Universidade Estadual Paulista “Júlio de Mesquita Filho” – UNESP

Rafael Vieira

Utilização de inteligência artificial e quimioinformática no desenvolvimento de ferramentas computacionais para o estudo da microdiversidade molecular em exsudatos da fermentação de cacau (*Theobroma cacao* L.)

Araraquara, 2023

Rafael Vieira

Utilização de inteligência artificial e quimioinformática no desenvolvimento de ferramentas computacionais para o estudo da microdiversidade molecular em exsudatos da fermentação de cacau (*Theobroma cacao* L.)

Tese de Doutorado apresentada ao Instituto de Química, Universidade Estadual Paulista, como parte dos requisitos para obtenção do título de doutor em Química.

Prof. Dr. Ian Castro-Gamboa

Araraquara, 2023

V658u

Vieira, Rafael

Utilização de inteligência artificial e quimioinformática no desenvolvimento de ferramentas computacionais para o estudo da microdiversidade molecular em exsudatos da fermentação de cacau (Theobroma cacao L.) / Rafael Vieira. -- Araraquara, 2023

309 p.

Tese (doutorado) - Universidade Estadual Paulista (Unesp), Instituto de Química, Araraquara

Orientador: Ian Castro Gamboa

Coorientadora: Kally Alves de Sousa

1. Aprendizado de máquinas. 2. Cacau. 3. Redes neurais (Computação). 4. Dinâmica molecular. 5. Produtos naturais. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca do Instituto de Química, Araraquara. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

IMPACTO POTENCIAL DA PESQUISA

Os resultados alcançados nesta pesquisa de doutorado têm potencial em inserir uma série de repercussões importantes no campo da química de produtos naturais e nas ciências ômicas contemporâneas, tais como:

- **Identificação de novos compostos terapêuticos:** A identificação de 10 moléculas distintas em diferentes fases do processo de fermentação do cacau, com potencial atividade contra alvos biomacromoleculares associados a doenças respiratórias, é um avanço significativo. Isso poderia levar ao desenvolvimento de novos tratamentos para doenças como asma e covid-19, que são grandes desafios de saúde global.
- **Automatização e eficiência:** A oferta de produtos em forma de web plataformas inteligentes que fazem uso de inteligência artificial e técnicas computacionais avançadas, como o LUMIOS, Chemistika e CHEIC, auxiliam na automatização do processamento de dados espectrais e na análise de dinâmica molecular. Ferramentas que permitem não apenas economizar tempo e esforço, mas também minimizar o risco de erros humanos, permitindo que os pesquisadores se concentrem em interpretações e aplicações mais avançadas dos dados.
- **Interdisciplinaridade:** A integração de técnicas de química de produtos naturais, quimioinformática, estatística, docagem e dinâmica molecular na análise de matrizes complexas demonstra a natureza interdisciplinar da pesquisa. Contribuindo de maneira fundamental para as ciências ômicas do século XXI, onde a complexidade dos sistemas biológicos (e das matrizes complexas) requerem uma abordagem integrada.
- **Aplicabilidade:** As ferramentas desenvolvidas são altamente aplicáveis a outras áreas de pesquisa em ciências ômicas e química de produtos naturais. Isso poderia facilitar a descoberta de novos compostos bioativos em outras matrizes biológicas complexas, acelerando o desenvolvimento de novos fármacos e tratamentos.

- **Democratização do acesso:** Ao tornar as plataformas gratuitas, permite-se que pesquisadores, independentemente de seus recursos financeiros ou institucionais, tenham acesso a ferramentas avançadas de análise. Isso é especialmente importante para pesquisadores em países em desenvolvimento ou instituições com recursos limitados.
- **Promoção da colaboração:** A disponibilidade de plataformas gratuitas desenvolvidas durante a elaboração desta tese pode facilitar a colaboração entre pesquisadores de diferentes áreas e instituições, uma vez que todos podem acessar e utilizar as mesmas ferramentas. Isso pode levar a uma maior integração e cooperação na comunidade acadêmica.
- **Contribuição para a economia local:** A exploração da cadeia produtiva, por meio dos processos fermentativos, como os do cacau, para a identificação de moléculas com atividade terapêutica, é um exemplo de economia circular. Isso não apenas contribui para a saúde humana, mas também para a sustentabilidade ambiental e econômica de regiões menos favorecidas, como os estados do centro-norte do país.
- **Aceleração da pesquisa:** As ferramentas computacionais podem processar grandes volumes de dados muito mais rapidamente do que seria possível manualmente. Ao disponibilizar essas ferramentas gratuitamente, pode-se acelerar significativamente o progresso da pesquisa em várias áreas.
- **Inovação:** Ao disponibilizar tais plataformas de maneira gratuita, cria-se um ambiente propício para a inovação, pois permite que pesquisadores de diferentes áreas e perspectivas tenham acesso e contribuam para o desenvolvimento e aprimoramento das ferramentas.
- **Educação:** Plataformas gratuitas são recursos valiosos para a educação e treinamento de estudantes e jovens pesquisadores, o quais podem aprender e praticar novas técnicas e métodos de análise sem a barreira de custos associados ao software proprietário.
- **Transparência e reprodutibilidade:** A disponibilidade de plataformas gratuitas pode ajudar a aumentar a transparência e a reprodutibilidade na

pesquisa, pois permite que outros pesquisadores testem e validem os resultados uns dos outros usando as mesmas ferramentas.

Portanto, esta pesquisa não apenas contribui para a química de produtos naturais, fornecendo novos compostos potencialmente terapêuticos e ferramentas para sua identificação e análise, mas também para as ciências ômicas do século XXI, fornecendo uma abordagem interdisciplinar e integrada para a análise de sistemas biológicos complexos, promovendo também a colaboração e a inovação, com intuito de acelerar a pesquisa, melhorar a educação, e aumentar a transparência e a reprodutibilidade na pesquisa.

POTENTIAL IMPACT OF RESEARCH

The results achieved in this doctoral research have the potential to bring about a series of significant repercussions in the field of natural product chemistry and contemporary omic sciences, such as:

• **Identification of new therapeutic compounds:** The identification of 10 distinct molecules at different stages of cocoa fermentation, with potential activity against biomacromolecular targets associated with respiratory diseases, is a significant advancement. This could lead to the development of new treatments for diseases such as asthma and covid-19, which are major global health challenges.

• **Automation and efficiency:** The provision of products in the form of smart web platforms that use artificial intelligence and advanced computational techniques, like LUMIOS, Chemistika, and CHEIC, assist in the automation of spectral data processing and molecular dynamics analysis. Tools that not only save time and effort but also minimize the risk of human errors, allowing researchers to focus on more advanced data interpretations and applications.

• **Interdisciplinarity:** The integration of natural product chemistry techniques, cheminformatics, statistics, docking, and molecular dynamics in the analysis of complex matrices demonstrates the interdisciplinary nature of this research. Contributing fundamentally to the 21st-century omic sciences, where the complexity of biological systems (and of complex matrices) requires an integrated approach.

• **Applicability:** The tools developed are highly applicable to other areas of research in omic sciences and natural product chemistry. This could facilitate the discovery of new bioactive compounds in other complex biological matrices, speeding up the development of new drugs and treatments.

• **Democratizing access:** By making platforms available for free, it allows researchers, regardless of their financial or institutional resources, to access advanced analysis tools. This is especially important for researchers in developing countries or institutions with limited resources.

• **Promoting collaboration:** The availability of free platforms developed during this thesis can facilitate collaboration between researchers from different areas and institutions, as everyone can access and use the same tools. This can lead to greater integration and cooperation in the academic community.

• **Contribution to the local economy:** Exploring the production chain through fermentative processes, such as cocoa's, for identifying molecules with therapeutic activity is an example of a circular economy. This not only contributes to human health but also to the environmental and economic sustainability of less favored regions, like the north-central states of the country.

• **Accelerating research:** Computational tools can process large volumes of data much faster than manually possible. By making these tools available for free, research progress in various areas can be significantly accelerated.

• **Innovation:** By offering such platforms for free, it creates an environment conducive to innovation, as it allows researchers from different fields and perspectives to access and contribute to the development and enhancement of the tools.

• **Education:** Free platforms are valuable resources for educating and training students and young researchers, who can learn and practice new techniques and methods of analysis without the cost barrier associated with proprietary software.

• **Transparency and reproducibility:** The availability of free platforms can help increase transparency and reproducibility in research, as it allows other researchers to test and validate each other's results using the same tools.

Therefore, this research not only contributes to natural product chemistry by providing potentially therapeutic new compounds and tools for their identification and analysis, but also to the 21st-century omic sciences by providing an interdisciplinary and integrated approach to the analysis of complex biological systems, further promoting collaboration and innovation with the intent of accelerating research, improving education, and increasing transparency and reproducibility in research.

UNIVERSIDADE ESTADUAL PAULISTA

Câmpus de Araraquara

CERTIFICADO DE APROVAÇÃO

TÍTULO DA TESE: "Utilização de inteligência artificial e quimioinformática no desenvolvimento de ferramentas computacionais para o estudo da microdiversidade molecular em exsudatos da fermentação de cacau (*Theobroma cacao* L.)"

AUTOR: RAFAEL VIEIRA
ORIENTADOR: IAN CASTRO GAMBOA
COORIENTADORA: KALLY ALVES DE SOUSA

Aprovado como parte das exigências para obtenção do Título de Doutor em Química, pela Comissão Examinadora:

Prof. Dr. IAN CASTRO GAMBOA (Participaçao Presencial)
Departamento de Bioquimica e Quimica Organica / Instituto de Quimica UNESP Araraquara

Prof. Dr. RICARDO ROBERTO DA SILVA (Participaçao Presencial)
Departamento de Física e Química / Faculdade de Ciências Farmacêuticas de Ribeirão Preto - USP - Ribeirão Preto

Dr. LUCIANO DA SILVA PINTO (Participaçao Presencial)
Departamento de Química / Centro de Ciências Exatas e de Tecnologia - UFSCAR - São Carlos

Profa. Dra. ERICA REGINA FILLETTI NASCIMENTO (Participaçao Presencial)
Departamento de Engenharia, Física e Matemática / Instituto de Química - UNESP - Araraquara

Profa. Dra. KELLY JOHANA DUSSAN MEDINA (Participaçao Presencial)
Departamento de Engenharia, Fisica e Matematica / Instituto de Quimica - UNESP - Araraquara

Araraquara, 11 de outubro de 2023

IDENTIFICAÇÃO

Nome: Rafael Vieira

Site: www.vieira-rafael.com

Nome em citações bibliográficas: VIEIRA, R.

ORCID: https://orcid.org/0000-0001-9003-3209

Endereço profissional: Rua Prof. Francisco Degni, nº 55, Bairro Quitandinha, Araraquara-SP, CEP: 14800-060

FORMAÇÃO ACADÊMICA/TITULAÇÃO

2019 – 2023 – Doutorado em Química:

Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, São Paulo, Brasil

Título: Relações moleculares da microdiversidade presente em exsudatos da fermentação de cacau (*Theobroma cacao L.*) utilizando *machine learning,* quimioinformática e técnicas de desreplicação.

