. Pouga, L. *et al.* New resistance mutations to nucleoside reverse transcriptase inhibitors at codon 184 of HIV-1 reverse transcriptase (M184L and M184T). *Chem Biol Drug Des* **93**, 50–59 (2019).
343. Peiffer-Smadja, N. *et al.* Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. *Clinical Microbiology and Infection* vol. 26 584–595 Preprint at <https://doi.org/10.1016/j.cmi.2019.09.009> (2020).
344. Alves, M. A. *et al.* Explaining machine learning based diagnosis of COVID-19 from routine blood tests with decision trees and criteria graphs. *Comput Biol Med* **132**, (2021).
345. Zuin, G. *et al.* Prediction of SARS-CoV-2-positivity from million-scale complete blood counts using machine learning. *Communications Medicine* **2**, (2022).
346. Torres, P. B. *et al.* Standardization of a protocol to extract and analyze chlorophyll a and carotenoids in *Gracilaria tenuistipitata* var. *liui*. zhang and xia (rhodophyta). *Braz J Oceanogr* **62**, 57–63 (2014).
347. Vekey´chemical, K. *Mass Spectrometry and Mass-Selective Detection in q Chromatography*. *Journal of Chromatography A* www.elsevier.com/locate/chroma (2001).

348. Albahri, A. S. *et al.* Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review. *Journal of Medical Systems* vol. 44 Preprint at <https://doi.org/10.1007/s10916-020-01582-x> (2020).
349. Caballé, N. C., Castillo-Sequera, J. L., Gómez-Pulido, J. A., Gómez-Pulido, J. M. & Polo-Luque, M. L. Machine learning applied to diagnosis of human diseases: A systematic review. *Applied Sciences (Switzerland)* vol. 10 Preprint at <https://doi.org/10.3390/app10155135> (2020).
350. Marcus, J. L., Sewell, W. C., Balzer, L. B. & Krakower, D. S. Artificial Intelligence and Machine Learning for HIV Prevention: Emerging Approaches to Ending the Epidemic. *Current HIV/AIDS Reports* vol. 17 171–179 Preprint at <https://doi.org/10.1007/s11904-020-00490-6> (2020).
351. Nieto-Torres, J. L. *et al.* Severe acute respiratory syndrome coronavirus E protein transports calcium ions and activates the NLRP3 inflammasome. *Virology* **485**, 330–339 (2015).
352. Deng, B. *et al.* Cytokine and chemokine levels in patients with severe fever with thrombocytopenia syndrome virus. *PLoS One* **7**, (2012).
353. Alemzadeh, E., Alemzadeh, E., Ziaee, M., Abedi, A. & Salehiniya, H. The effect of low serum calcium level on the severity and mortality of Covid patients: A systematic review and meta-analysis. *Immunity, Inflammation and Disease* vol. 9 1219–1228 Preprint at <https://doi.org/10.1002/iid3.528> (2021).
354. Yang, C. *et al.* Low serum calcium and phosphorus and their clinical performance in detecting COVID-19 patients. *J Med Virol* **93**, 1639–1651 (2021).
355. Cappellini, F. *et al.* Low levels of total and ionized calcium in blood of COVID-19 patients. *Clinical Chemistry and Laboratory Medicine* vol. 58 E171–E173 Preprint at <https://doi.org/10.1515/cclm-2020-0611> (2020).
356. Nieto-Torres, J. L. *et al.* Severe Acute Respiratory Syndrome Coronavirus Envelope Protein Ion Channel Activity Promotes Virus Fitness and Pathogenesis. *PLoS Pathog* **10**, (2014).
357. Rabaan, A. A. *et al.* SARS-CoV-2, SARS-CoV, and MERS-CoV: A Comparative Overview.
358. Rabaan, A. A. *et al.* SARS-CoV-2, SARS-CoV, and MERS-CoV: A Comparative Overview.

359. Martha, J. W., Wibowo, A. & Pranata, R. Prognostic value of elevated lactate dehydrogenase in patients with COVID-19: a systematic review and meta-analysis. *Postgrad Med J* **98**, 422–427 (2022).
360. Hariyanto, T. I. *et al.* Inflammatory and hematologic markers as predictors of severe outcomes in COVID-19 infection: A systematic review and meta-analysis. *American Journal of Emergency Medicine* **41**, 110–119 (2021).
361. Abay, F., Yalew, A., Shibabaw, A. & Enawgaw, B. Hematological Abnormalities of Pulmonary Tuberculosis Patients with and without HIV at the University of Gondar Hospital, Northwest Ethiopia: A Comparative Cross-Sectional Study. *Tuberc Res Treat* **2018**, 1–6 (2018).
362. López-Pereira, P. *et al.* Can COVID-19 cause severe neutropenia? *Clin Case Rep* **8**, 3349–3351 (2020).
363. Hernandez, J. M. *et al.* Pancytopenia and Profound Neutropenia as a Sequela of Severe SARS-CoV-2 Infection (COVID-19) with Concern for Bone Marrow Involvement. *Open Forum Infect Dis* **8**, (2021).
364. Enoh, J. E., Cho, F. N., Manfo, F. P., Ako, S. E. & Akum, E. A. Abnormal Levels of Liver Enzymes and Hepatotoxicity in HIV-Positive, TB, and HIV/TB-Coinfected Patients on Treatment in Fako Division, Southwest Region of Cameroon. *Biomed Res Int* **2020**, (2020).
365. Mocroft, A. *et al.* *Reasons for Stopping Antiretrovirals Used in an Initial Highly Active Antiretroviral Regimen: Increased Incidence of Stopping Due to Toxicity or Patient/Physician Choice in Patients with Hepatitis C Coinfection.* *AIDS RESEARCH AND HUMAN RETROVIRUSES* vol. 21 http://www.cphiv.dk/pdf_folder/EuroSIDA_enrol_2001.pdf (2005).
366. Hensley-McBain, T. & Klatt, N. R. The Dual Role of Neutrophils in HIV Infection. *Current HIV/AIDS Reports* vol. 15 Preprint at <https://doi.org/10.1007/s11904-018-0370-7> (2018).
367. Banda, N. K. *et al.* *HIV-Gp120 Induced Cell Death in Hematopoietic Progenitor CD34 + Cells.* *Apoptosis* vol. 2 (1997).
368. Carter, C. C. *et al.* HIV-1 infects multipotent progenitor cells causing cell death and establishing latent cellular reservoirs. *Nat Med* **16**, 446–451 (2010).
369. Calenda, V., Graber, P., Delamarter, J. -F & Chermann, J. -C. Involvement of HIV nef protein in abnormal hematopoiesis in AIDS: in vitro study on bone marrow progenitor cells. *Eur J Haematol* **52**, 103–107 (1994).

370. Chu, K. A. *et al.* Association of iron deficiency anemia with tuberculosis in Taiwan: A nationwide population-based study. *PLoS One* **14**, (2019).
371. Bergamaschi, G. *et al.* Anemia in patients with Covid-19: pathogenesis and clinical significance. *Clin Exp Med* **21**, 239–246 (2021).
372. Volberding, P. A. *et al.* *Anemia in HIV Infection: Clinical Impact and Evidence-Based Management Strategies*. *Clinical Infectious Diseases* vol. 38 <https://academic.oup.com/cid/article/38/10/1454/347195> (2004).
373. Kibaru, E. G., Nduati, R., Wamalwa, D. & Kariuki, N. Impact of highly active antiretroviral therapy on hematological indices among HIV-1 infected children at Kenyatta National Hospital-Kenya: Retrospective study. *AIDS Res Ther* **12**, (2015).
374. Mihiretie, H., Taye, B. & Tsegaye, A. Magnitude of anemia and associated factors among pediatric HIV/aids patients attending Zewditu memorial hospital art clinic, Addis Ababa, Ethiopia. *Anemia* **2015**, (2015).
375. Mitiku, H. & Mesfin, F. Prevalence of anemia and nutritional status among HIV-positive children receiving antiretroviral therapy in harar, eastern Ethiopia. *HIV/AIDS - Research and Palliative Care* **7**, 191–196 (2015).
376. Volberding, P. *The Impact of Anemia on Quality of Life in Human Immunodeficiency Virus-Infected Patients*. https://academic.oup.com/jid/article/185/Supplement_2/S110/887029.
377. Macallan, D. C. Malnutrition in Tuberculosis. (1999).
378. de Oliveira, M. A. L. *et al.* DIAGNOSTIC TESTS FOR SARS-COV-2: A CRITICAL REFLECTION. *Quim Nova* **45**, 760–766 (2022).
379. Menezes, D., Fonseca, P. L. C., de Araújo, J. L. F. & Souza, R. P. de. SARS-CoV-2 Genomic Surveillance in Brazil: A Systematic Review with Scientometric Analysis. *Viruses* vol. 14 Preprint at <https://doi.org/10.3390/v14122715> (2022).
380. Hirabara, S. M. *et al.* SARS-COV-2 Variants: Differences and Potential of Immune Evasion. *Frontiers in Cellular and Infection Microbiology* vol. 11 Preprint at <https://doi.org/10.3389/fcimb.2021.781429> (2022).
381. Bigoni, A. *et al.* Brazil's health system functionality amidst of the COVID-19 pandemic: An analysis of resilience. *The Lancet Regional Health - Americas* **10**, 100222 (2022).

382. de MELO, A. C., Trindade, G. M., de FREITAS, A. R., Resende, K. A. & Palhano, T. J. Community pharmacies and pharmacists in Brazil: A missed opportunity. *Pharm Pract (Granada)* **19**, (2021).
383. Costa, S. *et al.* Pharmacy interventions on COVID-19 in Europe: Mapping current practices and a scoping review. *Research in Social and Administrative Pharmacy* **18**, 3338–3349 (2022).
384. Carpenter, D. M. *et al.* Rural community pharmacies' preparedness for and responses to COVID-19. *Research in Social and Administrative Pharmacy* **17**, 1327–1331 (2021).
385. Uebbing, E., Lacroix, M., Bratberg, J. & Federico, C. Pharmacists' response during a pandemic: A survey on readiness to test during COVID-19. *Journal of the American Pharmacists Association* **61**, e80–e84 (2021).
386. Alwhaibi, A. *et al.* Role of pharmacist during COVID-19 pandemic: A retrospective study focused on critically ill COVID-19 patients. *Saudi Pharmaceutical Journal* **29**, 1050–1055 (2021).
387. Cobre, A. de F. *et al.* Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators? *Comput Biol Med* **134**, (2021).
388. Stiglic, G. *et al.* Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* vol. 10 Preprint at <https://doi.org/10.1002/widm.1379> (2020).
389. Joyce, D. W., Kormilitzin, A., Smith, K. A. & Cipriani, A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *npj Digital Medicine* vol. 6 Preprint at <https://doi.org/10.1038/s41746-023-00751-9> (2023).
390. Heinen, S., Schwilk, M., Von Rudorff, G. F. & Von Lilienfeld, O. A. Machine learning the computational cost of quantum chemistry. *Mach Learn Sci Technol* **1**, (2020).
391. Böger, B. *et al.* Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19. *Am J Infect Control* **49**, 21–29 (2021).
392. Wang, L. *et al.* Artificial Intelligence for COVID-19: A Systematic Review. *Frontiers in Medicine* vol. 8 Preprint at <https://doi.org/10.3389/fmed.2021.704256> (2021).

393. Müller, D., Soto-Rey, I. & Kramer, F. Robust chest CT image segmentation of COVID-19 lung infection based on limited data. *Inform Med Unlocked* **25**, (2021).
394. Sejuti, Z. A. & Islam, M. S. A hybrid CNN–KNN approach for identification of COVID-19 with 5-fold cross validation. *Sensors International* **4**, (2023).
395. Allen, W. E. *et al.* Population-scale longitudinal mapping of COVID-19 symptoms, behaviour and testing. *Nat Hum Behav* **4**, 972–982 (2020).
396. Alimohamadi, Y., Sepandi, M., Taghdir, M. & Hosamirudsari, H. Determine the most common clinical symptoms in COVID-19 patients: A systematic review and meta-analysis. *Journal of Preventive Medicine and Hygiene* vol. 61 E304–E312 Preprint at <https://doi.org/10.15167/2421-4248/jpmh2020.61.3.1530> (2020).
397. Fernández-de-las-Peñas, C. *et al.* Prevalence of post-COVID-19 symptoms in hospitalized and non-hospitalized COVID-19 survivors: A systematic review and meta-analysis. *Eur J Intern Med* **92**, 55–70 (2021).
398. Yang, J. *et al.* Prevalence of comorbidities and its effects in coronavirus disease 2019 patients: A systematic review and meta-analysis. *International Journal of Infectious Diseases* **94**, 91–95 (2020).
399. Cares-Marambio, K. *et al.* Prevalence of potential respiratory symptoms in survivors of hospital admission after coronavirus disease 2019 (COVID-19): A systematic review and meta-analysis. *Chronic Respiratory Disease* vol. 18 Preprint at <https://doi.org/10.1177/147997312111002240> (2021).
400. Pascarella, G. *et al.* COVID-19 diagnosis and management: a comprehensive review. *Journal of Internal Medicine* vol. 288 192–206 Preprint at <https://doi.org/10.1111/joim.13091> (2020).
401. Mesas, A. E. *et al.* Predictors of in-hospital COVID-19 mortality: A comprehensive systematic review and meta-analysis exploring differences by age, sex and health conditions. *PLoS One* **15**, (2020).
402. Cobre, A. de F. *et al.* Risk factors associated with delay in diagnosis and mortality in patients with covid-19 in the city of rio de janeiro, brazil. *Ciencia e Saude Coletiva* **25**, 4131–4140 (2020).
403. Jin, J. M. *et al.* Gender Differences in Patients With COVID-19: Focus on Severity and Mortality. *Front Public Health* **8**, (2020).
404. Kopel, J. *et al.* Racial and Gender-Based Differences in COVID-19. *Frontiers in Public Health* vol. 8 Preprint at <https://doi.org/10.3389/fpubh.2020.00418> (2020).