Orientador: Ian Castro-Gamboa

2014 – 2015 – Mestrado em Química

Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, São Paulo, Brasil

Título: Título: Exploração racional da rede metabólica de *Xylaria* sp. visando a produção de metabólitos de interesse farmacológico, através de ferramentas quimiométricas e técnicas de desreplicação, **Ano de obtenção**: 2015

Orientador: Ian Castro-Gamboa

2008 – 2012 – Graduação em Química

Universidade Estadual Paulista Júlio de Mesquita Filho, UNESP, São Paulo, Brasil

FORMAÇÃO COMPLEMENTAR

2021 – 2023:

Especialização em Ciência de Dados e Inteligência Artificial

Instituto Federal de Educação, Ciência e Tecnologia de São Paulo – Campus Campinas

Orientador: Prof. Dr. Samuel Botter Martins

2021 – Curso de Curta Duração (virtual):

Amazon Web Service – AWS ACADEMY MACHINE LEARNING (20 horas)

Amazon Web Service – AWS ACADEMY CLOUD FOUNDATIONS (20 horas)

ATUAÇÃO PROFISSIONAL

Período: 16/01/2016 – 31/06/2017

Profissão: Professor e coordenador de curso de Química no Centro Universitário FAEMA – Ariquemes – RO

Período: 17/07/2017 – 31/11/2017

Profissão: Professor efetivo no Instituto Federal de Educação, Ciência e Tecnologia do ACRE (IFAC), Câmpus Tarauacá.

Período: 29/12/2017 – atual

Profissão: Professor efetivo no Instituto Federal de Educação, Ciência e Tecnologia de Rondônia (IFRO), Campus Ji-Paraná.

Coordenador de Pós-Graduação vinculado ao Departamento de Pesquisa, Inovação e Pós-Graduação (DEPESP) do *campus* Guajará-Mirim – IFRO (**PORTARIA Nº 29/GJM - CGAB/IFRO, DE 08 DE FEVEREIRO DE 2018)**

PRODUÇÃO BIBLIOGRÁFICA

ARTIGOS COMPLETOS PUBLICADOS EM PERIÓDICO:

1. **VIEIRA, RAFAEL**; ALVES DE SOUSA, KALLY; SOUZA DA SILVA, GIVALDO ; HELENA SIQUEIRA SILVA, DULCE ; CASTRO-GAMBOA, IAN . CHEIC: CHEMICAL IMAGE CLASSIFICATOR An intelligent system for identification of volatiles compounds with potential for respiratory diseases using Deep Learning. **EXPERT SYSTEMS WITH APPLICATIONS**. Fator de Impacto (JCR 2022: 8,5), v. 234, p. 121178, **2023**.

2. **VIEIRA, RAFAEL**; DE SOUSA, KALLY ALVES; MONTEIRO, AFIF FELIX; PINTO, LUCIANO SILVA; CASTRO-GAMBOA, IAN. Induction of metabolic variability of the endophytic fungus *Xylaria* sp. by OSMAC approach and experimental design. **ARCHIVES OF MICROBIOLOGY**. Fator de Impacto (JCR 2021: 2,6670, 203, 3025-3022, **2021**.

3. **VIEIRA, R.;** SOUZA, J. M.; SILVA, G. S.; SOUSA, C. O. Identification of Tannins in Amazon Biodiversity Plants: Application Possibilities as a Natural Coagulant. **JOURNAL OF APPLIED OF PHARMACEUTICAL SCIENCE.** , v.5, p.17 - 23, **2018**.

4. HONORATO DE JESUS, JOCIEL; DO CARMO SILVA DE OLIVEIRA, MARIA; MARIA MINETTO BRONDANI, FILOMENA; ROSSI OLIVEIRA LIMA, REGIANE; **VIEIRA, RAFAEL.** PROPRIEDADES FÍSICO-QUÍMICAS DO AMIDO DO CARÁ (*Dioscorea cayennensis*) NATIVO E MODIFICADO POR ACETILAÇÃO. The Journal of Engineering and Exact Sciences. , v.4, p.0429 - 0436, **2018**.

APRESENTAÇÃO DE TRABALHOS

2022 – (Cartagena das Índias – Colômbia):

Apresentação de Poster/Painel no **IV-LAMPS – Latin American Metabolic Profiling Society:** Exploratory analysis of complex matrices of cupuaçu fermented (Theobroma grandiflorum (Willd. ex Spreng.) Schum.): the use of cheminformatics for the bioprospecting of molecules of high added value.

2022 – (Apresentação de palestra (WEBINAR) internacional) organizado pela IV-LAMPS – Latin American Metabolic Profiling Society

Molecular relationship of microdiversity present in cocoa fermentation exudates (*Theobroma cacao* L.) using *Machine Learning* and *Big Data* Technologies

PROJETOS DE PESQUISA ENVOLVIDO (PERÍODO DO DOUTORADO)

2019 – Atual (Pesquisador Responsável) – Projeto Universal – Áreas prioritárias (EDITAL Nº 6/2021/FAPERO-DC)

Análise exploratória das matrizes complexas de fermentados de cupuaçu (Theobroma grandiflorum (Willd. ex Spreng.) Schum.): o uso da quimioinformática para a bioprospecção de moléculas de alto valor agregado

2018 – 2019 (Projetos como Co-orientador):

- Desreplicação metabólica de *Fusarium* sp. e *Aspergillus* sp. cultivados em farelo de cacau (*Theobroma cacao* L.) (Edital no 14/2018/REIT-PROPESP/IFRO, de 10 de maio de 2018) (SEI no 0318471);
- Exploração racional do fermentado de cupuaçu (*Theobroma grandiflorum*) visando a identificação de metabólitos secundários de interesse comercial através de desreplicação (Edital no 14/2018/REIT-PROPESP/IFRO, de 10 de maio de 2018) (SEI no 0318471);
- Exploração racional da espécie *Himatanthus sucuuba* (Spruce) Woodson visando a identificação de metabólitos secundários de interesse comercial (Edital no 48/2018/GJM-CGAB/IFRO, de 22 de junho de 2018) (SEI no 0335263).

PARTICIPAÇÃO EM BANCAS (Últimos 5 anos)

Participação em bancas de curso de Mestrado – 1

Participação em banca de Tiago Teodoro de Lima Souza. Desreplicação do processo fermentativo espontâneo de cacau (*Theobroma cacao* L.) para identificação de bioativos de interesse comercial, **2022**.

Participação em bancas de trabalho de Conclusão de Curso – 4

1 Participação em banca de DANDARA DA SILVA PEREIRA. Avaliação da qualidade físico-química e microbiológica da água consumida no Instituto Federal de Rondônia campus Ji-Paraná, **2019**. (Química) Instituto Federal de Educação Ciência e Tecnologia de Rondônia

2 Participação em banca de STEPHANIE JEDOZ STEIN. Índice de balneabilidade no Rio Machado na área urbana do município de Ji-Paraná - Rondônia**, 2019**. (Química) Instituto Federal de Educação Ciência e Tecnologia de Rondônia

3 Participação em banca de FABIANA DE OLIVEIRA DA SILVA. O uso de Achachairu (*Garcinia humilis*) como Indicador Ácido-Base Natural, **2019** (Química) Instituto Federal de Educação Ciência e Tecnologia de Rondônia.

4 Participação em banca de Danilo Moura Santos. O uso da tecnologia 3D no ensino de química através da confecção de um dominó sobre funções orgânicas, **2018**. (Química) Instituto Federal de Educação Ciência e Tecnologia de Rondônia

ORIENTAÇÕES

Mayara Pacheco Figueiredo. Perfil metabólico da fermentação da amêndoa de cacau (*Theobroma cacao*) utilizando quimiometria e técnicas cromatográficas. **2019**. Curso (Química) - Instituto Federal de Educação Ciência e Tecnologia de Rondônia

PARTICIPAÇÃO EM BANCAS E ORIENTAÇÕES NA CARREIRA:

Orientações concluídas (TCC) – **15**

Participação em banca de trabalhos de conclusão (mestrado) – **1**

Participação em banca (curso de aperfeiçoamento/especialização) – **6**

Participação em banca de trabalhos de conclusão (graduação) – **11**

DEDICATÓRIA

Os filhos que criou trilharam seus caminhos (como dizem que deve ser) ...

Ele não se conformou e andou em nossos calcanhares. Lançou-se em mil direções e disse que no final daquele ano (2019) iria para Rondônia e voltaríamos de carro para Matão. Quando chegássemos, iríamos a São Paulo assistir um jogo do Palmeiras no Allianz Parque.

Toda noite, também sei que ele espreitava nossos antigos quartos para ver se as memórias dormiam direito; se escovamos dentes, se estávamos descobertos...

Ele se fragmentou ainda mais quando estivemos longe, mas no meio de 2019, infelizmente, virou poeira de gente e foi soprado entre nós.

Talvez não teríamos dito um ao outro o quanto nos amávamos (ou tínhamos? Acredito que ao nosso modo, sim), e como pai excessivo que era, não se importou em nenhum momento em pegar as rodovias desse país com seu velho caminhão e nos trazer alimentos quando crianças; ele também nunca pensara em perda e nem permanência; só buscou (ao seu modo simples) nos dar o melhor que pôde.

Não viu seu filho do meio, "*mesmo não sendo médico, virar doutor*" (como ele dizia orgulhosamente aos vizinhos) e, apesar dos percalços da vida, nos mostrou, do seu jeito simples, o caminho da estrada... e eu segui. E ainda sigo.

Se o silêncio da morte é grande, o do nosso coração, pai, é maior ainda...

Esta tese de doutorado é dedicada, *in memoriam,* a Luís Vieira, meu velho pai... Quantas histórias...

AGRADECIMENTOS

Gostaria de expressar minha profunda gratidão ao Professor Ian Castro-Gamboa, que não apenas aceitou meu retorno à UNESP após uma empreitada acadêmica muito valiosa no norte do país, mas também pela amizade e por incentivar que eu seguisse por esta linha de pesquisa.

Agradeço aos meus colegas do NUBBE e da pós-graduação, especialmente Givaldo Souza Silva, Tiago Teodoro de Lima Souza, Luciano da Silva Pinto, Camila Cunha, Ana Zanata e Helena Russo, pela camaradagem e apoio contínuo. Sou igualmente grato aos técnicos administrativos e de laboratório, cuja dedicação incansável facilitou grandemente essas análises.

Um agradecimento especial ao Instituto Federal de Rondônia, campus de Ji-Paraná, por permitir minha ausência das atividades acadêmicas por três anos e meio. Sou eternamente grato à Prof. Dra. Kally Alves de Sousa, minha coorientadora e querida amiga, com a esperança de que este seja apenas o começo de muitos outros projetos e artigos colaborativos, e que nossa amizade continue a se fortalecer.

Agradeço também à Fundação de Amparo à Pesquisa de Rondônia (FAPERO) e ao CNPQ pelo suporte financeiro concedido através do edital Universal. Muitos dos resultados apresentados neste trabalho foram possíveis graças a este projeto.

Por fim, agradeço a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho. Cada um de vocês desempenhou um papel crucial na minha jornada, e sou eternamente grato por isso.

RESUMO

Este trabalho está subdividido em seis capítulos. O capítulo 1 tem por objetivo explorar a variabilidade metabólica de produtos de síntese microbiana, provenientes da fermentação natural de sementes de cacau (*Theobroma cacao* L.), na identificação de bioativos de interesse industrial e/ou biotecnológico. Para isso, desenvolveu-se o aplicativo multitarefas LUMIOS, um sistema inteligente que consolida algoritmos para desreplicação, integrando modelos de *machine learning* e abordagens computacionais, o que inclui a docagem molecular, que visam o reconhecimento de feições moleculares em produtos naturais que possam atuar em alvos moleculares associados a doenças respiratórias, como asma e SARS-CoV-2. O capítulo 2 destaca a exploração das misturas complexas do inventário da diversidade metabólica dos exsudatos das sementes fermentadas (e não fermentadas), utilizando o aplicativo LUMIOS, com o intuito de realizar a identificação de sinais oriundos de moléculas de interesse comercial através de ferramentas de desreplicação. O algoritmo LUMIOS efetua comparações com mais de um milhão e duzentos mil espectros de massas, o qual possibilitou a identificação de 13 anotações moleculares, sendo que 10 delas (catequina, trealose, teobromina, procianidina, adenina, indol-3-acetamida, ácido ftálico, anidrido ftálico, fenilalanina e tirosina) apresentaram potenciais para atuarem em alvos de doenças respiratórias. O capítulo 3 apresenta o desenvolvimento e a testagem do aplicativo Chemistika, que, aliado às anotações fornecidas pelo LUMIOS, permite a automatização e tratamento de dados do planejamento de misturas do tipo Simplex-Lattice, do tipo 3x3. Os capítulos 4 e 5 apresentam a utilização do Chemistika para construir modelos que possam prever a intensidade relativa de cada anotação oriunda nas matrizes complexas de cacau e explorar a variabilidade metabólica nas diferentes fases do complexo processo fermentativo. Por fim, o capítulo 6 explora cada anotação molecular apontada pelo LUMIOS à luz da dinâmica molecular (utilizando o algoritmo Gromacs), em um estudo de trajetória de 100 nanossegundos, e faz uso da plataforma CHEIC para analisar os resultados ofertados pelo Gromacs. Os produtos computacionais desenvolvidos neste

trabalho, LUMIOS, Chemistika e CHEIC, representam avanços significativos na exploração de matrizes de produtos naturais. Essas ferramentas não apenas automatizam e agilizam o processo de análise, mas também proporcionam uma compreensão mais profunda das complexas interações moleculares presentes nos produtos naturais. A capacidade de identificar rapidamente moléculas de interesse comercial e biotecnológico, prever a intensidade relativa de anotações moleculares e explorar a dinâmica molecular de compostos promissores tem o potencial de acelerar a descoberta de novos bioativos e otimizar o processo de desenvolvimento de novos produtos. Além disso, ao facilitar a exploração racional da microdiversidade presente em sementes de cacau, essas ferramentas podem contribuir para a valorização deste recurso natural e para o desenvolvimento de iniciativas biotecnológicas inovadoras.

Palavras-chave: inteligência artificial; softwares multitarefas; planejamento de misturas; COVID-19; docking; dinâmica molecular

ABSTRACT

This work is divided into six chapters. Chapter 1 aims to explore the metabolic variability of microbial synthesis products, derived from the natural fermentation of cocoa seeds (Theobroma cacao L.), in the identification of bioactive compounds of industrial and/or biotechnological interest. For this purpose, the multi-tasking application LUMIOS was developed, an intelligent system that consolidates algorithms for dereplication, integrating machine learning models and computational approaches, which includes molecular docking, aimed at recognizing molecular features in natural products that may act on molecular targets associated with respiratory diseases, such as asthma and SARS-CoV-2. Chapter 2 highlights the exploration of the complex mixtures of the metabolic diversity inventory of the exudates of fermented (and non-fermented) seeds, using the LUMIOS application, with the aim of identifying signals originating from molecules of commercial interest through dereplication tools. The LUMIOS algorithm performs comparisons with more than one million two hundred thousand mass spectra, which enabled the identification of 13 molecular annotations, 10 of which (catechin, trehalose, theobromine, procyanidin, adenine, indole-3-acetamide, phthalic acid, phthalic anhydride, phenylalanine, and tyrosine) showed potential to act on targets of respiratory diseases. Chapter 3 presents the development and testing of the Chemistika application, which, together with the annotations provided by LUMIOS, allows the automation and data processing of the Simplex-Lattice mixture design, 3x3 type. Chapters 4 and 5 present the use of Chemistika to construct models that can predict the relative intensity of each annotation originating in the complex matrices of cocoa and explore the metabolic variability in the different phases of the complex fermentation process. Finally, chapter 6 explores each molecular annotation pointed out by LUMIOS in the light of molecular dynamics (using the Gromacs algorithm), in a 100 nanosecond trajectory study, and makes use of the CHEIC platform to analyze the results provided by Gromacs. The computational products developed in this work, LUMIOS, Chemistika, and CHEIC, represent significant advances in the exploration of natural product matrices. These tools not

only automate and speed up the analysis process but also provide a deeper understanding of the complex molecular interactions present in natural products. The ability to quickly identify molecules of commercial and biotechnological interest, predict the relative intensity of molecular annotations, and explore the molecular dynamics of promising compounds has the potential to accelerate the discovery of new bioactives and optimize the process of developing new products. Furthermore, by facilitating the rational exploration of the microdiversity present in cocoa seeds, these tools can contribute to the valorization of this natural resource and the development of innovative biotechnological initiatives.

Keywords: artificial intelligence; multitasking software; mixture design; COVID-19; docking; molecular dynamics.

LISTA DE FIGURAS

FIGURAS – MATERIAIS SUPLEMENTARES

LISTA DE TABELAS

LISTA DE TABELAS SUPLEMENTARES

SUMÁRIO

INTRODUÇÃO GERAL

Esta pesquisa envolveu a cadeia produtiva de cacau, principalmente de base familiar, inserida na mesorregião do leste rondoniense, e relacionou o emprego (e desenvolvimento) de técnicas e ferramentas inovadoras para identificar produtos oriundos do processo fermentativo de sementes de cacau. Este trabalho é fruto de uma colaboração nacional entre o Instituto Federal de Rondônia – IFRO (Ji-Paraná) e Instituto de Química da UNESP (Araraquara), com o propósito de criar uma sinergia de ferramentas científico-tecnológicas para o desenvolvimento de processos biotecnológicos de importância regional e/ou internacional.

No escopo de tais alternativas inovadoras foram empregadas ferramentas da Quimioinformática, da Ciência de Dados e da Inteligência Artificial, por meio do aprendizado de máquina (*Machine Learning)* e do aprendizado profundo (*Deep Learning)*, para resolver problemas no campo da química, como recuperação e extração de informações químicas, pesquisa de banco de dados, mineração de espaços químicos moleculares e criação de modelos capazes de prever faixas mais amplas de padrões de fármaco-similaridade, uma vez que a rápida explosão de *Big Data* de dados químicos e a crescente necessidade da redução de tempo para descoberta de moléculas-fármacos são recorrentes, abordagens computacionais se tornaram ferramenta indispensável para extrair informações de bancos e desenvolver medicamentos com propriedades biológicas importantes.

O *Machine Learning* e o *Deep Learning* são atualmente um dos tópicos mais importantes e em rápida evolução na descoberta de medicamentos auxiliados por computador. Em contraste com os modelos físicos que dependem de equações físicas explícitas, como Química Quântica ou simulações de dinâmica molecular, as abordagens de aprendizagem de máquina usam algoritmos de reconhecimento de padrões para discernir relações matemáticas entre observações empíricas de pequenas moléculas com o intuito de extrapolá-las para prever propriedades químicas, biológicas e físicas de novos compostos. Além disso, a mineração matemática de entidades químicas permite a derivação de uma constelação de descritores, que são empacotados como impressões digitais químicas em uma

variedade de modelos de aprendizado de máquina podendo explorar de maneira eficiente a variabilidade metabólica de um extrato bruto, prevendo padrões de estruturas que possam atuar em alvos biomacromoleculares de doenças diversas.

Desta forma, essa tese será dividida em capítulos, que em conjunto, abordarão: **(i)** curadoria de dados e desreplicação dos exsudatos de sementes fermentadas de cacau por meio de delineamento experimental do tipo misturas; **(ii)** análise de espaço químico, visualização, navegação e comparação das moléculas prospectadas nestes exsudatos, buscando anotações moleculares significativas; **(iii)** previsão de bioatividade das moléculas anotadas e que foram sinalizadas como promissoras por meio de inteligência artificial, docagem e dinâmica molecular **(v)** abordagens computacionais para automatização das análises químicas e desenvolvimento de softwares.

Tal abordagem enfatiza a importância do cacau não apenas na produção de chocolate, mas também como formas alternativas de obtenção de moléculas de interesse comercial para indústrias cosméticas, alimentícias e farmacêuticas.

1 REVISÃO DA LITERATURA

1.1 Cacau.

O cultivo, comercialização e industrialização do cacau (*Theobroma cacao* L.*)* e seus derivados têm apresentado, no decorrer dos anos, importante papel socioeconômico no cenário brasileiro e latino-americano (FRANZEN; BORGERHOFF MULDER, 2007).

A qualidade do chocolate, principal produto obtido do cacau, depende de uma grande variedade de fatores ambientais, agronômicos e tecnológicos (COOPER et al., 2008). Porém, este trabalho vai além dos potenciais de produção de chocolate, ele sinaliza o uso do cacau para indústria farmacêutica, alimentícia e cosmética, no sentido de obtenção de moléculas de alto valor agregado.

Dentre esses fatores, sabe-se que os micro-organismos, presentes na fermentação espontânea das sementes do cacau, desempenham função essencial no desenvolvimento dos metabólitos (MORENO-ZAMBRANO et al., 2018). Considerando que o processo de fermentação e secagem é realizado ainda nas fazendas, sem qualquer controle de processo, uma porcentagem significativa das sementes não sofre as alterações necessárias (principalmente a acidificação do pH e aumento da temperatura) para que as reações enzimáticas se processem de forma satisfatória. Uma possibilidade de remediar este problema é o acompanhamento e intervenção, principalmente no processo de fermentação, objetivando caracterizar os compostos aromáticos, enzimas e melhores condições de processo para melhor uniformizar e aumentar a qualidade das amêndoas de cacau produzidas (CASTRO-ALAYO et al., 2019).

A fermentação do cacau é um processo microbiológico espontâneo, no qual os micro-organismos metabolizam os açúcares fermentescíveis presentes na polpa, em ácido lático e etanol, o qual posteriormente, é oxidado a ácido acético através de reação exotérmica que envolve a atuação de bactérias acéticas. A polpa de cacau é um substrato rico para desenvolvimento microbiano, consistindo em 82-87% de água, 10-15% de açúcar, 2-3% de pentosanas, 1-3% de ácido cítrico e 1-

1,5% de pectina. Proteínas, aminoácidos, vitaminas e minerais também estão presentes (KONGOR et al., 2016a).

O ácido acético e o etanol, produtos do metabolismo microbiano, penetram na semente e em combinação com a ação do calor eliminam a capacidade germinativa do embrião quebrando as paredes celulares da semente. Estas alterações induzem reações bioquímicas dentro da amêndoa, gerando os precursores químicos do sabor e cor do chocolate (SANTANDER MUÑOZ et al., 2020).

O tempo requerido para a fermentação das sementes de cacau é variável. Para a ocorrência das principais reações, que levam à formação dos principais precursores moleculares, as sementes de cacaueiro do grupo Forastero, tipo predominante em todo o mundo, inclusive no Brasil, deve ser geralmente fermentado por períodos superiores a cinco dias (BRUNETTO et al., 2020a).

Muitos micro-organismos fermentadores são provenientes das mãos dos trabalhadores que manipulam os frutos durante os procedimentos para o rompimento das cascas. Ademais, os micro-organismos são oriundos dos cestos utilizados para transporte das sementes e da mucilagem seca, presente nas caixas, remanescente de fermentações anteriores (PUERARI; MAGALHÃES; SCHWAN, 2012; SCHWAN, 1998).