405. Hohl, H. T. *et al.* COVID-19 Testing Unit Munich: Impact of Public Health and Safety Measures on Patient Characteristics and Test Results, January to September 2020. *Front Public Health* **10**, (2022).
406. Bigoni, A. *et al.* Brazil's health system functionality amidst of the COVID-19 pandemic: An analysis of resilience. *The Lancet Regional Health - Americas* **10**, 100222 (2022).
407. Bigoni, A. *et al.* Brazil's health system functionality amidst of the COVID-19 pandemic: An analysis of resilience. *The Lancet Regional Health - Americas* **10**, 100222 (2022).
408. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *The BMJ* **369**, (2020).
409. Pang, Z., Zhou, G., Chong, J. & Xia, J. Comprehensive meta-analysis of covid-19 global metabolomics datasets. *Metabolites* **11**, 1–14 (2021).
410. Buyukozkan, M. *et al.* Integrative metabolomic and proteomic signatures define clinical outcomes in severe COVID-19. *iScience* **25**, (2022).
411. Hasan, M. R., Suleiman, M. & Pérez-López, A. Metabolomics in the Diagnosis and Prognosis of COVID-19. *Frontiers in Genetics* vol. 12 Preprint at <https://doi.org/10.3389/fgene.2021.721556> (2021).
412. Onoja, A. *et al.* Meta-Analysis of COVID-19 Metabolomics Identifies Variations in Robustness of Biomarkers. *Int J Mol Sci* **24**, (2023).
413. Pang, Z., Zhou, G., Chong, J. & Xia, J. Comprehensive meta-analysis of covid-19 global metabolomics datasets. *Metabolites* **11**, 1–14 (2021).
414. Galal, A., Talal, M. & Moustafa, A. Applications of machine learning in metabolomics: Disease modeling and classification. *Frontiers in Genetics* vol. 13 Preprint at <https://doi.org/10.3389/fgene.2022.1017340> (2022).
415. Pang, Z., Zhou, G., Chong, J. & Xia, J. Comprehensive meta-analysis of covid-19 global metabolomics datasets. *Metabolites* **11**, 1–14 (2021).
416. Astorino, A., Gorgone, E., Gaudio, M. & Pallaschke, D. Data preprocessing in semi-supervised SVM classification. *Optimization* **60**, 143–151 (2011).
417. Wienold, J. *et al.* Cross-validation and robustness of daylight glare metrics. *Lighting Research and Technology* **51**, 983–1013 (2019).
418. Favilla, S., Durante, C., Vigni, M. L. & Cocchi, M. Assessing feature relevance in NPLS models by VIP. *Chemometrics and Intelligent Laboratory Systems* **129**, 76–86 (2013).

419. Cocchi, M., Biancolillo, A. & Marini, F. Chemometric Methods for Classification and Feature Selection. in *Comprehensive Analytical Chemistry* vol. 82 265–299 (Elsevier B.V., 2018).
420. Pang, Z. *et al.* MetaboAnalyst 5.0: Narrowing the gap between raw spectra and functional insights. *Nucleic Acids Res* **49**, W388–W396 (2021).
421. Bruzzone, C. *et al.* SARS-CoV-2 Infection Dysregulates the Metabolomic and Lipidomic Profiles of Serum. *iScience* **23**, (2020).
422. Shi, D. *et al.* The serum metabolome of COVID-19 patients is distinctive and predictive. *Metabolism* **118**, (2021).
423. Barberis, E. *et al.* Large-scale plasma analysis revealed new mechanisms and molecules associated with the host response to sars-cov-2. *Int J Mol Sci* **21**, 1–25 (2020).
424. Blasco, H. *et al.* The specific metabolome profiling of patients infected by SARS-COV-2 supports the key role of tryptophan-nicotinamide pathway and cytosine metabolism. *Sci Rep* **10**, (2020).
425. Caterino, M. *et al.* The serum metabolome of moderate and severe covid-19 patients reflects possible liver alterations involving carbon and nitrogen metabolism. *Int J Mol Sci* **22**, (2021).
426. Barberis, E. *et al.* Understanding protection from SARS-CoV-2 using metabolomics. *Sci Rep* **11**, (2021).
427. Folch-Fortuny, A., Arteaga, F. & Ferrer, A. PCA model building with missing data: New proposals and a comparative study. *Chemometrics and Intelligent Laboratory Systems* **146**, 77–88 (2015).
428. Wang, R., Hozumi, Y., Yin, C. & Wei, G. W. Mutations on COVID-19 diagnostic targets. *Genomics* **112**, 5204–5213 (2020).
429. Voloch, C. M. *et al.* Genomic Characterization of a Novel SARS-CoV-2 Lineage from Rio de Janeiro, Brazil. *J Virol* **95**, (2021).
430. Nonaka, C. K. V. *et al.* Genomic evidence of SARS-CoV-2 reinfection involving E484K spike mutation, Brazil. *Emerg Infect Dis* **27**, 1522–1524 (2021).
431. Gromski, P. S. *et al.* A tutorial review: Metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. *Analytica Chimica Acta* vol. 879 10–23 Preprint at <https://doi.org/10.1016/j.aca.2015.02.012> (2015).

432. Mendez, K. M., Reinke, S. N. & Broadhurst, D. I. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics* **15**, (2019).
433. Mendez, K. M., Broadhurst, D. I. & Reinke, S. N. The application of artificial neural networks in metabolomics: a historical perspective. *Metabolomics* vol. 15 Preprint at <https://doi.org/10.1007/s11306-019-1608-0> (2019).
434. Dunn, W. B., Broadhurst, D. I., Atherton, H. J., Goodacre, R. & Griffin, J. L. Systems level studies of mammalian metabolomes: The roles of mass spectrometry and nuclear magnetic resonance spectroscopy. *Chem Soc Rev* **40**, 387–426 (2011).
435. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* vol. 559 547–555 Preprint at <https://doi.org/10.1038/s41586-018-0337-2> (2018).
436. Qin, S. J. & Chiang, L. H. Advances and opportunities in machine learning for process data analytics. *Comput Chem Eng* **126**, 465–473 (2019).
437. Fu, J. *et al.* Optimization of metabolomic data processing using NOREVA. *Nature Protocols* vol. 17 129–151 Preprint at <https://doi.org/10.1038/s41596-021-00636-9> (2022).
438. De Livera, A. M., Olshansky, G., Simpson, J. A. & Creek, D. J. NormalizeMets: assessing, selecting and implementing statistical methods for normalizing metabolomics data. *Metabolomics* **14**, (2018).
439. Tang, J. *et al.* ANPELA: Analysis and performance assessment of the label-free quantification workflow for metaproteomic studies. *Brief Bioinform* **21**, 621–636 (2020).
440. Tang, J. *et al.* Simultaneous Improvement in the Precision, Accuracy, and Robustness of Label-free Proteome Quantification by Optimizing Data Manipulation Chains. *Molecular and Cellular Proteomics* **18**, 1683–1699 (2019).
441. Yang, Q. *et al.* MMEASE: Online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis. *J Proteomics* **232**, (2021).
442. Bertol, G., Cobre, A. F. & Pontarolo, R. Differentiation of *Mikania glomerata* and *Mikania laevigata* Species Through Mid-infrared Spectroscopy and

- Chemometrics Guided by HPLC-DAD Analyses. *Revista Brasileira de Farmacognosia* **31**, 442–452 (2021).
443. Patterson, B. K. *et al.* Immune-Based Prediction of COVID-19 Severity and Chronicity Decoded Using Machine Learning. *Front Immunol* **12**, (2021).
444. Choudhary, S., Sreenivasulu, K., Mitra, P., Misra, S. & Sharma, P. Role of genetic variants and gene expression in the susceptibility and severity of COVID-19. *Annals of Laboratory Medicine* vol. 41 129–138 Preprint at <https://doi.org/10.3343/alm.2021.41.2.129> (2020).
445. Abu-Raddad, L. J., Chemaitelly, H. & Butt, A. A. Effectiveness of the BNT162b2 Covid-19 Vaccine against the B.1.1.7 and B.1.351 Variants. *New England Journal of Medicine* **385**, 187–189 (2021).
446. Patel, J. & Sampson, V. The role of oral bacteria in COVID-19. (2020) doi:10.31219/osf.io/jegwq.
447. Yamamoto, S. *et al.* The human microbiome and COVID-19: A systematic review. *PLoS ONE* vol. 16 Preprint at <https://doi.org/10.1371/journal.pone.0253293> (2021).
448. Visconti, A. *et al.* Interplay between the human gut microbiome and host metabolism. *Nat Commun* **10**, (2019).
449. Yang, D. *et al.* Implications of gut microbiota dysbiosis and metabolic changes in prion disease. *Neurobiol Dis* **135**, (2020).
450. Li, N., Ma, W. T., Pang, M., Fan, Q. L. & Hua, J. L. The commensal microbiota and viral infection: A comprehensive review. *Frontiers in Immunology* vol. 10 Preprint at <https://doi.org/10.3389/fimmu.2019.01551> (2019).
451. Garcia-Vidal, C. *et al.* Incidence of co-infections and superinfections in hospitalized patients with COVID-19: a retrospective cohort study. *Clinical Microbiology and Infection* **27**, 83–88 (2021).
452. Langford, B. J. *et al.* Bacterial co-infection and secondary infection in patients with COVID-19: a living rapid review and meta-analysis. *Clinical Microbiology and Infection* vol. 26 1622–1629 Preprint at <https://doi.org/10.1016/j.cmi.2020.07.016> (2020).
453. Timm, C. M., Lloyd, E. P., Egan, A., Mariner, R. & Karig, D. Direct growth of bacteria in headspace vials allows for screening of volatiles by gas chromatography mass spectrometry. *Front Microbiol* **9**, (2018).

454. Zou, J. N. *et al.* The characteristics and evolution of pulmonary fibrosis in COVID-19 patients as assessed by AI-assisted chest HRCT. *PLoS One* **16**, (2021).
455. Cartier, A. & Hla, T. Sphingosine 1-phosphate: Lipid signaling in pathology and therapy. *Science* vol. 366 Preprint at <https://doi.org/10.1126/science.aar5551> (2019).
456. Roslund, K. *et al.* Identifying volatile in vitro biomarkers for oral bacteria with proton-transfer-reaction mass spectrometry and gas chromatography–mass spectrometry. *Sci Rep* **11**, (2021).
457. Wang, Y. *et al.* Estimated assessment of dietary exposure to artificial sweeteners from processed food in Nanjing, China. *Food Addit Contam Part A Chem Anal Control Expo Risk Assess* **38**, 1105–1117 (2021).
458. Eichelbaum, M., Hengstmann, J. H., Rest, H. D., Brecht, T. & Dcngler, H. J. *Pharmacokinetics, Cardiovascular and Metabolic Actions of Cyclohexylamine in Man* ***. *Arch. Toxikol* vol. 31 (1974).
459. Liu, G., Lin, C. J., Yates, C. R. & Prasad, G. L. Metabolomic Analysis Identified Reduced Levels of Xenobiotics, Oxidative Stress, and Improved Vitamin Metabolism in Smokers Switched to Vuse Electronic Nicotine Delivery System. *Nicotine and Tobacco Research* **23**, 1133–1142 (2021).
460. Chen, X., Gu, M., Li, T. & Sun, Y. Metabolite reanalysis revealed potential biomarkers for COVID-19: A potential link with immune response. *Future Microbiol* **16**, 577–588 (2021).
461. Cobre, A. F. *et al.* Influence of foods and nutrients on COVID-19 recovery: A multivariate analysis of data from 170 countries using a generalized linear model. *Clinical Nutrition* **41**, 3077–3084 (2022).
462. Sikaroudi, M. K. *et al.* Assessment of anorexia and weight loss during the infection and recovery period of patients with coronavirus disease 2019 (COVID-19). *Clinical Nutrition Open Science* **40**, 102–110 (2021).
463. Di Filippo, L. *et al.* COVID-19 is associated with clinically significant weight loss and risk of malnutrition, independent of hospitalisation: A post-hoc analysis of a prospective cohort study. *Clinical Nutrition* **40**, 2420–2426 (2021).
464. van der Voort, P. H. J. *et al.* Leptin levels in SARS-CoV-2 infection related respiratory failure: A cross-sectional study and a pathophysiological framework on the role of fat tissue. *Heliyon* **6**, (2020).