1.2 Quimiometria e desreplicação molecular.

Inúmeras tecnologias têm sido desenvolvidas para explorar e identificar metabólitos secundários, de interesse industrial e farmacológico, produzidos durante processos fermentativos. Nesta perspectiva, destaca-se a quimiometria, que faz uso de ferramentas estatísticas, e a metabolômica, com a abordagem de desreplicação, que visa detectar novos compostos sem a necessidade de isolamento (HILLMAN; READNOUR; SOLOMON, 2017).

A quimiometria é uma área da ciência que utiliza conhecimentos de matemática e estatística para a identificação de informações relevantes de um problema em estudo, facilitando a obtenção de informações (KJELDAHL; BRO,

2010). Durante o processamento dos dados, pode-se realizar uma análise exploratória dos dados químicos fazendo uma varredura nos cromatogramas e nos espectros obtidos, buscando evidências de sinais de moléculas de alto valor agregado.

Adicionalmente, buscando realizar induções planejadas, a quimiometria contribuiu nessa investigação ao criar um delineamento experimental com o intuito de otimizar as condições de extração dos metabólitos oriundos da fermentação das sementes de cacau. Esse tipo de desenho experimental, conhecido em inglês por *Design of Experiments* (DoE), agrega o planejamento fatorial e possibilita o estudo das condições experimentais ideais de um dado problema (POLITIS et al., 2017). Para se obter sucesso na utilização desta técnica, é necessário estimar todos os possíveis fatores que podem impactar na fermentação das sementes de cacau. O planejamento experimental permite investigar, de forma robusta e econômica, os efeitos de vários fatores sobre as respostas de interesse, com o objetivo deste tipo de planejamento é encontrar a combinação ideal dos componentes que maximize a resposta desejada (CURTIS et al., 2022).

O trabalho de identificação, reconhecimento e elucidação estrutural é uma tarefa árdua, fazendo-se uso de técnicas modernas baseadas no princípio da química verde e o uso de simuladores *in silico* (computadores de alta performance) tem atraído atenção de renomados grupos de pesquisa em todo o mundo e tal abordagem tem sido conduzida de forma crescente no Brasil. Recentemente, a desreplicação pode ser auxiliada pela técnica computacionais, que faz relações moleculares com as fragmentações propiciadas pelo espectrômetro de massas (TARTAGLIONE et al., 2023). Essa abordagem metodológica pode ser destinada a distinguir, em matrizes complexas, os compostos já conhecidos daqueles ainda desconhecidos e, que possivelmente apresentem interesse de exploração (KIND; FIEHN, 2017).

Para tal, diversas técnicas hifenadas de separação e detecção são utilizadas, nas quais destacam-se a Cromatografia Gasosa Acoplada à Espectrometria de Massas (CG-EM) e Cromatografia Líquida de Alta Eficiência Acoplada à Espectrometria de Massas (CLAE-EM), bioensaios e análises que permitam a

comparação dos conjuntos de dados obtidos através do uso de base de dados (JOUANEH et al., 2022; QIN et al., 2023).

Tais técnicas, auxiliam na busca de moléculas bioativas já descritas na literatura e disponíveis em bases de dados, possibilitando correlacionar os metabólitos com a cocobiota presente nas sementes fermentadas de cacau fazendo uso de abordagens de ciência de dados, atreladas à quimioinformática, química computacional e inteligência artificial para reconhecer feições moleculares de compostos que possam atuar em alvos de doenças diversas (que neste trabalho, se direcionará à busca de moléculas capazes de modular alvos biomacromoleculares de doenças respiratórias, como bronquite, asma e SARS-CoV-2).

1.3 Abordagens computacionais (Ciência de Dados e Inteligência Artificial).

Para um maior interesse acadêmico e farmacológico, a versatilidade do uso de produtos naturais (PN) em diferentes áreas, como polímeros, suplementos alimentares, agricultura e cosméticos, impulsionou o aumento no número de bancos de dados moleculares abertos e restritos (comerciais) (AHMED et al., 2010; CROTEAU et al., 2000; KULKARNI VISHAKHA; BUTTE KISHOR; RATHOD SUDHA, 2012; SPARKS et al., 2019) e informações químicas agregadas de vários organismos, biomas, doenças específicas e usos tradicionais (SOROKINA et al., 2021).

Um banco de dados químico pode ser definido como uma coleção de moléculas que contém informações sobre compostos e seus descritores químicos, bem como reatividade e diversos recursos biológicos (KOULOURIDI et al., 2019a). Esses bancos de dados são alimentados pela análise de artigos científicos que contêm resultados do processo de isolamento e elucidação estrutural de produtos naturais. Algumas plataformas de banco de dados são sistematicamente analisadas e revisadas, visando racionalizar e organizar as informações disponíveis sobre moléculas orgânicas, às vezes não publicadas, de diferentes origens biológicas.

Informações como estrutura, propriedades biológicas, origem e localização geográfica são inseridas manualmente nos bancos de dados, enquanto as propriedades moleculares e o nome IUPAC, bem como a geração de espectros de ressonância magnética nuclear, são gerados automaticamente (PILON et al., 2017). Assim, a necessidade de armazenar, gerenciar e processar essas informações criou um "Big Data" que contempla um espaço químico diverso (SALDIVAR-GONZALEZ et al., 2018), que pode ser explorado por diferentes métodos *in silico*, permeando ferramentas da inteligência artificial, da quimioinformática, da docagem e da dinâmica molecular.

1.4 Machine learning – ML

A Inteligência Artificial (IA) tem se estabelecido como uma das tecnologias mais influentes do século XXI (ÖZDEMIR; HEKIM, 2018) e o *Machine Learning*, um subconjunto essencial da IA, desempenha um papel fundamental nessa revolução tecnológica (SARKER, 2021). À medida que o mundo se torna cada vez mais dependente de dados, a capacidade de extrair informações valiosas desses dados se torna uma habilidade crítica (QIU et al., 2016).

Machine Learning não é apenas uma ferramenta, mas uma disciplina que revolucionou a maneira como as máquinas compreendem e interpretam informações (SOORI; AREZOO; DASTRES, 2023). Ao invés de depender de regras de programação rígidas, os algoritmos de *Machine Learning* aprendem com dados, refinam suas respostas e se adaptam a novos cenários (WOSCHANK; RAUCH; ZSIFKOVITS, 2020). Isso abre caminho para uma ampla gama de aplicações que vão desde a automação de tarefas rotineiras até a solução de desafios complexos em áreas como medicina, finanças, e indústria (ÇELIK; ALTUNAYDIN, 2018).

Dentro do campo do *Machine Learning*, existem três grandes áreas que constituem seus pilares fundamentais (DAS; DEY; ROY, 2015):

- **Aprendizado supervisionado:** onde um modelo é treinado em dados rotulados, é amplamente utilizado em tarefas como reconhecimento de fala, classificação de imagens e diagnóstico médico.

- **Aprendizado não supervisionado:** O aprendizado não supervisionado, que explora a estrutura latente de dados não rotulados, desempenha um papel crucial em tarefas como segmentação de mercado, análise de redes sociais e detecção de anomalias.
- **Aprendizado por reforço:** O aprendizado por reforço é fundamental em domínios como jogos, robótica e controle de processos. É uma abordagem para a aprendizagem de máquina em que um agente interage com seu ambiente e aprende a tomar ações para maximizar alguma forma de recompensa cumulativa (DAYAN; NIV, 2008).

Em suma, o campo do aprendizado de máquina, ou *machine learning*, está revolucionando a maneira de interagir com a tecnologia e solucionar problemas em diversas áreas. À medida que a pesquisa e a aplicação prática continuam a avançar, pode-se esperar que o aprendizado de máquina desempenhe um papel cada vez mais crucial na sociedade, impulsionando a automação, a tomada de decisões inteligentes e a inovação em todos os setores (RUDIN; WAGSTAFF, 2014). A jornada do *machine learning* está longe de ser concluída, e seu potencial quase que ilimitado promete um futuro empolgante e repleto de possibilidades, sobretudo, na área das ciências.

1.5 Deep Learning – DL

Deep learning, em português "aprendizado profundo," é uma subárea do aprendizado de máquina (*machine learning*) que se concentra na criação e treinamento de redes neurais artificiais profundas para realizar tarefas complexas de análise de dados e tomada de decisão (LECUN; BENGIO; HINTON, 2015). O termo "profundo" refere-se ao fato de que essas redes neurais são compostas por múltiplas camadas de unidades de processamento, conhecidas como neurônios artificiais (MIN; LEE; YOON, 2017).

A principal característica do *deep learning* é a capacidade de aprender automaticamente representações de dados de nível hierárquico, o que permite a

extração de características complexas e abstratas dos dados de entrada (ELHARROUSS et al., 2022). Isso é especialmente útil em tarefas de visão computacional (DEL CAMPO; CARLSON; MANNINGER, 2021), processamento de linguagem natural (KANG et al., 2020), reconhecimento de padrões (BAI et al., 2021), entre outras. Essas redes são treinadas usando algoritmos de otimização para minimizar o erro entre as previsões do modelo e os rótulos (labels) verdadeiros dos dados de treinamento (SUN, 2020).

1.5.1. Arquitetura das redes neurais convolucionais – CNNs

As CNNs são compostas por camadas convolucionais que aplicam operações de convolução para extrair características relevantes dos dados de entrada. Essas camadas são seguidas por camadas de *pooling* para reduzir a dimensionalidade e camadas totalmente conectadas para realizar a classificação ou regressão (BHATT et al., 2021). A arquitetura típica de uma CNN inclui:

- **Camadas de Convolução**: Essas camadas aplicam filtros (*kernels*) a regiões locais da entrada para extrair características relevantes. Cada filtro é aprendido durante o treinamento para detectar padrões específicos (YAMASHITA et al., 2018).
- **Camadas de *Pooling***: As camadas de *pooling* reduzem o tamanho espacial da representação, mantendo as características mais importantes. Isso ajuda a tornar a rede mais eficiente e robusta (SINGH; RAJ; NAMBOODIRI, 2020).
- **Camadas Totalmente Conectadas**: No final da rede, camadas totalmente conectadas são usadas para realizar a classificação ou regressão com base nas características extraídas.

Uma das razões pelas quais o *Deep Learning* se tornou tão poderoso é o uso de conjuntos de dados grandes e avanços em hardware de GPU (*graphics processing units*), que aceleram o treinamento de redes neurais profundas (GAWEHN et al., 2018), propiciando o contínuo desenvolvimento da área, com

avanços constantes em arquiteturas de rede, algoritmos de treinamento e aplicações práticas.

1.6 Inteligência artificial em química

A inteligência artificial, por meio de algoritmos de *machine* e *deep learning*, tem moldado inúmeras áreas do conhecimento, inclusive a química. Suas aplicações são vastas e seu potencial é ilimitado. Em particular, na química, o *Machine Learning* tem desempenhado um papel crucial na aceleração da descoberta de novos compostos (MEUWLY, 2021), na previsão de propriedades moleculares (COVA; PAIS, 2019), na otimização de processos químicos (HORWOOD; NOUTAHI, 2020) e na identificação de padrões em grandes conjuntos de dados experimentais (CHOI et al., 2021).