465. Feehan, J., de Courten, M., Apostolopoulos, V. & de Courten, B. Nutritional interventions for covid-19: A role for carnosine? *Nutrients* vol. 13 Preprint at <https://doi.org/10.3390/nu13051463> (2021).
466. Wang, Q. *et al.* *V I R O L O G Y O-GlcNAc Transferase Promotes Influenza A Virus-Induced Cytokine Storm by Targeting Interferon Regulatory Factor-5.* <https://www.science.org> (2020).
467. Ayres, J. S. A metabolic handbook for the COVID-19 pandemic. *Nat Metab* **2**, 572–585 (2020).
468. Laviada-Molina, H. A., Leal-Berumen, I., Rodriguez-Ayala, E. & Bastarrachea, R. A. Working Hypothesis for Glucose Metabolism and SARS-CoV-2 Replication: Interplay Between the Hexosamine Pathway and Interferon RF5 Triggering Hyperinflammation. Role of BCG Vaccine? *Front Endocrinol (Lausanne)* **11**, (2020).
469. Kryukov, E. V. *et al.* Association of Low Molecular Weight Plasma Amino thiols with the Severity of Coronavirus Disease 2019. *Oxid Med Cell Longev* **2021**, (2021).
470. Pei, L. *et al.* Plasma Metabolomics Reveals Dysregulated Metabolic Signatures in HIV-Associated Immune Reconstitution Inflammatory Syndrome. *Front Immunol* **12**, (2021).
471. Jiang, H. & Mei, Y. F. Sars–cov–2 spike impairs dna damage repair and inhibits v(D)j recombination in vitro. *Viruses* **13**, (2021).
472. Rees, C. A. *et al.* Altered amino acid profile in patients with SARS-CoV-2 infection. *Proc Natl Acad Sci U S A* **118**, (2021).
473. Zhang, J. *et al.* Esophageal cancer metabolite biomarkers detected by LC-MS and NMR methods. *PLoS One* **7**, (2012).
474. Rees, C. A. *et al.* Altered amino acid profile in patients with SARS-CoV-2 infection. *Proc Natl Acad Sci U S A* **118**, (2021).
475. Luporini, R. L. *et al.* Phenylalanine and COVID-19: Tracking disease severity markers. *Int Immunopharmacol* **101**, (2021).
476. Kamel, K. S., Oh, M. S. & Halperin, M. L. L-lactic acidosis: pathophysiology, classification, and causes; emphasis on biochemical and metabolic basis. *Kidney International* vol. 97 75–88 Preprint at <https://doi.org/10.1016/j.kint.2019.08.023> (2020).

477. Nechipurenko, Y. D. *et al.* biology The Role of Acidosis in the Pathogenesis of Severe Forms of COVID-19. (2021) doi:10.3390/biology.
478. De Backer, D. Lactic acidosis. *Intensive Care Med* **29**, 699–702 (2003).
479. Li, J. *et al.* COVID-19 infection may cause ketosis and ketoacidosis. *Diabetes Obes Metab* **22**, 1935–1941 (2020).
480. Carpenè, G. *et al.* Blood lactate concentration in COVID-19: A systematic literature review. *Clinical Chemistry and Laboratory Medicine* vol. 60 332–337 Preprint at <https://doi.org/10.1515/cclm-2021-1115> (2022).
481. Cruzat, V., Rogero, M. M., Keane, K. N., Curi, R. & Newsholme, P. Glutamine: Metabolism and immune function, supplementation and clinical translation. *Nutrients* vol. 10 Preprint at <https://doi.org/10.3390/nu10111564> (2018).
482. Leite, J. S. M., Cruzat, V. F., Krause, M. & Homem de Bittencourt, P. I. Physiological regulation of the heat shock response by glutamine: implications for chronic low-grade inflammatory diseases in age-related conditions. *Nutrire* vol. 41 Preprint at <https://doi.org/10.1186/s41110-016-0021-y> (2016).
483. Shi, D. *et al.* The serum metabolome of COVID-19 patients is distinctive and predictive. *Metabolism* **118**, (2021).
484. Wu, D. *et al.* Plasma metabolomic and lipidomic alterations associated with COVID-19. *Natl Sci Rev* **7**, 1157–1168 (2020).
485. Chanda, B. *et al.* Glycerol-3-phosphate is a critical mobile inducer of systemic immunity in plants. *Nat Genet* **43**, 421–429 (2011).
486. Abbas, A. K. *et al.* S-sulfo-cysteine is an endogenous amino acid in neonatal rat brain but an unlikely mediator of cysteine neurotoxicity. *Neurochem Res* **33**, 301–307 (2008).
487. Cai, Y. *et al.* Kynurenic Acid May Underlie Sex-Specific Immune Responses to COVID-19. *Sci. Signal* vol. 14 <https://www.science.org> (2021).
488. Mohamed, K., Yazdanpanah, N., Saghazadeh, A. & Rezaei, N. Computational drug discovery and repurposing for the treatment of COVID-19: A systematic review. *Bioorganic Chemistry* vol. 106 Preprint at <https://doi.org/10.1016/j.bioorg.2020.104490> (2021).
489. Walls, A. C. *et al.* Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **181**, 281-292.e6 (2020).
490. Mannar, D. *et al.* SARS-CoV-2 Omicron Variant: Antibody Evasion and Cryo-EM Structure of Spike Protein-ACE2 Complex. <https://www.science.org>.

491. Harvey, W. T. *et al.* SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* vol. 19 409–424 Preprint at <https://doi.org/10.1038/s41579-021-00573-0> (2021).
492. Vemula, D., Jayasurya, P., Sushmitha, V., Kumar, Y. N. & Bhandari, V. CADD, AI and ML in drug discovery: A comprehensive review. *European Journal of Pharmaceutical Sciences* vol. 181 Preprint at <https://doi.org/10.1016/j.ejps.2022.106324> (2023).
493. Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data. *Nucleic Acids Res* **35**, (2007).
494. Yan, R. *et al.* Structural Basis for the Recognition of SARS-CoV-2 by Full-Length Human ACE2. *Science* vol. 367 <https://www.science.org> (2020).
495. Mannar, D. *et al.* SARS-CoV-2 Omicron Variant: Antibody Evasion and Cryo-EM Structure of Spike Protein-ACE2 Complex. <https://www.science.org>.
496. Ravi, L. & Krishnan, K. A Handbook On Protein-Ligand Docking Tool: AutoDock4 A HANDBOOK ON PROTEINLIGAND DOCKING TOOL: AUTODOCK4. (2016).
497. Søndergaard, C. R., Olsson, M. H. M., Rostkowski, M. & Jensen, J. H. Improved treatment of ligands and coupling effects in empirical calculation and rationalization of p K a values. *J Chem Theory Comput* **7**, 2284–2295 (2011).
498. Olsson, M. H. M., Søndergaard, C. R., Rostkowski, M. & Jensen, J. H. PROPKA3: Consistent treatment of internal and surface residues in empirical p K a predictions. *J Chem Theory Comput* **7**, 525–537 (2011).
499. Yan, R. *et al.* Structural Basis for the Recognition of SARS-CoV-2 by Full-Length Human ACE2. *Science* vol. 367 <https://www.science.org> (2020).
500. Mannar, D. *et al.* SARS-CoV-2 Omicron Variant: Antibody Evasion and Cryo-EM Structure of Spike Protein-ACE2 Complex. <https://www.science.org>.
501. Sterling, T. & Irwin, J. J. ZINC 15 - Ligand Discovery for Everyone. *J Chem Inf Model* **55**, 2324–2337 (2015).
502. Dallakyan, S. & Olson, A. J. Small-molecule library screening by docking with PyRx. *Methods in Molecular Biology* **1263**, 243–250 (2015).
503. Gasteiger, J. & Marsili, M. ITERATIVE PARTIAL EQUALIZATION OF ORBITAL ELECTRONEGATIVITY-A RAPID ACCESS TO ATOMIC CHARGES. (1980).
504. O’Boyle, N. M. *et al.* Open Babel: An Open chemical toolbox. *J Cheminform* **3**, (2011).

505. Kirchmair, J., Markt, P., Distinto, S., Wolber, G. & Langer, T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection - What can we learn from earlier mistakes? *Journal of Computer-Aided Molecular Design* vol. 22 213–228 Preprint at <https://doi.org/10.1007/s10822-007-9163-6> (2008).
506. Allouche, A. R. Gabedita - A graphical user interface for computational chemistry softwares. *J Comput Chem* **32**, 174–182 (2011).
507. Cuzzolin, A., Sturlese, M., Malvacio, I., Ciancetta, A. & Moro, S. DockBench: An integrated informatic platform bridging the gap between the robust validation of docking protocols and virtual screening simulations. *Molecules* **20**, 9977–9993 (2015).
508. Vieira, T. F. & Sousa, S. F. Comparing AutoDock and Vina in ligand/decoy discrimination for virtual screening. *Applied Sciences (Switzerland)* **9**, (2019).
509. Alnajjar, R., Mostafa, A., Kandeil, A. & Al-Karmalawy, A. A. Molecular docking, molecular dynamics, and in vitro studies reveal the potential of angiotensin II receptor blockers to inhibit the COVID-19 main protease. *Heliyon* **6**, (2020).
510. Hosseini, F. S. & Amanlou, M. Anti-HCV and anti-malaria agent, potential candidates to repurpose for coronavirus infection: Virtual screening, molecular docking, and molecular dynamics simulation study. *Life Sci* **258**, (2020).
511. Brink, T. Ten & Exner, T. E. Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. *J Chem Inf Model* **49**, 1535–1546 (2009).
512. Ten Brink, T. & Exner, T. E. PKa based protonation states and microspecies for protein-ligand docking. *J Comput Aided Mol Des* **24**, 935–942 (2010).
513. Xiong, G. *et al.* ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Res* **49**, W5–W14 (2021).
514. Abraham, M. J. *et al.* Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
515. Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *Journal of Computational Chemistry* vol. 26 1781–1802 Preprint at <https://doi.org/10.1002/jcc.20289> (2005).

516. Brodie, B. R., Gilette, J. R. & Ladu, B. N. *1071 Choline-Induced Constriction of the Trachea. References and Notes (1) Presented at the Seventh Northeast Regional Meeting of The. Journal of Medicinal Chemistry* vol. 20 (1977).
517. Cheatham, T. E., Miller, J. L., Fox, T., Darden, T. A. & Kollman, P. A. Molecular Dynamics Simulations on Solvated Biomolecular Systems: The Particle Mesh Ewald Method Leads to Stable Trajectories of DNA, RNA, and Proteins. *J. Am. Chem. SOC* **117**, 226–245 (1995).
518. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* **31**, 671–690 (2010).
519. Justin A. Lemkul. From Proteins to Perturbed Hamiltonians: A Suite of Tutorials for the GROMACS-2018 Molecular Simulation Package [Article v1.0]. *Living J Comput Mol Sci* (2018).
520. Field, M. J., Albe, M., Bret, C., Martin, P.-D. & Thomas, A. *The Dynamo Library for Molecular Simulations Using Hybrid Quantum Mechanical and Molecular Mechanical Potentials. Journal of Computational Chemistry* vol. 21 <http://www.ibs.fr/ext/labos/LDM/projet6/> (2000).
521. Oany, A. R. *et al.* Design of novel viral attachment inhibitors of the spike glycoprotein (S) of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) through virtual screening and dynamics. *Int J Antimicrob Agents* **56**, (2020).
522. Rakib, A. *et al.* Immunoinformatics-guided design of an epitope-based vaccine against severe acute respiratory syndrome coronavirus 2 spike glycoprotein. *Comput Biol Med* **124**, (2020).
523. Islam, M. S., Mahmud, S., Sultana, R. & Dong, W. Identification and in silico molecular modelling study of newly isolated *Bacillus subtilis* SI-18 strain against S9 protein of *Rhizoctonia solani*. *Arabian Journal of Chemistry* **13**, 8600–8612 (2020).
524. Humphrey, W., Dalke, A. & Schulten, K. *VMD: Visual Molecular Dynamics*. (1996).
525. Valdés-Tresanco, M. S., Valdés-Tresanco, M. E., Valiente, P. A. & Moreno, E. Gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS. *J Chem Theory Comput* **17**, 6281–6291 (2021).

526. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery* vol. 10 449–461 Preprint at <https://doi.org/10.1517/17460441.2015.1032936> (2015).
527. Elsbaey, M., Ibrahim, M. A. A., Bar, F. A. & Elgazar, A. A. Chemical constituents from coconut waste and their in silico evaluation as potential antiviral agents against SARS-CoV-2. *South African Journal of Botany* **141**, 278–289 (2021).
528. Pang, J., Gao, S., Sun, Z. & Yang, G. Discovery of small molecule PLpro inhibitor against COVID-19 using structure-based virtual screening, molecular dynamics simulation, and molecular mechanics/Generalized Born surface area (MM/GBSA) calculation. *Struct Chem* **32**, 879–886 (2021).
529. Simeon, S. *et al.* Probing the origins of human acetylcholinesterase inhibition via QSAR modeling and molecular docking. *PeerJ* **4**, e2322 (2016).
530. Yap, C. W. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* **32**, 1466–1474 (2011).
531. Jaganathan, K., Tayara, H. & Chong, K. T. An Explainable Supervised Machine Learning Model for Predicting Respiratory Toxicity of Chemicals Using Optimal Molecular Descriptors. *Pharmaceutics* **14**, (2022).
532. Lipinski, C. A. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol* **1**, 337–341 (2004).
533. Hira, Z. M. & Gillies, D. F. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data. *Adv Bioinformatics* **2015**, 198363 (2015).
534. May, R. J., Maier, H. R. & Dandy, G. C. Data splitting for artificial neural networks using SOM-based stratified sampling. *Neural Netw* **23**, 283–294 (2010).
535. Chicco, D., Warrens, M. J. & Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput Sci* **7**, e623 (2021).
536. Wei, W. *et al.* Application of a Combined Model with Autoregressive Integrated Moving Average (ARIMA) and Generalized Regression Neural Network (GRNN) in Forecasting Hepatitis Incidence in Heng County, China. *PLoS One* **11**, e0156768 (2016).
537. Futagami, K., Fukazawa, Y., Kapoor, N. & Kito, T. Pairwise acquisition prediction with SHAP value interpretation. *The Journal of Finance and Data Science* **7**, 22–44 (2021).

538. Meng, Y., Yang, N., Qian, Z. & Zhang, G. What makes an online review more helpful: An interpretation framework using xgboost and shap values. *Journal of Theoretical and Applied Electronic Commerce Research* **16**, 466–490 (2021).
539. Chtita, S. *et al.* QSAR study of anti-Human African Trypanosomiasis activity for 2-phenylimidazopyridines derivatives using DFT and Lipinski's descriptors. *Heliyon* e01304 (2019) doi:10.1016/j.heliyon.2019.
540. Paliwal, S., Seth, D., Yadav, D., Yadav, R. & Paliwal, S. Development of a robust QSAR model to predict the affinity of pyrrolidine analogs for dipeptidyl peptidase IV (DPP-IV). *J Enzyme Inhib Med Chem* **26**, 129–140 (2011).
541. Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* **32**, 1466–1474 (2011).
542. Senawi, A., Wei, H. L. & Billings, S. A. A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. *Pattern Recognit* **67**, 47–61 (2017).
543. Lindner, T., Puck, J. & Verbeke, A. Beyond addressing multicollinearity: Robust quantitative analysis and machine learning in international business research. *Journal of International Business Studies* vol. 53 1307–1314 Preprint at <https://doi.org/10.1057/s41267-022-00549-z> (2022).
544. Garg, A. & Tai, K. *Comparison of Statistical and Machine Learning Methods in Modelling of Data with Multicollinearity*. *Int. J. Modelling, Identification and Control* vol. 18 (2013).
545. Chekol Abebe, E. *et al.* Mutational Pattern, Impacts and Potential Preventive Strategies of Omicron SARS-CoV-2 Variant Infection. *Infect Drug Resist* **15**, 1871–1887 (2022).
546. Callaway, E. Heavily mutated Omicron variant puts scientists on alert. *Nature* vol. 600 21 Preprint at <https://doi.org/10.1038/d41586-021-03552-w> (2021).
547. Brüssow, H. COVID-19: Omicron – the latest, the least virulent, but probably not the last variant of concern of SARS-CoV-2. *Microbial Biotechnology* vol. 15 1927–1939 Preprint at <https://doi.org/10.1111/1751-7915.14064> (2022).
548. Mannar, D. *et al.* SARS-CoV-2 Omicron Variant: Antibody Evasion and Cryo-EM Structure of Spike Protein-ACE2 Complex. <https://www.science.org>.
549. Ao, D. *et al.* SARS-CoV-2 Omicron variant: Immune escape and vaccine development. *MedComm* vol. 3 Preprint at <https://doi.org/10.1002/mco2.126> (2022).

550. Fan, Y. *et al.* SARS-CoV-2 Omicron variant: recent progress and future perspectives. *Signal Transduction and Targeted Therapy* vol. 7 Preprint at <https://doi.org/10.1038/s41392-022-00997-x> (2022).
551. Badavath, V. N. *et al.* Determination of potential inhibitors based on isatin derivatives against SARS-CoV-2 main protease (mpro): a molecular docking, molecular dynamics and structure-activity relationship studies. *J Biomol Struct Dyn* **40**, 3110–3128 (2022).
552. Sen Gupta, P. S., Biswal, S., Panda, S. K., Ray, A. K. & Rana, M. K. Binding mechanism and structural insights into the identified protein target of COVID-19 and importin- α with in-vitro effective drug ivermectin. *J Biomol Struct Dyn* **40**, 2217–2226 (2022).
553. Gangadevi, S. *et al.* Kobophenol A Inhibits Binding of Host ACE2 Receptor with Spike RBD Domain of SARS-CoV-2, a Lead Compound for Blocking COVID-19. *Journal of Physical Chemistry Letters* **12**, 1793–1802 (2021).
554. kumar, B. H., Manandhar, S., Mehta, C. H., Nayak, U. Y. & Pai, K. S. R. Structure-based docking, pharmacokinetic evaluation, and molecular dynamics-guided evaluation of traditional formulation against SARS-CoV-2 spike protein receptor bind domain and ACE2 receptor complex. *Chemical Papers* **76**, 1063–1083 (2022).
555. Poli, G., Martinelli, A. & Tuccinardi, T. Reliability analysis and optimization of the consensus docking approach for the development of virtual screening studies. *J Enzyme Inhib Med Chem* **31**, 167–173 (2016).
556. Tuccinardi, T., Poli, G., Romboli, V., Giordano, A. & Martinelli, A. Extensive consensus docking evaluation for ligand pose prediction and virtual screening studies. *J Chem Inf Model* **54**, 2980–2986 (2014).
557. Brodie, B. R., Gilette, J. R. & Ladu, B. N. 1071 Choline-Induced Constriction of the Trachea. *References and Notes (1) Presented at the Seventh Northeast Regional Meeting of The. Journal of Medicinal Chemistry* vol. 20 (1977).
558. Kozakov, D., Clodfelter, K. H., Vajda, S. & Camacho, C. J. Optimal clustering for detecting near-native conformations in protein docking. *Biophys J* **89**, 867–875 (2005).
559. Mukherjee, S., Balias, T. E. & Rizzo, R. C. Docking validation resources: Protein family and ligand flexibility experiments. *J Chem Inf Model* **50**, 1986–2000 (2010).

560. Kalathiya, U., Padariya, M., Fahraeus, R., Chakraborti, S. & Hupp, T. R. Multivalent display of sars-cov-2 spike (Rbd domain) of covid-19 to nanomaterial, protein ferritin nanocages. *Biomolecules* **11**, 1–12 (2021).
561. Sabzian-Molaei, F. *et al.* Urtica dioica Agglutinin: A plant protein candidate for inhibition of SARS-COV-2 receptor-binding domain for control of Covid19 Infection. *PLoS One* **17**, (2022).
562. Taka, E. *et al.* Critical Interactions between the SARS-CoV-2 Spike Glycoprotein and the Human ACE2 Receptor. *Journal of Physical Chemistry B* **125**, 5537–5548 (2021).
563. Jamroz, M., Kolinski, A. & Kmiecik, S. CABS-flex predictions of protein flexibility compared with NMR ensembles. *Bioinformatics* **30**, 2150–2154 (2014).
564. Vardhan, S. & Sahoo, S. K. In silico ADMET and molecular docking study on searching potential inhibitors from limonoids and triterpenoids for COVID-19. *Comput Biol Med* **124**, (2020).
565. Alnajjar, R., Mostafa, A., Kandeil, A. & Al-Karmalawy, A. A. Molecular docking, molecular dynamics, and in vitro studies reveal the potential of angiotensin II receptor blockers to inhibit the COVID-19 main protease. *Heliyon* **6**, (2020).
566. Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery* vol. 10 449–461 Preprint at <https://doi.org/10.1517/17460441.2015.1032936> (2015).
567. Tutunchi, H., Naeini, F., Ostadrahimi, A. & Hosseinzadeh-Attar, M. J. Naringenin, a flavanone with antiviral and anti-inflammatory effects: A promising treatment strategy against COVID-19. *Phytotherapy Research* vol. 34 3137–3147 Preprint at <https://doi.org/10.1002/ptr.6781> (2020).
568. Zeng, W., Jin, L., Zhang, F., Zhang, C. & Liang, W. Naringenin as a potential immunomodulator in therapeutics. *Pharmacological Research* vol. 135 122–126 Preprint at <https://doi.org/10.1016/j.phrs.2018.08.002> (2018).
569. Majumder, R. & Mandal, M. Screening of plant-based natural compounds as a potential COVID-19 main protease inhibitor: an in silico docking and molecular dynamics simulation approach. *J Biomol Struct Dyn* **40**, 696–711 (2022).
570. Mishra, D. *et al.* Structurally modified compounds of hydroxychloroquine, remdesivir and tetrahydrocannabinol against main protease of SARS-CoV-2, a possible hope for COVID-19: Docking and molecular dynamics simulation studies. *J Mol Liq* **335**, (2021).

571. Rashdan, H. R. M. & Abdelmonsef, A. H. In silico study to identify novel potential thiadiazole-based molecules as anti-Covid-19 candidates by hierarchical virtual screening and molecular dynamics simulations. *Struct Chem* **33**, 1727–1739 (2022).
572. Sepay, N., Sekar, A., Halder, U. C., Alarifi, A. & Afzal, M. Anti-COVID-19 terpenoid from marine sources: A docking, admet and molecular dynamics study. *J Mol Struct* **1228**, (2021).
573. Farhadi, F., Iranshahi, M., Mohtashami, L., Shakeri Asil, S. & Iranshahi, M. Metabolic differences of two *Ferula* species as potential sources of galbanum: An NMR-based metabolomics study. *Phytochemical Analysis* **32**, 811–819 (2021).
574. Amin, A. *et al.* Antiprotozoal and antiglycation activities of sesquiterpene coumarins from *Ferula narthex* exudate. *Molecules* **21**, (2016).
575. Amin, A. *et al.* Studies on effects of umbelliferon derivatives against periodontal bacteria; antibiofilm, inhibition of quorum sensing and molecular docking analysis. *Microb Pathog* **144**, (2020).
576. Mollazadeh, S. *et al.* The enhancement of vincristine cytotoxicity by combination with feselol. *J Asian Nat Prod Res* **12**, 569–575 (2010).
577. Iranshahi, M. *et al.* Drimane-type sesquiterpene coumarins from *ferula gummosa* fruits enhance doxorubicin uptake in doxorubicin-resistant human breast cancer cell line. *J Tradit Complement Med* **4**, 118–125 (2014).
578. Dearden, J. C., Cronin, M. T. D., Zhao, Y.-H. & Raevsky, O. A. *QSAR Studies of Compounds Acting by Polar and Non-Polar Narcosis: An Examination of the Role of Polarisability and Hydrogen Bonding* {
<http://www.ibmh.msk.su/qsar/molpro>.
579. Brink, T. Ten & Exner, T. E. Influence of protonation, tautomeric, and stereoisomeric states on protein-ligand docking results. *J Chem Inf Model* **49**, 1535–1546 (2009).
580. Jena, N. R. Role of different tautomers in the base-pairing abilities of some of the vital antiviral drugs used against COVID-19. *Physical Chemistry Chemical Physics* **22**, 28115–28122 (2020).
581. Umar, Y. Theoretical studies of the rotational and tautomeric states, electronic and spectroscopic properties of favipiravir and its structural analogues: a

- potential drug for the treatment of COVID-19. *Journal of Taibah University for Science* **14**, 1613–1625 (2020).
582. Fischer, A., Sellner, M., Neranjan, S., Smieško, M. & Lill, M. A. Potential inhibitors for novel coronavirus protease identified by virtual screening of 606 million compounds. *Int J Mol Sci* **21**, (2020).
583. Lutgens, A. *et al.* Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *J Am Chem Soc* **144**, 2905–2920 (2022).
584. Meyer-Almes, F. J. Repurposing approved drugs as potential inhibitors of 3CL-protease of SARS-CoV-2: Virtual screening and structure based drug design. *Comput Biol Chem* **88**, (2020).
585. Kalliokoski, T., Salo, H. S., Lahtela-Kakkonen, M. & Poso, A. The effect of ligand-based tautomer and protomer prediction on structure-based virtual screening. *J Chem Inf Model* **49**, 2742–2748 (2009).
586. De Camp, W. H. *The FDA Perspective on the Development of Stereoisomers. CHIRALITY* vol. 1 (1989).
587. Pifferi, G. & Perucca, E. *The Cost Benefit Ratio of Enantiomeric Drugs. EUROPEAN JOURNAL OF DRUG METABOLISM AND PHARMACOKINETICS* vol. 20 (1995).
588. Al-Asmari, K. M. *et al.* Arabica coffee and olive oils mitigate malathion-induced nephrotoxicity in rat: In silico, immunohistochemical and biochemical evaluation. *Saudi J Biol Sci* **29**, (2022).
589. El-Moamly, A. A. Scabies as a part of the World Health Organization roadmap for neglected tropical diseases 2021–2030: what we know and what we need to do for global control. *Tropical Medicine and Health* vol. 49 Preprint at <https://doi.org/10.1186/s41182-021-00348-6> (2021).
590. World Health Organization. HIV and AIDS. <https://www.who.int/news-room/fact-sheets/detail/hiv-aids> (2023).
591. Olaru, I. D. *et al.* Global prevalence of hepatitis B or hepatitis C infection among patients with tuberculosis disease: systematic review and meta-analysis. *EClinicalMedicine* **58**, (2023).
592. Alberts, C. J. *et al.* Worldwide prevalence of hepatitis B virus and hepatitis C virus among patients with cirrhosis at country, region, and global levels: a systematic review. *Lancet Gastroenterol Hepatol* **7**, 724–735 (2022).

593. De Chasse, B., Meyniel-Schicklin, L., Aublin-Gex, A., André, P. & Lotteau, V. New horizons for antiviral drug discovery from virus-host protein interaction networks. *Current Opinion in Virology* vol. 2 606–613 Preprint at <https://doi.org/10.1016/j.coviro.2012.09.001> (2012).
594. Jing, Y., Bian, Y., Hu, Z., Wang, L. & Xie, X. Q. S. Deep Learning for Drug Design: an Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. *AAPS Journal* vol. 20 Preprint at <https://doi.org/10.1208/s12248-018-0210-0> (2018).
595. Wishart, D. S. Emerging applications of metabolomics in drug discovery and precision medicine. *Nature Reviews Drug Discovery* vol. 15 473–484 Preprint at <https://doi.org/10.1038/nrd.2016.32> (2016).
596. Wishart, D. S. *et al.* HMDB 5.0: The Human Metabolome Database for 2022. *Nucleic Acids Res* **50**, D622–D631 (2022).
597. Yasutake, Y. *et al.* HIV-1 with HBV-associated Q151M substitution in RT becomes highly susceptible to entecavir: Structural insights into HBV-RT inhibition by entecavir. *Sci Rep* **8**, (2018).
598. National Institutes of Health (USA). *Guidelines for the Use of Antiretroviral Agents in Adults and Adolescents with HIV Developed by the DHHS Panel on Antiretroviral Guidelines for Adults and Adolescents-A Working Group of the Office of AIDS Research Advisory Council (OARAC) How to Cite the Adult and Adolescent Guidelines: Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the Use of Antiretroviral Agents in Adults And.* <http://hivinfo.nih.gov>.
599. Matthews, S. J. & Lancaster, J. W. Telaprevir: A Hepatitis C NS3/4A Protease Inhibitor. *Clinical Therapeutics* vol. 34 1857–1882 Preprint at <https://doi.org/10.1016/j.clinthera.2012.07.011> (2012).
600. Lok, A. S. F. *et al.* Antiviral Therapy for Chronic Hepatitis B Viral Infection in Adults: A Systematic Review and Meta-Analysis. (2015) doi:10.1002/hep.28280/supinfo.
601. Tang, H., Griffin, J., Innaimo, S., Lehman-Mckeeman, L. & Llamoso, C. The discovery and development of a potent antiviral drug, entecavir, for the treatment of chronic hepatitis b. *Journal of Clinical and Translational Hepatology* vol. 1 51–58 Preprint at <https://doi.org/10.14218/JCTH.2013.00006> (2013).

602. Yang, J., Sun, H. & Liu, Q. The comparative efficacy and safety of entecavir and lamivudine in patients with HBV-associated acute-on-chronic liver failure: A systematic review and meta-analysis. *Gastroenterology Research and Practice* vol. 2016 Preprint at <https://doi.org/10.1155/2016/5802674> (2016).
603. Huang, K. W., Tam, K. W., Luo, J. C. & Kuan, Y. C. Efficacy and Safety of Lamivudine Versus Entecavir for Treating Chronic Hepatitis B Virus-related Acute Exacerbation and Acute-on-Chronic Liver Failure. *J Clin Gastroenterol* **51**, 539–547 (2017).
604. Kang, K. Bin *et al.* Jubanines F-J, cyclopeptide alkaloids from the roots of *Ziziphus jujuba*. *Phytochemistry* **119**, 90–95 (2015).
605. Orhan, I. E. & Senol Deniz, F. S. Natural Products as Potential Leads Against Coronaviruses: Could They be Encouraging Structural Models Against SARS-CoV-2? *Natural Products and Bioprospecting* vol. 10 171–186 Preprint at <https://doi.org/10.1007/s13659-020-00250-4> (2020).
606. Bray, M. Highly pathogenic RNA viral infections: Challenges for antiviral research. *Antiviral Research* vol. 78 1–8 Preprint at <https://doi.org/10.1016/j.antiviral.2007.12.007> (2008).
607. Noske, G. D. *et al.* An in-solution snapshot of SARS-COV-2 main protease maturation process and inhibition. *Nat Commun* **14**, (2023).
608. Barnes-Seeman, D. *et al.* Design and synthesis of lactam-thiophene carboxylic acids as potent hepatitis C virus polymerase inhibitors. *Bioorg Med Chem Lett* **24**, 3979–3985 (2014).
609. Ando, I. *et al.* Preclinical characterization of JTK-853, a novel nonnucleoside inhibitor of the hepatitis C virus RNA-dependent RNA polymerase. *Antimicrob Agents Chemother* **56**, 4250–4256 (2012).

APÊNDICE I – BIOGRAFIA DO AUTOR

Nascido em 30 de abril de 1993, em um vilarejo na montanhosa Província de Nampula, no Norte de Moçambique, Alexandre de Fátima Cobre é o mais jovem de uma família simples de camponeses, com dois irmãos e duas irmãs. Ele foi o primeiro de sua vila a se tornar doutor e o primeiro de sua família a frequentar o ensino superior.

Entre 2000 e 2007, concluiu o ensino primário em uma escola pública, percorrendo diariamente cerca de 20 quilômetros descalço, atravessando mata e enfrentando diversos perigos, como chuvas e animais selvagens. Assim como a realidade de muitas escolas em Moçambique, as salas de aulas eram embaixo das árvores, e recebeu todas as aulas no ensino primário sentando-se no chão, e quando chovia, as aulas ficavam encerradas.

De 2008 a 2011, frequentou o ensino médio também em uma escola pública, sendo o único do vilarejo a continuar os estudos. Apesar dos desafios, tornou-se monitor e foi premiado como o melhor aluno da escola.

Em 2012, prestou vestibular para o curso de Farmácia em uma Universidade pública da sua cidade (Universidade Lúrio), onde foi aprovado em primeiro lugar e conquistou uma bolsa de estudos.

De 2014 a 2016, durante a sua graduação em Farmácia, destacou-se nas jornadas acadêmicas da universidade, sendo seus trabalhos científicos classificados em primeiro lugar por três anos consecutivos. Em 2016, terminou a graduação como o melhor estudante.

No início de 2017, candidatou-se a uma bolsa de estudo de mestrado no Brasil através do programa PEC-PG do CNPq. Para isso, precisou comprovar proficiência em português fazendo o exame CELP-BRAS na embaixada brasileira em Moçambique, localizada na capital do País. Durante a viagem para realizar o exame, o seu ônibus foi atacado por rebeldes, ficando parado por quatro dias no meio da mata. Apesar de ter passado no exame, não foi contemplado com a bolsa, e, como não conhecia ninguém na capital, passou duas noites dormindo na rua.

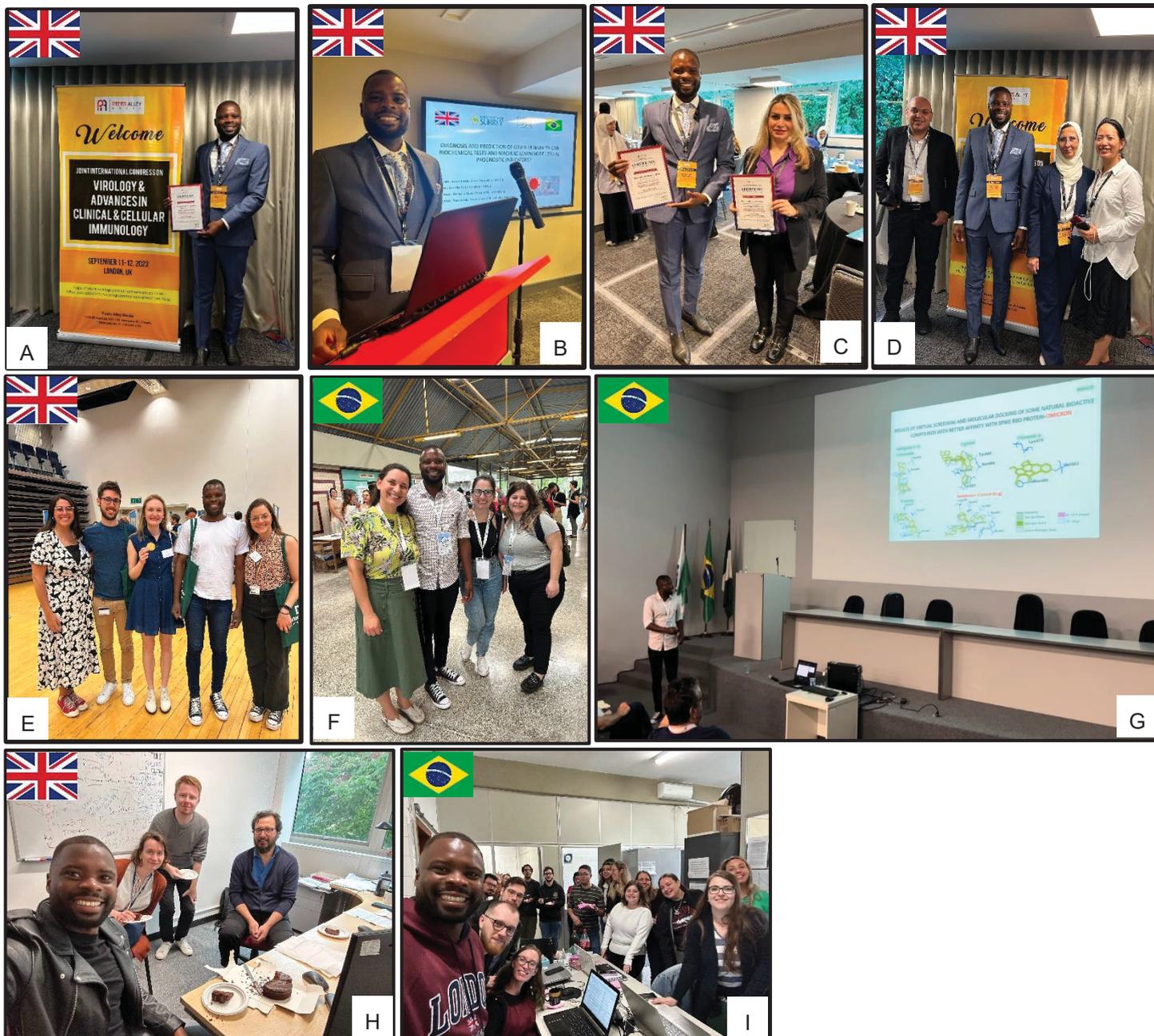
Em 2018, chegou ao Brasil com a ajuda do Professor Pontarolo, que o apoiou em todas as necessidades básicas. Concluiu o mestrado em fevereiro de 2020 e, em abril do mesmo ano, foi aceito no programa de doutorado. Apesar das dificuldades com bolsas, o Professor Pontarolo o sustentou para evitar sua repatriação.

Entre 2022 e 2023, concluiu o curso de especialização em *Data Science & Big Data* no departamento de Estatística da Universidade Federal do Paraná. Seu trabalho final de conclusão do curso concentrou-se no desenvolvimento de uma inteligência artificial e modelos de machine learning para a descoberta de novos fármacos visando o tratamento do HIV/AIDS.