Uma das contribuições mais significativas do *Machine Learning* na química é a capacidade de projetar moléculas com propriedades específicas, o que é essencial no desenvolvimento de novos medicamentos e materiais (CHAN et al., 2019). Algoritmos de *Machine Learning* podem analisar vastos bancos de dados de compostos químicos e identificar estruturas promissoras que atendem a critérios específicos, economizando tempo e recursos no processo de pesquisa, tornando-a mais eficiente e precisa (PAUL et al., 2021).

À medida que a colaboração entre cientistas e especialistas em *Machine Learning* continua a crescer, é provável que existam avanços ainda mais notáveis na interseção entre a química e o *Machine Learning*. Assim, a área em questão tem experimentado uma significativa e crescente importância nos últimos anos, emergindo como um campo de pesquisa de vanguarda. As pesquisas avançadas realizadas nesse domínio refletem seu impacto duradouro e seu potencial para moldar o futuro de inúmeras esferas da sociedade, sobretudo, na área das ciências. À medida que se continua a explorar as inúmeras implicações e desdobramentos desse campo em constante evolução, pode-se antecipar avanços significativos que irão impactar positivamente uma ampla gama de setores e disciplinas.

1.7 Plataformas web.

A revolução digital no campo da ciência de dados tem transformado várias disciplinas, incluindo a química (STENTA, 2021). Uma consequência direta desta revolução é o acúmulo massivo de dados químicos, consolidados em diversos bancos de dados moleculares disponíveis publicamente ou privadamente. Este acúmulo de dados impulsionou a construção de plataformas web especializadas, projetadas para facilitar e automatizar análises de dados químicos complexos (LI MANNI et al., 2023).

Essas plataformas não apenas tornam o acesso e a análise de dados mais eficientes, mas também facilitam a colaboração entre pesquisadores de diferentes partes do mundo, promovendo assim o avanço da ciência de forma mais globalizada (BATTISTELLA; NONINO, 2012). Além disso, a automação de análises de dados químicos por meio dessas plataformas ajuda a minimizar erros humanos e permite que os pesquisadores se concentrem em aspectos mais criativos e interpretativos de seus estudos (SCHMIDT; LIPSON, 2009).

A disponibilidade dessas plataformas tem implicações significativas para a pesquisa e desenvolvimento em química de produtos naturais, química medicinal, ciências ômicas, entre outras áreas. Elas facilitam a identificação de novos compostos bioativos, a predição de suas propriedades e atividades, e a análise de seus mecanismos de ação. Tudo isso contribui para a aceleração do processo de descoberta de novos fármacos e outros produtos de interesse (LIU et al., 2019) (Gonzalez et al., 2020).

Neste contexto, esta pesquisa de doutorado tem como produto a elaboração de três inovadores aplicativos disponíveis em plataforma web: LUMIOS, Chemistika e CHEIC. Estes foram desenvolvidos com o objetivo de automatizar de maneira inteligente as análises exploratórias dos dados químicos de matrizes complexas de cacau, em suas diversas fases de fermentação. Tais aplicativos não só facilitam e diminuem significativamente o tempo necessário para a análise de dados, mas também integra tecnologias avançadas como inteligência artificial, desreplicação, planejamento experimental, docagem e dinâmica molecular. Dessa forma, os

aplicativos desenvolvidos não apenas otimizam a eficiência do processo de análise, mas também potencializam a qualidade e a relevância dos dados obtidos, contribuindo significativamente para um melhor entendimento das matrizes complexas oriundas durante a fermentação do cacau. Este avanço, por sua vez, tem o potencial de descoberta de novas aplicações para os compostos presentes no cacau.

OBJETIVO GERAL

Explorar matrizes complexas oriundas do processo fermentativo do cacau utilizando delineamento experimental, técnicas de *Machine Learning* e simulações computacionais através de docagem e dinâmica molecular das anotações encontradas frente a alvos biomacromoleculares de doenças respiratórias.

Objetivo 1. Prospectar, através de técnicas de desreplicação, Cromatografia Líquida de Alta Eficiência (CLAE), hifenado ao Espectrômetro de Massas (EM), compostos bioativos sintetizados por micro-organismos ao longo do processo fermentativo espontâneo de sementes de cacau;

Objetivo 2. Avaliar, através de delineamento experimental do tipo misturas, as melhores condições de extração dos metabólitos anotados;

Objetivo 3. Prospectar plataformas de Big Data, qualificadas à implementação de algoritmos de *Machine Learning*, para o processo de mineração molecular;

Objetivo 4. Empregar algoritmos de *Machine Learning* e de *Deep Learning* no processo de mineração molecular, dentro de bancos de dados de acesso aberto, visando congregar informações químicas consistentes (padrões de similaridade) sobre ampla variedade de fontes de origem biológica;

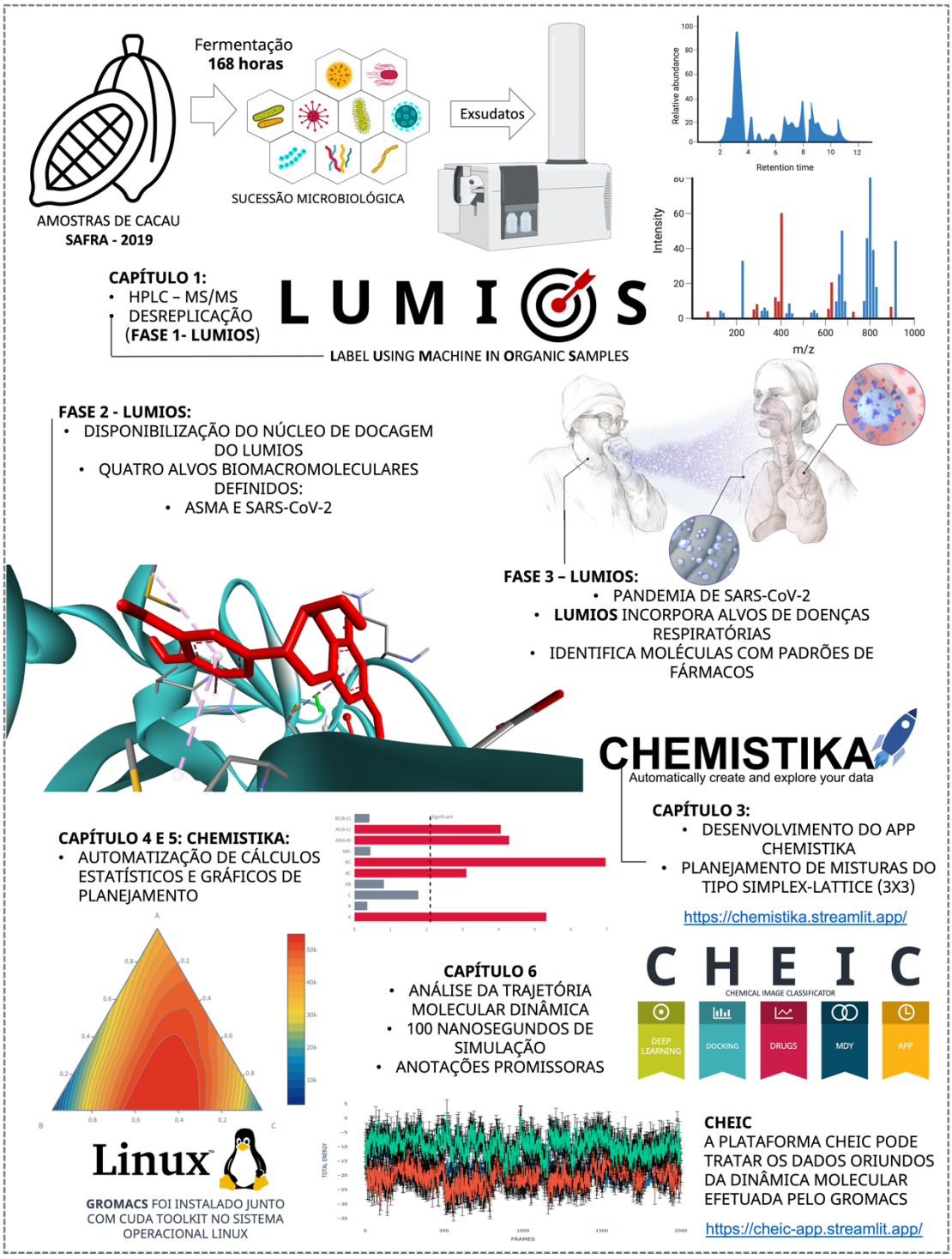
Objetivo 5. Avaliar estruturalmente as anotações classificadas pelos algoritmos quanto às possibilidades de aplicações como fármacos.

Objetivo 6. Avaliar os possíveis metabólitos em alvos biomacromoleculares por meios de técnicas computacionais *in silico*, como docagem e dinâmica molecular.

Objetivo 7. Desenvolver produtos computacionais em forma de web plataformas, que possam ser acessados de maneira gratuita e utilizadas de maneira intuitiva em estudos de desreplicação, aplicação em modelos de *machine learning* e testes computacionais automatizados como docagem e análises de dinâmica molecular.

A Figura 1 faz a representação visual dos pontos principais abordados em cada capítulo desta tese.

Figura 1 - Figura representativa dos capítulos contemplados na tese de doutorado.



Fonte: Elaborado pelo autor (2023).

CAPÍTULO 1 – LUMIOS: LABEL USING MACHINE IN ORGANIC SAMPLES. UM SOFTWARE PARA DESREPLICAÇÃO, DOCAGEM MOLECULAR E COMBINAÇÃO DE MACHINE E DEEP LEARNING[*1]

RESUMO

Com o intuito de realizar o processo de desreplicação e análise exploratória dos dados dos extratos de cacau, criou-se o software LUMIOS, que é o acrônimo para "*Label Using Machine In Organic Samples*". Trata-se de um software multitarefa criado a partir de bibliotecas de linguagem Python que tem como objetivo ajudar profissionais e estudantes na área de química orgânica que realizam abordagens computacionais para a exploração racional de produtos naturais (PNs), sugerindo rótulos moleculares, conhecidos como anotações. LUMIOS não exige codificação ou conhecimento de linguagens computacionais, permitindo uma aplicação amigável e "*no-code*". Esta aplicação recebe informações moleculares a partir do pré-processamento de espectros de massas e realiza a desreplicação molecular, comparando-os com um banco de dados com mais de 1,2 milhão de espectros. As moléculas anotadas passam por um filtro de verificação de que a estrutura pertence à classe de PNs, garantindo que a desreplicação aponte apenas para moléculas indicadas nesta categoria. Tais estruturas podem ser testadas em modelos de Aprendizado de Máquina (*Machine Learning* – (ML)) e Aprendizado Profundo (*Deep Learning* - (DL)), treinados para classificar moléculas de produtos naturais como tendo padrões de moléculas-fármacos, e assim, direcionadas de forma rápida e automatizada para alvos biomacromoleculares de doenças respiratórias (asma e SARS-CoV-2), realizando uma triagem computacional robusta dos dados estudados.

Palavras-chave: processamento de dados; anotações moleculares; produtos naturais; doenças respiratórias.[1]

[1]ARTIGO LUMIOS EM PRÉ-PRINT: REVISTA **EXPERT SYSTEMS WITH APPLICATION**
https://dx.doi.org/10.2139/ssrn.4341603

ABSTRACT

LUMIOS is the acronym for Label Using Machine In Organic Samples. It is a multitasking software created from Python language libraries (adding already consolidated libraries) that aims to help professionals and students in the area of organic chemistry who carry out computational approaches to the rational exploration of natural products (NPs), suggesting molecular labels, known as annotations. LUMIOS does not require coding or knowledge of computational languages, allowing a user-friendly and NO-CODE application. This application receives molecular information from the preprocessing of mass spectra and performs molecular dereplication comparing spectra with a database of more than 1,200,000 chemical information. The annotated molecules pass through a verification filter that the structure belongs to the class of NPs, guaranteeing that the dereplication only points to molecules indicated in this category. Such structures can be tested in Machine Learning (ML) and Deep Learning (DL) models, trained to classify molecules of natural products as having patterns of drug molecules, and thus, directed in a fast and automated way to biomacromolecular targets of respiratory diseases (asthma and SARS-CoV-2), performing a robust computational screening of the studied input data.