Ao longo dos quatro anos do doutorado, produziu mais de 30 artigos científicos com o apoio dos orientadores, Professor Pontarolo e Profa. Fernanda.

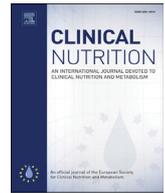
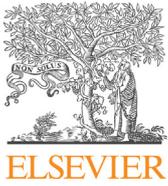
Em 2023, financiado pelo Professor Pontarolo, realizou um estágio na University of Surrey, na Inglaterra. Devido ao seu melhor desempenho acadêmico na Inglaterra, no dia 23 de março de 2024 foi contratado pelo governo Britânico para trabalhar como professor e cientista na The University of Manchester.

APÊNDICE II – GALERIA DE FOTOS DURANTE A PARTICIPAÇÃO COMO PALESTRANTE EM EVENTOS CIENTÍFICOS INTERNACIONAIS NO REINO UNIDO (INGLATERRA) E NO BRASIL



Legenda: Fotos A, B, C e D - Londres (Reino Unido), 11-12 de setembro: Palestra em evento científico internacional de virologia e imunologia, apresentando um dos artigos do doutorado. Foto E - Guildford, Reino Unido, 15 de junho de 2023: Participação em evento internacional na University of Surrey, onde realizei o estágio de doutorado sanduíche. Fotos F e G - UFPR (Brasil), 14 de novembro: Palestra em evento internacional organizado pela UFPR, apresentando um dos artigos da tese de doutorado. Fotos H e I - Grupos de pesquisa na Inglaterra e no Brasil, respectivamente.

**APÊNDICE III – ARTIGOS CIENTÍFICOS PUBLICADOS
(PRIMEIRA PÁGINA)**



Covid-19

Influence of foods and nutrients on COVID-19 recovery: A multivariate analysis of data from 170 countries using a generalized linear model



Alexandre F. Cobre ^a, Monica Surek ^a, Raquel O. Vilhena ^a, Beatriz Böger ^a, Mariana M. Fachi ^a, Danilo R. Momade ^a, Fernanda S. Tonin ^a, Flavia M. Sarti ^b, Roberto Pontarolo ^{c,*}

^a Pharmaceutical Sciences Postgraduate Program, Federal University of Paraná, Curitiba, Brazil

^b Complex Systems Modelling Postgraduate Program, University of Sao Paulo, Sao Paulo, Brazil

^c Department of Pharmacy, Federal University of Parana, Curitiba, Brazil

ARTICLE INFO

Article history:

Received 18 January 2021

Accepted 15 March 2021

Keywords:

Coronavirus

Macronutrients

Nutrients

Hunger

Multivariate analysis

SUMMARY

Background & aims: COVID-19 is an emergency public health problem of global importance. This study aimed to investigate the effect of foods and nutrients as complementary approaches on the recovery from COVID-19 in 170 countries, especially considering the complexity of the disease and the current scarcity of active treatments.

Methods: A retrospective study was performed using the Kaggle database, which links the consumption of various foods with recovery from COVID-19 in 170 countries, using multivariate analysis based on a generalized linear model.

Results: The results showed that certain foods had a positive effect on recovery from COVID-19: eggs, fish and seafood, fruits, meat, milk, starchy roots, stimulants, vegetable products, nuts, vegetable oil and vegetables. In general, consumption of higher levels of proteins and lipids had a positive effect on COVID-19 recovery, whereas high consumption of alcoholic beverages had a negative effect. In developed countries, where hunger had been eradicated, the effect of food on recovery from COVID-19 had a greater magnitude than in countries with a higher global hunger index (GHI), where there was almost no identifiable effect.

Conclusion: Several foods had a positive effect on COVID-19 recovery in developed countries, especially food groups with a higher content of lipids, proteins, antioxidants and micronutrients (e.g., selenium and zinc). In countries with extreme poverty (high GHI), foods presented little effect on recovery from COVID-19.

© 2021 Elsevier Ltd and European Society for Clinical Nutrition and Metabolism. All rights reserved.

1. Introduction

Coronavirus disease 2019 (COVID-19) was first reported last year in Wuhan, Hubei (China), and the etiologic agent was identified as a new coronavirus, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) [1]. The disease has been categorized as a

global pandemic by the World Health Organization (WHO), totaling 114,853,685 confirmed cases and 2,554,694 deaths worldwide by March 4, 2021 [2]. The typical clinical manifestations of COVID-19 include fever, dry cough and fatigue, often with pulmonary involvement. In severe cases of the disease, increases in the levels of cytokines (IL-2, IL-7, IL-10), granulocyte colony stimulating factor (G-CSF), monocyte chemotactic protein (MCP) and TNF- α occur. In this sense, the inflammatory cascade that triggers this cytokine storm appears to be a key factor in the cause of severe acute respiratory syndrome and extra-pulmonary organ failure, demonstrating the importance of the immune system in the progression of this disease [3,4].

In addition, evidence shows that there are certain population groups with higher vulnerability to the disease, especially patients

Abbreviations: COVID-19, Coronavirus disease 2019; ARS-CoV-2, severe acute respiratory syndrome coronavirus 2; FAO, Food and Agriculture Organization of the United Nations; IQR, interquartile range; GLM, generalized multivariate linear model.

* Corresponding author. Department of Pharmacy, Federal University of Parana, Av. Lothário Meissner, 632, 80210-170, Curitiba, Paraná, Brazil. Fax: +55 41 33604101.

E-mail address: pontarolo@ufpr.br (R. Pontarolo).

<https://doi.org/10.1016/j.clnu.2021.03.018>

0261-5614/© 2021 Elsevier Ltd and European Society for Clinical Nutrition and Metabolism. All rights reserved.

Risk factors associated with delay in diagnosis and mortality in patients with COVID-19 in the city of Rio de Janeiro, Brazil

Fatores de risco associados ao atraso no diagnóstico e mortalidade em pacientes com COVID-19 na cidade do Rio de Janeiro, Brasil

Alexandre de Fátima Cobre (<https://orcid.org/0000-0001-6642-3928>)¹
Beatriz Böger (<https://orcid.org/0000-0003-0025-2315>)¹
Mariana Millan Fachi (<https://orcid.org/0000-0001-5918-4738>)¹
Raquel de Oliveira Vilhena (<https://orcid.org/0000-0002-8942-0591>)¹
Eric Luiz Domingos (<https://orcid.org/0000-0001-8474-3984>)¹
Fernanda Stumpf Tonin (<https://orcid.org/0000-0003-4262-8608>)¹
Roberto Pontarolo (<https://orcid.org/0000-0002-7049-4363>)¹

Abstract We investigated the predictors of delay in the diagnosis and mortality of patients with COVID-19 in Rio de Janeiro, Brazil. A cohort of 3,656 patients were evaluated (Feb-Apr 2020) and patients' sociodemographic characteristics, and social development index (SDI) were used as determinant factors of diagnosis delays and mortality. Kaplan-Meier survival analyses, time-dependent Cox regression models, and multivariate logistic regression analyses were conducted. The median time from symptoms onset to diagnosis was eight days (interquartile range [IQR] 7.23-8.99 days). Half of the patients recovered during the evaluated period, and 8.3% died. Mortality rates were higher in men. Delays in diagnosis were associated with male gender ($p = 0.015$) and patients living in low SDI areas ($p < 0.001$). The age groups statistically associated with death were: 70-79 years, 80-89 years, and 90-99 years. Delays to diagnosis greater than eight days were also risk factors for death. Delays in diagnosis and risk factors for death from COVID-19 were associated with male gender, age under 60 years, and patients living in regions with lower SDI. Delays superior to eight days to diagnosis increased mortality rates.

Key words COVID-19, Multivariate analysis, Mortality

Resumo Investigamos os preditores de atraso no diagnóstico e mortalidade de pacientes com COVID-19 no Rio de Janeiro, Brasil. Uma coorte de 3.656 pacientes foi avaliada (fevereiro-abril de 2020) e as características sociodemográficas dos pacientes, o bairro e o índice de desenvolvimento social (IDS) foram usados como fatores determinantes dos atrasos no diagnóstico e da mortalidade. Foram realizadas análises de sobrevivência de Kaplan-Meier, modelos de regressão Cox dependentes do tempo e análises de regressão logística multivariada. O tempo mediano desde o início dos sintomas até o diagnóstico foi de oito dias (intervalo interquartil [IQR] 7,23-8,99 dias). Metade dos pacientes se recuperou no período avaliado e 8,3% faleceram. As taxas de mortalidade foram maiores nos homens. Atrasos no diagnóstico foram associados ao sexo masculino ($p = 0,015$) e pacientes que moravam em áreas com baixo IDS ($p < 0,001$). As faixas etárias estatisticamente associadas à morte foram: 70-79 anos, 80-89 anos e 90-99 anos. Atrasos no diagnóstico superiores a oito dias também foram fatores de risco para óbito. Atrasos no diagnóstico e fatores de risco para morte por COVID-19 foram associados ao sexo masculino, idade abaixo de 60 anos e pacientes que vivem em regiões com menor IDS. Atrasos superiores a oito dias no diagnóstico aumentam as taxas de mortalidade.

Palavras-chave COVID-19, Análise multivariada, Mortalidade

¹ Programa de Pós-Graduação em Ciências Farmacêuticas, Departamento de Farmácia, Universidade Federal do Paraná. Av. Lothário Meissner 632, Jardim Botânico. 80210-170 Curitiba PR Brasil. alexandrecobre@gmail.com



A multivariate analysis of risk factors associated with death by Covid-19 in the USA, Italy, Spain, and Germany

Alexandre de Fátima Cobre¹ · Beatriz Böger¹ · Raquel de Oliveira Vilhena² · Mariana Millan Fachi¹ · Josiane Marlei Muller Fernandes dos Santos¹ · Fernanda Stumpf Tonin¹

Received: 13 July 2020 / Accepted: 5 October 2020 / Published online: 19 October 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

Aim Our aim was to investigate the risk factors associated with death from COVID-19 in four countries: The USA, Italy, Spain, and Germany.

Subject and methods We used data from the Institute for Health Metrics and Evaluation with projection information from January–August 2020. A multivariate analysis of logistic regression was performed. The following factors were analyzed (per day): number of beds needed for the hospital services, number of intensive care units (ICU) beds required, number of ventilation devices, number of both hospital and ICU admissions due to COVID-19. Nagelkerke's R^2 coefficient of determination was used to evaluate the model's predictive ability. The quality of the model's fit was assessed by the Hosmer–Lemeshow and the chi-square tests.

Results Among the evaluated countries, Italy presented greater need for ICU beds/day (≤ 98 ; OR = 2315.122; CI 95% [334.767–16,503.502]; $p < 0.001$) and daily ventilation devices (≤ 118 ; OR = 1784.168; CI 95% [250.217–12,721.995]; $p < 0.001$). It is expected that both Italy and Spain have a higher ICU admission rate due to COVID-19 ($n = 14/\text{day}$). Spain will need more beds/day (≤ 357 ; OR = 146.838; CI 95% [113.242–190.402]; $p < 0.001$) and probably will have a higher number of daily hospital admissions ($n = 48/\text{day}$). All the above-mentioned factors have an important impact on patients' mortality due to COVID-19 in all four countries.

Conclusions Further investments in hospitals' infrastructure, as well as the development of innovative devices for patient's ventilation, are paramount to fight the pandemic in the USA, Italy, Spain, and Germany.