Keywords: data processing; molecular annotations; natural products; respiratory diseases.

1 INTRODUÇÃO

Produtos naturais (PNs) isolados de plantas e micro-organismos são fontes promissoras de medicamentos. Eles contribuíram para melhorar a qualidade e estender a vida útil (DEMAIN, 2014), especialmente após a descoberta da penicilina e outros antibióticos (THAKARE et al., 2020). Recentemente, PNs têm sido incluídos em estudos avançados de inúmeros alvos biomacromoleculares, com ênfase em doenças como o câncer (HUANG; LU; DING, 2021), HIV (VONRANKE et al., 2022) e SARS-CoV-2 (BOOZARI; HOSSEINZADEH, 2021). Além disso, é fundamental na descoberta de possíveis medicamentos (NEWMAN; CRAGG, 2012) que podem ser baseados em modificações estruturais de produtos naturais (CRAGG; NEWMAN, 2013).

Nas últimas décadas, os avanços na área computacional e a possibilidade de armazenar grandes quantidades de dados possibilitaram a sistematização de informações moleculares sobre PNs, na forma de bancos de dados químicos, como LOTUS (RUTZ et al., 2022a), COCONUT (SOROKINA et al., 2021), e dados espectrais, como o Banco de Massas da América do Norte – MoNA (https://mona.fiehnlab.ucdavis.edu). Esses bancos de dados compilam milhares de informações químicas sobre PNs abertamente e permitem acesso rápido a essas extensas coleções moleculares, abrindo novos caminhos para analisar o espaço químico que eram inacessíveis sem o poder computacional atual.

Assim, por um lado, a quantidade e complexidade desses dados, principalmente genômicos, proteômicos, metabolômicos e de ensaios clínicos, implica obstáculos nas abordagens de descoberta de medicamentos (GUPTA et al., 2021), por outro lado, o *Big Data* Químico permite a construção de modelos de *Machine Learning* (ML) e *Deep Learning* (DL) capazes de reconhecer padrões estruturais usando descritores moleculares extraídos de bancos de dados (ou calculados pelo software) ou até mesmo examinando imagens moleculares por meio de Redes Neurais Convolucionais (RNC) para seleção de características (GU et al., 2018) das estruturas, o que acelera a obtenção de insights valiosos e significativos durante os estudos computacionais (KEITH et al., 2021), possibilitando

gerenciamento de custos e redução de tempo (ZHANG et al., 2017), tornando a exploração de PNs mais eficiente.

Uma maneira de explorar racionalmente matrizes complexas de produtos naturais é realizar uma triagem computacional baseada nas informações químicas e espectrais desses metabólitos, por meio de técnicas de desreplicação molecular. O termo "desreplicação" foi mencionado pela primeira vez na década de 80 para reconhecer e eliminar estruturas bioativas que já haviam sido previamente elucidadas (LANGLYKKE, 1980), sendo uma maneira eficiente de direcionar o estudo para sinais moleculares ainda não reportados.

A desreplicação também permite o reconhecimento de metabólitos com alto valor agregado (MOHIMANI et al., 2017), quando o interesse é obter estruturas com essas características para estudos futuros, sejam eles biotecnológicos ou sintéticos. Atualmente, o trabalho envolvendo desreplicação molecular tem sido realizado com base em plataformas como a Global Natural Products Social Molecular Networking – GNPS (https://gnps.ucsd.edu) (DE QUEIROZ et al., 2022), MetaboLights (http://www.ebi.ac.uk/metabolights) (HAUG et al., 2020), ou o Metabolomics Workbench (www.metabolomicsworkbench.org) (SUD et al., 2016) com a ajuda de softwares como o Sirius 4 e MS-Finder (DÜHRKOP et al., 2019; MALLMANN; DE OLIVEIRA RIOS; RODRIGUES, 2022).

As ferramentas computacionais também podem otimizar o processo de prospecção de moléculas candidatas a fármacos oriundas de produtos naturais. Embora os recursos relacionados à pesquisa e desenvolvimento de medicamentos tenham crescido, houve uma diminuição na probabilidade de sucesso na obtenção de uma molécula candidata a fármaco (GARCÍA-ORTEGÓN et al., 2022). Para superar esse problema, foi proposto o uso de técnicas de *docking* molecular (CIEPLINSKI et al., 2020). Tais técnicas desenvolvem testes computacionais (rápidos) com as moléculas indicadas no processo de desreplicação contra vários alvos biomacromoleculares.

A internet permitiu a criação de projetos de serviços web dedicados à realização de simulações de *docking* molecular, como o SwissDock (http://www.swissdock.ch), Dockthor (https://dockthor.lncc.br/v2/) e ArgusLab (http://www.arguslab.com).

Nesse cenário, destaca-se o LUMIOS. Um acrônimo para **L**abel **U**sing **M**achine **I**n **O**rganic **S**amples. Diferente das plataformas e softwares já mencionados, o LUMIOS é uma interface de programação de aplicativos (API) multitarefa com interface amigável, rápida e intuitiva (Figura 2), que visa colaborar cientificamente com diferentes áreas do conhecimento de forma interdisciplinar, contribuindo para:

• **Processamento de dados**. Conversão de dados de entrada de formatos como .msp (e outros, como .mgf, .mzml) para .csv (consolidando apenas informações relevantes necessárias, permitindo análises exploratórias dessas informações e visualizações de dados).

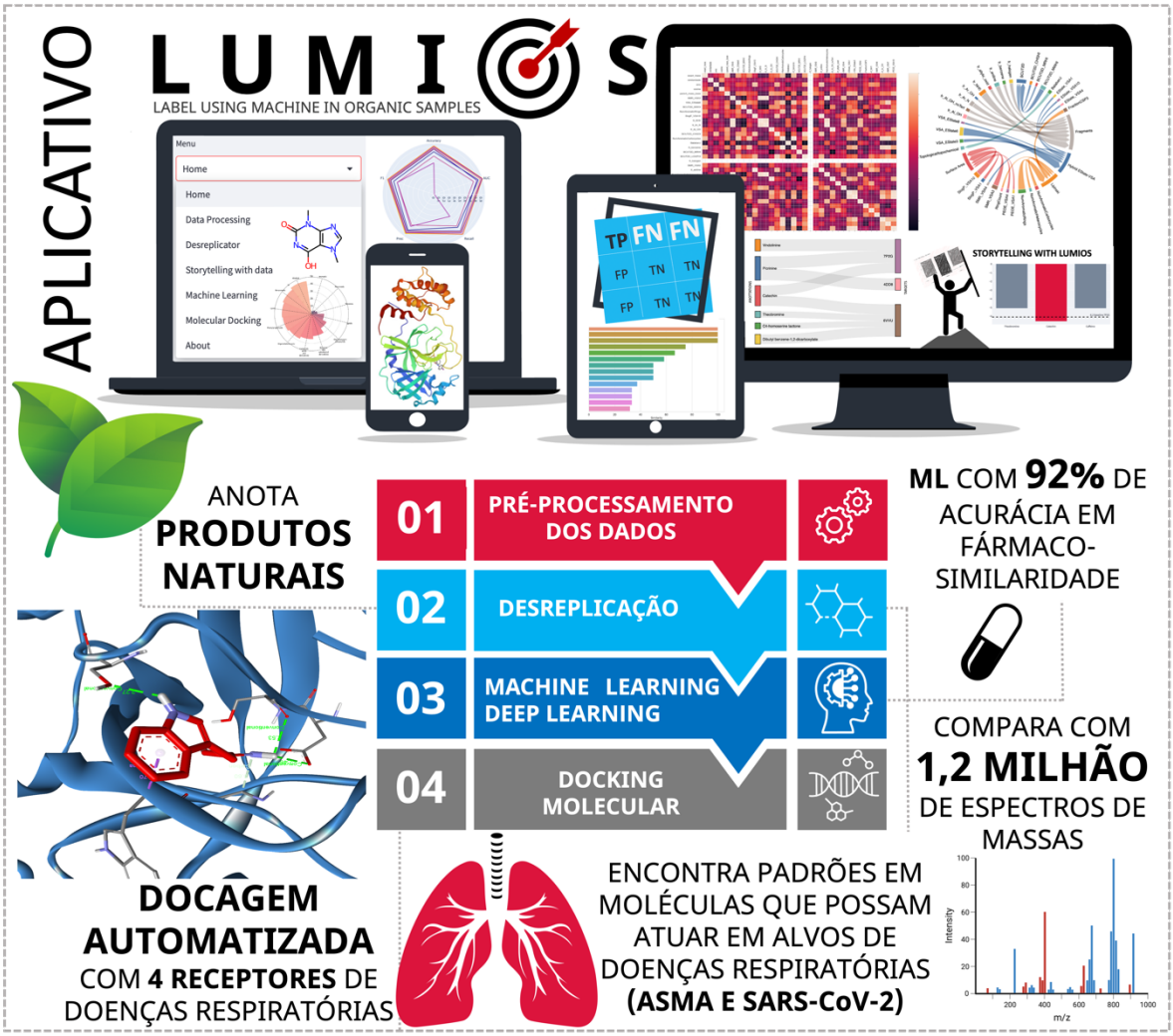
• **Dereplicação**. Dereplicação de matrizes complexas, por meio de comparação com mais de 1,2 milhão de dados espectrais;

• **Storytelling**. Visualização automatizada de informações químicas derivadas de anotações moleculares.

• **Inteligência Artificial**. Modelos de *machine* e *deep learning* que permitem o reconhecimento de padrões de estruturas químicas com potencial para modular proteínas de doenças respiratórias, como asma e bronquite.

• **Docking**. Docagem molecular automatizada, auxiliado por algoritmos consolidados, como o AutoDock Vina que permite boa interpretabilidade, baixos custos computacionais e relevância na triagem exploratória de anotações moleculares.

Figura 2 - Funcionalidades e interface do API LUMIOS.



Fonte: Elaborado pelo autor (2023).

O processo de docking molecular analisa interações entre complexos proteína-proteína ou proteína-ligante. Pode ser usado em triagem virtual em grandes bibliotecas de compostos, classificando os resultados e propondo hipóteses estruturais sobre como os ligantes modulam o alvo (PAGADALA; SYED; TUSZYNSKI, 2017). Atualmente, trabalhos envolvendo docking molecular podem ser realizados em softwares tradicionais como SAnDReS (MORRONE XAVIER et al., 2016), GOLD (VERDONK et al., 2003) e AutoDock Vina (TROTT; OLSON, 2010).

2 METODOLOGIA

O algoritmo LUMIOS foi criado usando a linguagem de alto nível Python e bibliotecas consolidadas atribuídas a esse tipo de programação multiparadigmática. O LUMIOS processa dados de análises de espectrometria de massas e permite entradas nos formatos .MGF, .MZML, .MSP e .MZML. Os dados inseridos passam por um pré-processamento, no qual é utilizada a biblioteca open-access matchms (HUBER et al., 2020), que permite a extração de metadados a partir de informações espectrais, como intensidade de sinal, número de picos, modo de aquisição, precursor e aduto.