Keywords COVID-19 · Risk factors · USA · Italy · Spain · Germany

Introduction

The present study was carried out on April 19, 2020, then all the data used in this study refer to the period from the beginning of the COVID-19 pandemic to the month of April of this year. The World Health Organization recorded more than 4 million confirmed cases and 224 thousand deaths by the new Coronavirus disease 2019 (COVID-19) in more than 208

countries. The largest number of confirmed cases was reported by the United States of America (USA) ($n = 1,060,572$), Spain ($n = 212,917$), Italy ($n = 203,597$), France ($n = 128,442$), and Germany ($n = 161,187$) (Arabi et al. 2020; WHO 2020a). The exponential increase in the number of infected individuals, especially critically ill patients, is challenging public health systems worldwide. The infections caused by the SARS-COV-2 virus can be asymptomatic or cause mild

✉ Beatriz Böger
beatrizboger@gmail.com

Alexandre de Fátima Cobre
alexandrecobre@gmail.com

Raquel de Oliveira Vilhena
raquel.vilhena@hotmail.com

Mariana Millan Fachi
marianamfachi@gmail.com

Josiane Marlei Muller Fernandes dos Santos
josianemullerfernandes@gmail.com

Fernanda Stumpf Tonin
ffstonin@gmail.com

¹ Pharmaceutical Sciences Postgraduate Program, Federal University of Paraná, Curitiba, Brazil

² Department of Pharmacy, Pharmaceutical Sciences Postgraduate Program, Federal University of Paraná, Av. Lothário Meissner, 632, Paraná, Curitiba 80210-170, Brazil



Novel COVID-19 biomarkers identified through multi-omics data analysis: *N*-acetyl-4-*O*-acetylneuraminic acid, *N*-acetyl-L-alanine, *N*-acetyltryptophan, palmitoylcarnitine, and glycerol 1-myristate

Alexandre de Fátima Cobre¹ · Alexsander Couto Alves² · Ana Raquel Manuel Gotine³ ·
Karime Zeraik Abdalla Domingues¹ · Raul Edison Luna Lazo¹ · Luana Mota Ferreira⁴ ·
Fernanda Stumpf Tonin⁵ · Roberto Pontarolo⁴

Received: 1 November 2023 / Accepted: 16 January 2024
© The Author(s), under exclusive licence to Società Italiana di Medicina Interna (SIMI) 2024

Abstract

This study aims to apply machine learning models to identify new biomarkers associated with the early diagnosis and prognosis of SARS-CoV-2 infection. Plasma and serum samples from COVID-19 patients (mild, moderate, and severe), patients with other pneumonia (but with negative COVID-19 RT-PCR), and healthy volunteers (control) from hospitals in four different countries (China, Spain, France, and Italy) were analyzed by GC-MS, LC-MS, and NMR. Machine learning models (PCA and PLS-DA) were developed to predict the diagnosis and prognosis of COVID-19 and identify biomarkers associated with these outcomes. A total of 1410 patient samples were analyzed. The PLS-DA model presented a diagnostic and prognostic accuracy of around 95% of all analyzed data. A total of 23 biomarkers (e.g., spermidine, taurine, L-aspartic, L-glutamic, L-phenylalanine and xanthine, ornithine, and ribothimidine) have been identified as being associated with the diagnosis and prognosis of COVID-19. Additionally, we also identified for the first time five new biomarkers (*N*-Acetyl-4-*O*-acetylneuraminic acid, *N*-Acetyl-L-Alanine, *N*-Acetyltryptophan, palmitoylcarnitine, and glycerol 1-myristate) that are also associated with the severity and diagnosis of COVID-19. These five new biomarkers were elevated in severe COVID-19 patients compared to patients with mild disease or healthy volunteers. The PLS-DA model was able to predict the diagnosis and prognosis of COVID-19 around 95%. Additionally, our investigation pinpointed five novel potential biomarkers linked to the diagnosis and prognosis of COVID-19: *N*-Acetyl-4-*O*-acetylneuraminic acid, *N*-Acetyl-L-Alanine, *N*-Acetyltryptophan, palmitoylcarnitine, and glycerol 1-myristate. These biomarkers exhibited heightened levels in severe COVID-19 patients compared to those with mild COVID-19 or healthy volunteers.

Keywords COVID-19 · Diagnosis · Prognosis · Biomarker · Machine learning

✉ Roberto Pontarolo
pontarolo@ufpr.br

Alexandre de Fátima Cobre
alexandrecobre@gmail.com

Alexsander Couto Alves
a.coutoalves@surrey.ac.uk

Ana Raquel Manuel Gotine
anaraquelmanuel@gmail.com

Karime Zeraik Abdalla Domingues
karimezeraik@gmail.com

Raul Edison Luna Lazo
raulluna@ufpr.br

Luana Mota Ferreira
luanamota@ufpr.br

Fernanda Stumpf Tonin
stumpf.tonin@ufpr.br

- ¹ Universidade Federal do Paraná, Curitiba, Brazil
- ² School of Biosciences and Medicine, Faculty of Health and Medical Sciences, University of Surrey, Guildford, UK
- ³ Public Health College, Universidade de São Paulo, São Paulo, Brazil
- ⁴ Department of Pharmacy, Universidade Federal do Paraná, Campus III, Av. Pref. Lothário Meissner, 632, Jardim Botânico, Curitiba, PR 80210-170, Brazil
- ⁵ H&TRC - Health & Technology Research Centre, ESTeSL, Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Lisbon, Portugal



Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/complbiomed

Diagnosis and prediction of COVID-19 severity: can biochemical tests and machine learning be used as prognostic indicators?

Alexandre de Fátima Cobre^a, Dile Pontarolo Stremel^b, Guilhermina Rodrigues Noletto^c, Mariana Millan Fachi^a, Monica Surek^a, Astrid Wiens^d, Fernanda Stumpf Tonin^a, Roberto Pontarolo^{d,*}

^a Pharmaceutical Sciences Postgraduate Programme, Universidade Federal Do Paraná, Curitiba, Brazil

^b Department of Forest Engineering and Technology, Universidade Federal Do Paraná, Curitiba, Brazil

^c Department of Biochemistry, Universidade Federal Do Paraná, Curitiba, Brazil

^d Department of Pharmacy, Universidade Federal Do Paraná, Curitiba, Brazil

ARTICLE INFO

Keywords:

COVID-19

Diagnosis

Severity

Blood test

Urine test

Machine learning model

ABSTRACT

Objective: This study aimed to implement and evaluate machine learning based-models to predict COVID-19 diagnosis and disease severity.

Methods: COVID-19 test samples (positive or negative results) from patients who attended a single hospital were evaluated. Patients diagnosed with COVID-19 were categorised according to the severity of the disease. Data were submitted to exploratory analysis (principal component analysis, PCA) to detect outlier samples, recognise patterns, and identify important variables. Based on patients' laboratory tests results, machine learning models were implemented to predict disease positivity and severity. Artificial neural networks (ANN), decision trees (DT), partial least squares discriminant analysis (PLS-DA), and K nearest neighbour algorithm (KNN) models were used. The four models were validated based on the accuracy (area under the ROC curve).

Results: The first subset of data had 5,643 patient samples (5,086 negatives and 557 positives for COVID-19). The second subset included 557 COVID-19 positive patients. The ANN, DT, PLS-DA, and KNN models allowed the classification of negative and positive samples with >84% accuracy. It was also possible to classify patients with severe and non-severe disease with an accuracy >86%. The following were associated with the prediction of COVID-19 diagnosis and severity: hyperferritinaemia, hypocalcaemia, pulmonary hypoxia, hypoxemia, metabolic and respiratory acidosis, low urinary pH, and high levels of lactate dehydrogenase.

Conclusion: Our analysis shows that all the models could assist in the diagnosis and prediction of COVID-19 severity.

1. Introduction

Coronavirus disease (COVID-19) remains an emergency of global interest; up to 21 May 2021, a total of 164.52 million confirmed cases and 3.42 million deaths had accumulated from the disease [1]. Social disparity and the scarcity of hospital resources for the treatment of patients in hospital units have been identified among the main factors associated with an increased number of deaths from this disease [2–7]. Thus, it is essential to identify potential prognostic biomarkers towards earlier and more targeted care, especially considering that some patients

with COVID-19 develop severe disease, which is associated with a higher risk of hospitalisation. Biomarkers provide a dynamic and powerful approach to understanding the spectrum of disease with applications in observational and analytic epidemiology, randomised clinical trials, screening and diagnosis, and prognosis [8]. Recently, studies investigating biomarkers to diagnose COVID-19 in early stages have been encouraged worldwide, aiming to provide a faster referral to treatment and reducing health-related problems associated with the disease [17, 18].

Machine learning (ML) is an effective and innovative tool able to

* Corresponding author.

E-mail addresses: alexandreobre@gmail.com (A.F. Cobre), dile.stremel@gmail.com (D.P. Stremel), guilherminanoletto@ufpr.br (G.R. Noletto), marianamfachi@gmail.com (M.M. Fachi), monicasurek13@gmail.com (M. Surek), astridwiens@hotmail.com (A. Wiens), stumpf.tonin@ufpr.br (F.S. Tonin), pontarolo@ufpr.br (R. Pontarolo).

<https://doi.org/10.1016/j.complbiomed.2021.104531>

Received 9 March 2021; Received in revised form 21 May 2021; Accepted 25 May 2021

Available online 29 May 2021

0010-4825/© 2021 Elsevier Ltd. All rights reserved.



Naringenin-4'-glucuronide as a new drug candidate against the COVID-19 Omicron variant: a study based on molecular docking, molecular dynamics, MM/PBSA and MM/GBSA

Alexandre de Fátima Cobre^a , Moisés Maia Neto^b , Eduardo Borges de Melo^c, Mariana Millan Fachi^a , Luana Mota Ferreira^d , Fernanda Stumpf Tonin^e  and Roberto Pontarolo^d 

^aPharmaceutical Sciences Postgraduate Programme, Universidade Federal do Paraná, Curitiba, Brazil; ^bDepartment of Pharmacy, Fаметro University Centre (UNIFAMETRO), Fortaleza-Ceará, Brazil; ^cDepartment of Pharmacy, Universidade Estadual do Oeste do Paraná (UNIOESTE), Cascavel-PR, Brazil; ^dDepartment of Pharmacy, Universidade Federal do Paraná, Curitiba, Brazil; ^eH&TRC - Health & Technology Research Centre, ESTeSL, Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, Lisbon, Portugal

Communicated by Ramaswamy H. Sarma

ABSTRACT

This study aimed to identify natural bioactive compounds (NBCs) as potential inhibitors of the spike (S1) receptor binding domain (RBD) of the COVID-19 Omicron variant using computer simulations (*in silico*). NBCs with previously proven biological *in vitro* activity were obtained from the ZINC database and analyzed through virtual screening, molecular docking, molecular dynamics (MD), molecular mechanics/Poisson–Boltzmann surface area (MM/PBSA), and molecular mechanics/generalized Born surface area (MM/GBSA). Remdesivir was used as a reference drug in docking and MD calculations. A total of 170,906 compounds were analyzed. Molecular docking screening revealed the top four NBCs with a high affinity with the spike (affinity energy < -7 kcal/mol) to be ZINC000045789238, ZINC00004098448, ZINC00008662732, and ZINC00003995616. In the MD analysis, the four ligands formed a complex with the highest dynamic equilibrium S1 (mean RMSD < 0.3 nm), lowest fluctuation of the complex amino acid residues (RMSF < 1.3), and solvent accessibility stability. However, the ZINC000045789238-spike complex (naringenin-4'-O glucuronide) was the only one that simultaneously had minus signal (-) MM/PBSA and MM/GBSA binding free energy values (-3.74 kcal/mol and -15.65 kcal/mol, respectively), indicating favorable binding. This ligand (naringenin-4'-O glucuronide) was also the one that produced the highest number of hydrogen bonds in the entire dynamic period (average = 4601 bonds per nanosecond). Six mutant amino acid residues formed these hydrogen bonds from the RBD region of S1 in the Omicron variant: Asn417, Ser494, Ser496, Arg403, Arg408, and His505. Naringenin-4'-O-glucuronide showed promising results as a potential drug candidate against COVID-19. *In vitro* and preclinical studies are needed to confirm these findings.

Abbreviations: ACE-2: angiotensin converting enzyme 2; CADD: computer-aided drug discovery; CHARMM: chemistry at Harvard macromolecular mechanics; COVID-19: coronavirus disease 2019; GROMACS: groningen machine for chemical simulations; HIV: human immunodeficiency virus; LINCOS: linear constraint solver; MD: molecular dynamics; MM/GBSA: molecular mechanics/generalised Born surface area; MM/PBSA: molecular mechanics/Poisson–Boltzmann surface area; NBCs: natural bioactive compounds; NVE: microcanonical; NPT: isothermal-isobaric; NVT: canonical; PDB: protein data bank; PDBQT: protein data bank partial charge and atom type format; QSAR: quantitative structure-activity relationship; RBD: receptor binding domain; RCSB: research collaborative for structural bioinformatics; Rg: radius of gyration; RMSD: root mean square deviation; RMSF: root mean square fluctuation; SASA: solvent accessible surface area; S1: subunit 1 spike glycoprotein; SARS-CoV-2: Severe acute respiratory syndrome coronavirus 2; VMD: visual molecular dynamics; WHO: world health organisation

ARTICLE HISTORY

Received 12 January 2023
Accepted 19 June 2023

KEYWORDS

Naringenin-4'-glucuronide;
SARS-CoV-2; treatment;
spike protein; *in silico*

1. Introduction

Two years into the global COVID-19 pandemic, with over 591 million infections and more than 6.5 million confirmed deaths by August 2022 (WHO, 2022), the focus of attention has shifted to the emergence and spread of new variants of SARS-CoV-2 that have now been associated with substantial

detrimental effects on the transmission and severity of the virus (Liu et al., 2022).

The Omicron variant (B.1.1.529), initially detected in Southern Africa in November 2021, became a dominant strain of concern, given its high transmissibility (Liu et al., 2022; Mannar et al., 2022; WHO, 2021). Due to the significant accumulation of mutations in the receptor binding domain

CONTACT Roberto Pontarolo  pontarolo@ufpr.br  Department of Pharmacy, Universidade Federal do Paraná, Campus III, Av. Prof. Lothário Meissner, 632, Jardim Botânico, Curitiba, PR 80210-170, Brazil

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/07391102.2023.2229446>.