Cada amostra é inicialmente comparada com uma coleção espectral retirada da plataforma MoNa (https://mona.fiehnlab.ucdavis.edu/), um repositório auto-curado e centrado em metadados, projetado para armazenamento e consulta eficiente de registros espectrais de massas. Tem como objetivo servir como estrutura para um banco de dados centralizado e colaborativo de espectros de massa de metabólitos, metadados e compostos associados. Mais de um milhão e duzentos mil espectros compõem o banco de dados principal do LUMIOS. Para indicar anotações moleculares, o LUMIOS segue o seguinte pipeline:

- **Filtragem por pesos moleculares:** cada sinal espectral apresenta um precursor chamado "precursor m/z ". A partir dessa informação, o algoritmo calcula o “*parent mass*” e inicialmente assume esse valor como a massa molecular do composto. Esse primeiro filtro permite reduzir a dimensionalidade do espaço químico explorado, pois, nesse momento, os dados de entrada serão comparados apenas com estruturas que atendem ao desvio de 0,01 Dalton (Da) na comparação entre massas moleculares, criando um subconjunto de dados chamado "candidatos à anotação".

- **Comparação MS/MS:** com base nas moléculas candidatas à anotação, são feitas comparações entre os espectros de massa (MS/MS), procurando encontrar correspondências de fragmentação para realizar cálculos de similaridade entre tais informações, permitindo um desvio de 0,02 Dalton (Da) entre fragmentos moleculares. A desreplicação baseada em espectrometria de massas é uma técnica analítica altamente sensível e seletiva, muito popular no campo de produtos naturais, devido à sensibilidade do método e à facilidade de aquisição de dados em grandes conjuntos de amostras.
- **Similaridade espectral:** a similaridade entre os espectros baseia-se na identificação de sinais padrão entre a amostra em avaliação e os dados de massas (MS2) dos candidatos à anotação molecular. Esse conjunto de sinais sinalizados como semelhantes é comparado com todos os sinais disponíveis para essa molécula. O algoritmo realiza a relação entre o tamanho dos vetores (sinais padrão e soma de todos os sinais MS2) para calcular a similaridade molecular. Ficou estabelecido que apenas estruturas com 50% de similaridade passam para a próxima etapa de desreplicação.
- **Anotações de produtos naturais**: LUMIOS apenas anota moléculas relatadas na Plataforma LOTUS (https://lotus.naturalproducts.net). Esta plataforma consolida mais de 276.000 estruturas químicas de produtos naturais e fornece informações bibliográficas sobre esses compostos. Além disso, os *SMILES* (Sistema Simplificado de Entrada de Linha Molecular) das moléculas anotadas são submetidos a um filtro comparativo que identifica sua presença no banco de dados LOTUS.

- **Modelo de *Machine Learning* (ML):** As moléculas anotadas são submetidas a um modelo de *machine learning* treinado para reconhecer estruturas químicas pertencentes a produtos naturais e moléculas em estágios avançados de estudos para o controle de doenças respiratórias, como asma, bronquite e SARS-CoV-2. Os algoritmos (presentes na figura 3) foram testados e o *Light Gradient Boosting Machine* (LGBM Classifier) foi obtido com as métricas mais altas para o reconhecimento de padrões moleculares e separação entre duas classes desejadas. É um algoritmo de aumento de gradiente baseado em árvores de decisão que aumenta a eficiência do modelo e reduz o uso de memória. O conjunto de dados foi compilado a partir de 6.700 estruturas de produtos naturais retiradas da plataforma LOTUS e mais de 6.700 estruturas aplicadas em doenças respiratórias obtidas da plataforma Cortellis (www.cortellis.com). Duzentos e oito descritores químicos diferentes foram calculados, os quais utilizando técnicas de seleção, obteve-se 30 descritores (Tabela 1 e Tabela 2) considerados mais relevantes para o reconhecimento dos padrões propostos: Descritores Constitucionais (0D) (TODESCHINI; CONSONNI, 2008), fragmentos de moléculas (1D) (HELGUERA et al., 2008; MARKOVIC; GUTMAN, 1999), além dos descritores 2D que calculam topologia (BALABAN, 1979; BAYADA; HAMERSMA; VAN GEERESTEIN, 1999), área de superfície (ERTL; ROHDE; SELZER, 2000; STANTON; JURS, 1990), composição híbrida de estruturas baseadas em Áreas de Superfície van der Waals (VSA) (LABUTE, 2000)(Labute, 2000) e descritores vetoriais baseados em valores próprios de Burden (GONZÁLEZ et al., 2005).

Os dados foram sequencialmente separados entre treinamento e teste, utilizando a proporção 80/20, obtendo uma precisão de 92%. Desta forma, os mesmos 30 descritores necessários para inserção no modelo de aprendizado de máquina são calculados para todas as anotações. Assim, eles são submetidos ao poder classificatório do LUMIOS.

Tabela 1 - Descritores constitucionais (DC) e Descritores de Fragmentos (DF) usados como features nos modelos de aprendizado de máquina do LUMIOS

FEATURES	TIPO	DESCRIÇÃO
AromaticRings	Descritor Constitucional	Nº de anéis alifáticos
FractionCSP^3	Descritor Constitucional	Fração de carbonos hibrid. sp^3
AromaticCarbocycles	Descritor Constitucional	Nº de carbociclos aromáticos
AromaticHeterocycles	Descritor Constitucional	Nº de aromáticos heterocíclicos
RingCount	Descritor Constitucional	Contagem de anéis
Fr-NH	Fragmento	Nº de aminas terciárias
Fr-anilina	Fragmento	Nº de anilinas
Fr-Ar-N	Fragmento	Nº de aromáticos nitrogenados
Fr-Al-OH	Fragmento	Nº de grupos hidroxilas alifáticos
Fr-Halogênios	Fragmento	Nº de halogênios
Fr-alílico-oxid	Fragmento	Nº de sítios de oxidação em alílicos
Fr-Alcool-não-tercb	Fragmento	Nº álcool alifático sem grupo terc-butílico
Fr-Benzeno	Fragmento	Nº de anéis benzênicos
Fr-Aromático-OH	Fragmento	Nº de fragmentos fenólicos

Tabela 2 - Variações utilizados como features dos modelos de aprendizado de máquina do LUMIOS

FEATURES	TIPO	DESCRIÇÃO
Slog-VSA10*[1]	Descritor VSA	Informações sobre a área de superfície lipofílica
SMR-VSA3[*2]	Descritor SMR	Capacidade de ressonância de uma molécula (3 frag)
SlogP-VSA1	Descritor VSA	Informações sobre a área de superfície lipofílica
PEOE-VSA4[*3]	Descritor PEOE	Medida da distr. da carga eletrostática da molécula
PEOE-VSA1	Descritor PEOE	Medida da distr. da carga eletrostática da molécula
SMR-VSA4	Descritor SMR	Capacidade de ressonância de uma molécula (4 frag)
VSA-EState6[*4]	Descritor E-State	Volume e área de superfície ($1{,}54 \leqslant x < 1{,}81$)
VSA-EState3	Descritor E-State	Volume e área de superfície ($5{,}00 \leqslant x < 5{,}41$)
VSA-EState8	Descritor E-State	Volume e área de superfície ($6{,}45 \leqslant x < 7{,}00$)
VSA-EState10	Descritor E-State	Volume e área de superfície ($9{,}17 \leqslant x < 15{,}00$)
VSA-EState1	Descritor E-State	Volume e área de superfície ($-\inf < x < 0{,}39$)
VSA-EState4	Descritor E-State	Volume e área de superfície ($0{,}72 < x < 1{,}17$)
BALABANJ[*5]	Desc. Topológico	Considera as distâncias entre os átomos vizinhos
BCUT2D-MWHI[*6]	BCUT2D	MM hidrofóbica ponderada bidimensional (2D)
BCUT2D-CHGHI[*7]	BCUT2D	Carga hidrofílica média ponderada (2D)
BCUT2D-MRHI[*8]	BCUT2D	Média da hidrofobicidade dos fragmentos

*1: VSA (Valence State Atom) – Leva em consideração a estrutura da molécula e a distribuição dos átomos e grupos funcionais envolvidos no fragmento molecular específico. A diferença numérica é relacionada ao tamanho do fragmento que é considerado.
*2: Surface-Mounted Resonance (SMR) – Leva em consideração os fragmentos moleculares e calcula a sua contribuição para o valor de SMR. O tamanho do fragmento molecular é determinado pela conectividade dos átomos e suas ligações dentro da estrutura molecular.
*3: Calcular a contribuição de fragmentos moleculares específicos para o valor de Partial Equalization of Orbital Electronegativity (PEOE). O descritor leva em consideração os fragmentos moleculares compostos de acordo com a numeração especificada e calcula a sua contribuição para o valor de PEOE. O tamanho do fragmento molecular é determinado pela conectividade dos átomos e suas ligações dentro da estrutura molecular.
*4: Descritores estão relacionados à distribuição de cargas eletrostáticas em torno dos átomos da molécula. O número associado a ele faz menção à regiões específicas da molécula, por exemplo, analisando regiões em que podem formar fragmentos com diferentes dimensões (medidas em Ângstrons)
*5: O índice de conectividade de Balaban é calculado considerando as distâncias entre átomos vizinhos na molécula. Quanto maior o valor do índice de conectividade de Balaban, mais conectados e menos ramificados são os átomos na molécula. O descritor "BalabanJ" é uma versão normalizada do índice de conectividade de Balaban, que varia entre 0 e 1. Valores mais próximos de 1 indicam uma maior conectividade e menor ramificação na molécula.
*6: O BCUT2D-MWHI é calculado utilizando um algoritmo que analisa a estrutura molecular e atribui pesos aos fragmentos baseados em sua contribuição para a hidrofobicidade. A soma desses pesos ponderados é então normalizada pela massa molecular total da molécula.
*7: O BCUT2D-CHGHI é calculado utilizando um algoritmo que analisa a estrutura molecular e atribui pesos aos fragmentos com base em sua contribuição para a carga hidrofílica. A soma desses pesos ponderados é então normalizada pela carga total da molécula
*8: O BCUT2D-MRHI é calculado analisando a estrutura molecular e atribuindo pesos aos fragmentos com base em sua contribuição para a hidrofobicidade relativa. A soma desses pesos ponderados é então normalizada pela hidrofobicidade total da molécula.