© 2023 Informa UK Limited, trading as Taylor & Francis Group

MACHINE LEARNING-BASED VIRTUAL SCREENING, MOLECULAR DOCKING, DRUG-LIKENESS, PHARMACOKINETICS AND TOXICITY ANALYSES TO IDENTIFY NEW NATURAL INHIBITORS OF THE GLYCOPROTEIN SPIKE (S1) OF SARS-CoV-2**Alexandre de F. Cobre^a, Beatriz Böger^a, Mariana M. Fachi^a, Carlos A. Ehrenfried^a, Dile P. Stremel^b, Eduardo B. De Melo^c, Fernanda S. Tonin^d and Roberto Pontarolo^a**^aDepartamento de Farmácia, Universidade Federal do Paraná, 80210-170 Curitiba – PR, Brasil^bDepartamento de Engenharia e Tecnologia Florestal, Universidade Federal do Paraná, 80210-170 Curitiba – PR, Brasil^cDepartamento de Farmácia, Universidade Estadual do Oeste do Paraná, 85819-110 Cascavel – PR, Brasil^dH&TRC- Health & Technology Research Center, ESTeSL, Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, 1990-096 Lisbon, Portugal

Recebido em 14/09/2022; aceito em 26/01/2023; publicado na web 31/03/2023

To identify natural bioactive compounds (NBCs) as potential inhibitors of spike (S1) by means of *in silico* assays. NBCs with previously proven biological *in vitro* activity were obtained from the ZINC database and analyzed through virtual screening and molecular docking to identify those with higher affinity to the spike protein. Eight machine learning models were used to validate the results: Principal Component Analysis (PCA), Artificial Neural Network (ANN), Support Vector Machine (SVM), k-Nearest Neighbors (KNN), Partial Least Squares-Discriminant Analysis (PLS-DA), Gradient Boosted Tree Discriminant Analysis (XGBoostDA), Soft Independent Modelling of Class Analogies (SIMCA) and Logistic Regression Discriminate Analysis (LREG). Selected NBCs were submitted to drug-likeness prediction using Lipinski's and Veber's rule of five. A prediction of pharmacokinetic parameters and toxicity was also performed (ADMET). Antivirals currently used for COVID-19 (remdesivir and molnupiravir) were used as a comparator. A total of 170,906 compounds were analyzed. Of these, 34 showed greater affinity with the S1 (affinity energy < -7 kcal mol⁻¹). Most of these compounds belonged to the class of coumarins (benzopyrones), presenting a benzene ring fused to a lactone (group of heterosides). The PLS-DA model was able to reproduce the results of the virtual screening and molecular docking (accuracy of 97.0%). Of the 34 compounds, only NBC5 (feselol), NBC14, NBC15 and NBC27 had better results in ADMET predictions. These had similar binding affinity to S1 when compared to remdesivir and molnupiravir. Feselol and three other NBCs were the most promising candidates for treating COVID-19. *In vitro* and *in vivo* studies are needed to confirm these findings.

Keywords: *in silico*; COVID-19; spike glycoprotein; treatment.**INTRODUCTION**

Several attempts have been made to manage the coronavirus disease 2019 (COVID-19) pandemic, caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), including the recent vaccination programs rolled out worldwide.¹ However, there is still a need to identify effective treatments, considering that to this day there is not even a drug with proven efficacy data. Current treatments are limited to treating symptoms only, they are palliative treatments.²⁻⁸

The scarce knowledge of the pathogenesis and immunological peculiarity of SARS-CoV-2, especially regarding the interaction between viral antigens and human receptors and the triggering of cytokine storms, poses additional challenges for the development of successful treatments.^{5,9} Although details of the cellular responses to this virus are unknown, a probable course of events can be postulated based on past studies with SARS-CoV-2.^{5,9,10} Infections are initiated by the virus binding to the angiotensin-converting enzyme receptor-2 (ACE-2) cell-surface receptors, which is followed by fusion of the virus and cell membranes to release the virus RNA genome into the host cell through receptor-mediated endocytosis.^{5,6} Both receptor binding and membrane fusion activities are mediated by the 'spike glycoprotein' of the virus.¹⁰ As with other class-I membrane-fusion proteins (alpha-helical), the spike protein is post-translationally cleaved, in this case by furin, into the S1 and S2 components that

remain associated after cleavage. Each S1 component consists of two large domains, the N-terminal domain (NTD) and the receptor-binding domain (RBD).¹¹ The interaction between viral antigens and host immune cells finally results in the induction of pro-inflammatory responses that trigger vasodilation, increased vascular permeability and the accumulation of humoral factors, causing fever and interrupting gas exchange (*i.e.*, respiratory distress).⁹

Given the global emergency caused by COVID-19, there is great interest in drug repurposing (*i.e.*, drug repositioning or rediscovery) to accelerate the identification of drugs that can cure or prevent this disease.^{1,8} One of the key drivers for the repositioning of drugs is the serendipitous discovery of pharmacological activity on new targets, which would then suggest a possible new indication of use. High-throughput screening of potential compounds available in databases is an emerging strategy that has already supported the discovering of new indications for marketed drugs (*e.g.*, lopinavir/ritonavir for HIV) and the development of additional therapeutic options against Ebola, hepatitis C and Zika virus infection.^{10,12}

To accelerate the drug discovery process, several open source *in silico* platforms are available in the literature for the prediction of pharmacokinetic parameters (absorption, distribution, metabolism and elimination), toxicity and drug-likeness. These platforms were built using machine learning models. For example, the SwissADME *in silico* platform, built using the Support Vector Machine (SVM) machine learning algorithm, allows for drug-likeness predictions (for example, Lipinski's rule), and pharmacokinetic parameters, with an

*e-mail: beatrizbogger@gmail.com



Systematic review and evidence gap mapping of biomarkers associated with neurological manifestations in patients with COVID-19

K. Z. A. Domingues¹ · A. F. Cobre¹ · R. E. L. Lazo¹ · L. S. Amaral¹ · L. M. Ferreira¹ · F. S. Tonin² · R. Pontarolo¹

Received: 11 August 2023 / Revised: 27 October 2023 / Accepted: 29 October 2023 / Published online: 28 November 2023
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany 2023

Abstract

Objective This study aimed to synthesize the existing evidence on biomarkers related to coronavirus disease 2019 (COVID-19) patients who presented neurological events.

Methods A systematic review of observational studies (any design) following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines and the Cochrane Collaboration recommendations was performed (PROSPERO: CRD42021266995). Searches were conducted in PubMed and Scopus (updated April 2023). The methodological quality of nonrandomized studies was assessed using the Newcastle–Ottawa Scale (NOS). An evidence gap map was built considering the reported biomarkers and NOS results.

Results Nine specific markers of glial activation and neuronal injury were mapped from 35 studies published between 2020 and 2023. A total of 2,237 adult patients were evaluated in the included studies, especially during the acute phase of COVID-19. Neurofilament light chain (NfL) and glial fibrillary acidic protein (GFAP) biomarkers were the most frequently assessed ($n=27$ studies, 77%, and $n=14$ studies, 40%, respectively). Although these biomarkers were found to be correlated with disease severity and worse outcomes in the acute phase in several studies ($p<0.05$), they were not necessarily associated with neurological events. Overall, 12 studies (34%) were judged as having low methodological quality, 9 (26%) had moderate quality, and 9 (26%) had high quality.

Conclusions Different neurological biomarkers in neurosymptomatic COVID-19 patients were identified in observational studies. Although the evidence is still scarce and conflicting for some biomarkers, well-designed longitudinal studies should further explore the pathophysiological role of NfL, GFAP, and tau protein and their potential use for COVID-19 diagnosis and management.

Keywords SARS-CoV-2 · Neurological · Biomarker · Neurofilament light chain · Tau protein

Abbreviations

A β	Amyloid beta	NfL	Neurofilament light chain
BBB	Blood–brain barrier	NSE	Neuron-specific enolase
CNS	Central nervous system	PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
GFAP	Glial fibrillary acidic protein	PNS	Peripheral nervous system
NfH	Neurofilament heavy chain	S100B	S100 calcium-binding protein B
		sTREM2	Soluble triggering receptor expressed on myeloid cells 2
		UCH-L1	Ubiquitin C-terminal hydrolase L1

K. Z. A. Domingues, A. F. Cobre have contributed equally to this work.

✉ R. Pontarolo
pontarolo@ufpr.br

¹ Programa de Pós-Graduação em Ciências Farmacêuticas, Universidade Federal do Paraná, Curitiba, PR 80210-170, Brazil

² H&TRC- Health & Technology Research Center, ESTeSL, Escola Superior de Tecnologia da Saúde, Instituto Politécnico de Lisboa, 1990-096 Lisbon, Portugal

Introduction

SARS-CoV-2, the virus responsible for causing COVID-19, can invade human cells through the interaction between the Spike (S) protein and the angiotensin-converting enzyme 2 receptor, which is expressed in different organs, including



ELSEVIER

Contents lists available at ScienceDirect

American Journal of Infection Control

journal homepage: www.ajicjournal.org

Major Article

Systematic review with meta-analysis of the accuracy of diagnostic tests for COVID-19



Beatriz Böger MSc^a, Mariana M. Fachi MSc^a, Raquel O. Vilhena PhD^b, Alexandre F. Cobre MSc^a,
Fernanda S. Tonin PhD^a, Roberto Pontarolo PhD^{b,*}

^a Pharmaceutical Sciences Postgraduate Program, Health Sciences Sector, Federal University of Paraná, Curitiba, Brazil

^b Department of Pharmacy, Federal University of Paraná, Curitiba, Brazil

Key Words:
SARS-CoV-2
Coronavirus
Evidence
Sensitivity
Specificity

Objective: To collate the evidence on the accuracy parameters of all available diagnostic methods for detecting SARS-CoV-2.

Methods: A systematic review with meta-analysis was performed. Searches were conducted in Pubmed and Scopus (April 2020). Studies reporting data on sensitivity or specificity of diagnostic tests for COVID-19 using any human biological sample were included.

Results: Sixteen studies were evaluated. Meta-analysis showed that computed tomography has high sensitivity (91.9% [89.8%–93.7%]), but low specificity (25.1% [21.0%–29.5%]). The combination of IgM and IgG antibodies demonstrated promising results for both parameters (84.5% [82.2%–86.6%]; 91.6% [86.0%–95.4%], respectively). For RT-PCR tests, rectal stools/swab, urine, and plasma were less sensitive while sputum (97.2% [90.3%–99.7%]) presented higher sensitivity for detecting the virus.

Conclusions: RT-PCR remains the gold standard for the diagnosis of COVID-19 in sputum samples. However, the combination of different diagnostic tests is highly recommended to achieve adequate sensitivity and specificity.

© 2020 Association for Professionals in Infection Control and Epidemiology, Inc. Published by Elsevier Inc. All rights reserved.

INTRODUCTION

After the first case reports of an acute respiratory syndrome of unknown etiology in the city of Wuhan, Hubei province (December 31, 2019), Chinese authorities identified a new coronavirus (SARS-CoV-2) that causes the clinical disease COVID-19. The virus outbreak spread quickly, significantly affecting all continents with more than 2 million people infected and thousands of deaths.^{1,2} Consequently, nations are facing the overwhelming of health care systems and both psychological and economic burdens. The lack of effective treatments or prevention strategies has contributed toward the increase in the number of cases, enhancing health care expenses with hospitalizations and palliative therapies. Additionally, there are limited diagnostic tests available, which favors the growth of under-reporting of cases.^{2,3}

Patients report fever and cough, and most develop chest discomfort, difficulty in breathing or pneumonia, being clinically diagnosed by imaging tests such as chest X-ray or computed tomography (CT). CT equipment is widespread worldwide and the scan process is relatively simple and quick, which enables rapid screening for suspected patients. The typical findings of chest CT images for individuals with COVID-19 are multifocal bilateral patchy ground-glass opacities or consolidation with interlobular septal and vascular thickening in the peripheral areas of the lungs. However, CT findings can change as the disease progresses and these manifestations may also be compatible with other viral pneumonias.^{4,5}

In this context, the current gold standard for diagnosing COVID-19 is based on a molecular test of the reverse transcription polymerase chain reaction (RT-PCR), aimed at detecting the RNA of the virus in respiratory samples such as nasopharyngeal swabs or bronchial aspirate.⁶ The real-time RT-PCR test provides a sensitive (the ability of the test to correctly identify those patients with the disease^{7,8}) and specific (the ability of the test to correctly identify those patients without the disease⁸) method to detect SARS-COV-2, with different diagnosis protocols including sequences of target primers available in the World Health Organization public database.^{6,9} However,

* Address correspondence to Roberto Pontarolo, PhD, Department of Pharmacy, Federal University of Paraná, Av. Lothário Meissner, 632, 80210-170, Curitiba, Paraná, Brazil.

E-mail address: pontarolo@ufpr.br (R. Pontarolo).

Conflict of interest: None to report.