2.1 Machine learning – ML

Realizou-se estudos classificatórios binários com algoritmos de aprendizado de máquina supervisionados para reconhecer padrões em anotações que poderiam agir em alvos biomacromoleculares de doenças respiratórias. Os algoritmos de aprendizado de máquina testados estão reportados a seguir:

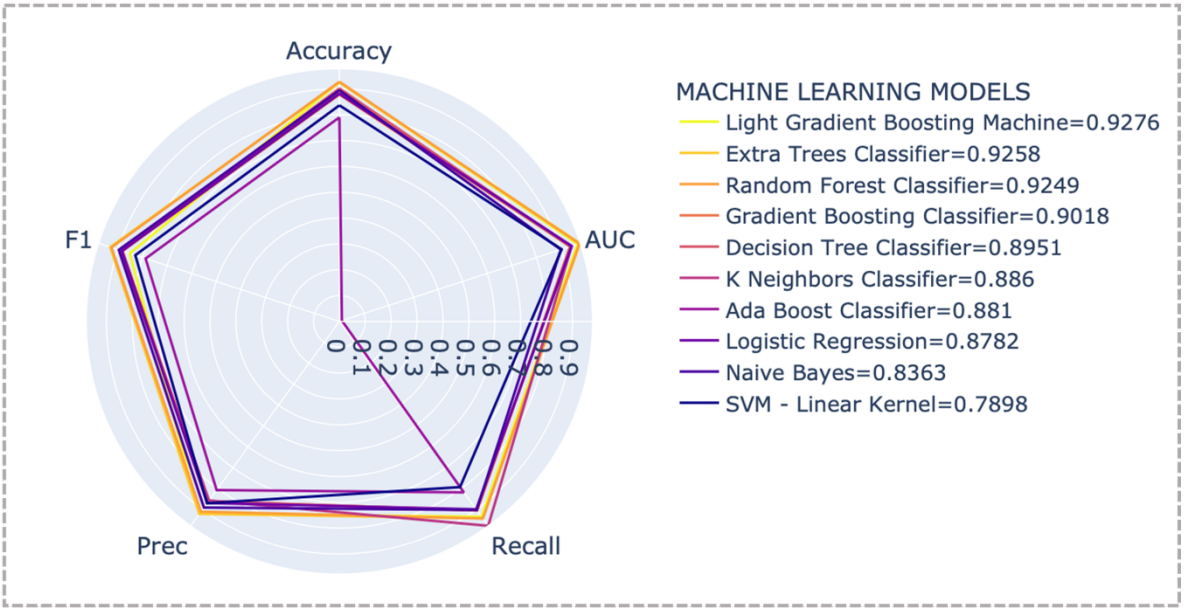
- **Classificador de Regressão Logística - Logistic Regression (LRC)** – um modelo de regressão que permite estimar a probabilidade associada à ocorrência de um determinado evento diante de um conjunto de variáveis explicativas (SONG et al., 2021);
- **K-Nearest Neighbor (KNN)** – o algoritmo busca os 'k' pontos de dados mais próximos no conjunto de dados de treinamento para um ponto específico que se quer classificar (KHANFAR; TAHA, 2013).
- **Árvores de Classificação e Regressão ou Decision Tree (CART)** – Elas funcionam dividindo os dados em subconjuntos cada vez menores baseados nos diferentes valores das variáveis de entrada. Isso é feito através da criação de uma estrutura de árvore, onde cada nó representa um teste em um atributo, cada ramo representa o resultado desse teste e cada folha representa um rótulo de classe (na classificação) (KRZYWINSKI; ALTMAN, 2017);

- **Máquina de Vetores de Suporte - Support Vector Machine (SVM)** – ferramentas de classificação e regressão que constroem hiperplanos no espaço n-dimensional para classificar ou regredir dados (NOBLE, 2006);
- **Classificador de Árvores Extremamente Aleatórias** – Extra Trees Classifier – ET – insere um estimador que se ajusta a um número de árvores de decisão aleatórias (também conhecidas como árvores extras) em vários subamostras do conjunto de dados e usa a média para melhorar a classificação e evitar problemas de superajuste (GOETZ et al., 2014);
- **Floresta Aleatória – Random Forest (RF)** – técnica de classificação baseada em múltiplas árvores de decisão e regras de votação majoritária (SVETNIK et al., 2003);
- **Extreme Gradient Boost (XGBoost)** – O XGBoost (eXtreme Gradient Boosting) utiliza várias árvores de decisão conjuntas para construir um modelo de previsão e usa o reforço de gradiente para minimizar a função de perda, tentando corrigir iterativamente os erros dos passos anteriores. Além disso, ele inclui um termo de regularização na função de perda para controlar o overfitting, o que geralmente resulta em melhor desempenho (ZHANG et al., 2022);
- **Naive Bayes (NB)** – uma abordagem probabilística que utiliza o teorema de Bayes, assumindo independência nos atributos do objeto (BENDER et al., 2006);
- **Light Gradient Boosting Machine (LGBM)** – é uma versão aprimorada do quadro de aprendizado de gradiente baseado em árvores de decisão e na ideia de "aprendizes fracos", que são modelos relativamente simples que quando combinados formam um modelo mais poderoso (KE et al., 2017a);
- **Classificador Ada Boost (ABC)** – como um método geral para gerar um classificador robusto a partir de um conjunto de classificadores fracos (BÜHLMANN; YU, 2000; WYNER et al., 2017).

O modelo de classificação LGBM apresentou a melhor métrica de avaliação (Figura 3) e foi selecionado para verificar a possibilidade de reduzir a

dimensionalidade dos dados de entrada, selecionando 30 dos 208 descritores que mais contribuem para o aprendizado do modelo selecionado. Essa redução de dimensionalidade não causa uma mudança significativa no poder de classificação do modelo e permite uma diminuição no custo computacional para calcular descritores insignificantes para este modelo.

Figura 3 – Métricas de avaliação dos algoritmos de *Machine Learning* utilizados no API LUMIOS.



Fonte: Elaborado pelo autor (2023).

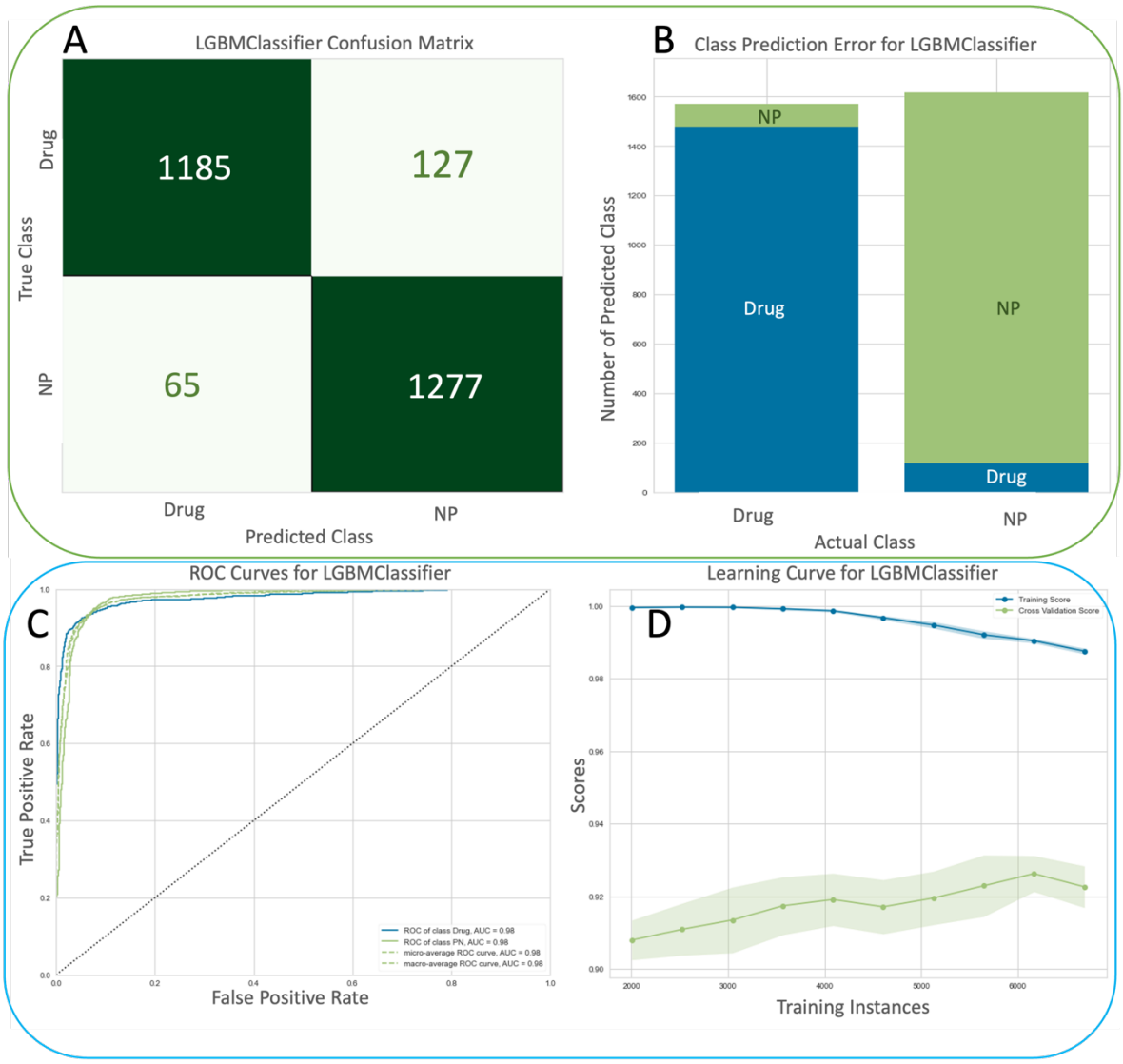
A Figura 4 apresenta as métricas de avaliação do modelo sistematizadas em representações gráficas. Na matriz de confusão da Figura 4-A, é verificado que a classe 0 (definida como classe negativa) são aquelas rotuladas como fármacos, e a classe 1 (classe positiva) são PNs. Entre as 1.312 moléculas que são genuinamente fármacos, o algoritmo classificou corretamente 1.185 estruturas, gerando o grupo de Verdadeiros Negativos (*True Negatives* (TN)). Consequentemente, as estruturas restantes (127) foram erroneamente identificadas como sendo PNs, gerando o grupo de Falsos Positivos (*False Positives* (FP)). Da mesma forma, entre as 1.342 moléculas que são genuinamente PNs, o algoritmo classificou corretamente 1.277 estruturas, gerando o grupo de

Verdadeiros Positivos (*True Positives* (TP)). Por outro lado, 65 estruturas foram erroneamente identificadas como sendo fármacos, gerando o grupo de Falsos Negativos (*False Negatives* (FN)). Na Figura 4-B, apresenta-se o erro de predição da classe para o modelo LGBM.

Na Figura 4-C, é possível observar a relação entre a taxa de falsos positivos e a taxa de verdadeiros positivos, visualizada pela curva ROC, demonstrando que os descritores utilizados no modelo de aprendizado de máquina são consideravelmente distintos para ambas as classes, e que o algoritmo identifica e classifica corretamente mais de 92% das moléculas. A análise exploratória das métricas proporcionou uma visão sobre a divisão entre acertos e falhas do algoritmo empregado. Portanto, devido à sua elevada capacidade de classificação, o modelo LGBM Classifier foi incorporado ao núcleo de classificação molecular do LUMIOS, com o objetivo de identificar moléculas com estruturas específicas passíveis de serem testadas em alvos biomacromoleculares de doenças respiratórias.

A Figura 4-D exibe a curva de aprendizado da máquina ao confrontar os dados de treinamento e teste. Nota-se que, a partir dos descritores escolhidos, o modelo é capaz de identificar padrões para categorizar tanto moléculas de produtos naturais (PNs) quanto moléculas utilizadas no tratamento de doenças respiratórias.

Figura 4 - Métricas de avaliação associadas ao modelo de *machine learning* do LUMIOS.



Fonte: Elaborado pelo autor (2023).

2.2 Deep Learning – DL no LUMIOS[2]

LUMIOS fornece um modelo de Rede Neural Convolucional (*Convolutional Neural Network* – CNN) no núcleo de Inteligência Artificial (IA).

[2]: https://github.com/vieira86/LUMIOS/tree/main/LUMIOS-Notebook
Algoritmos utilizados na construção do modelo de Deep Learning.

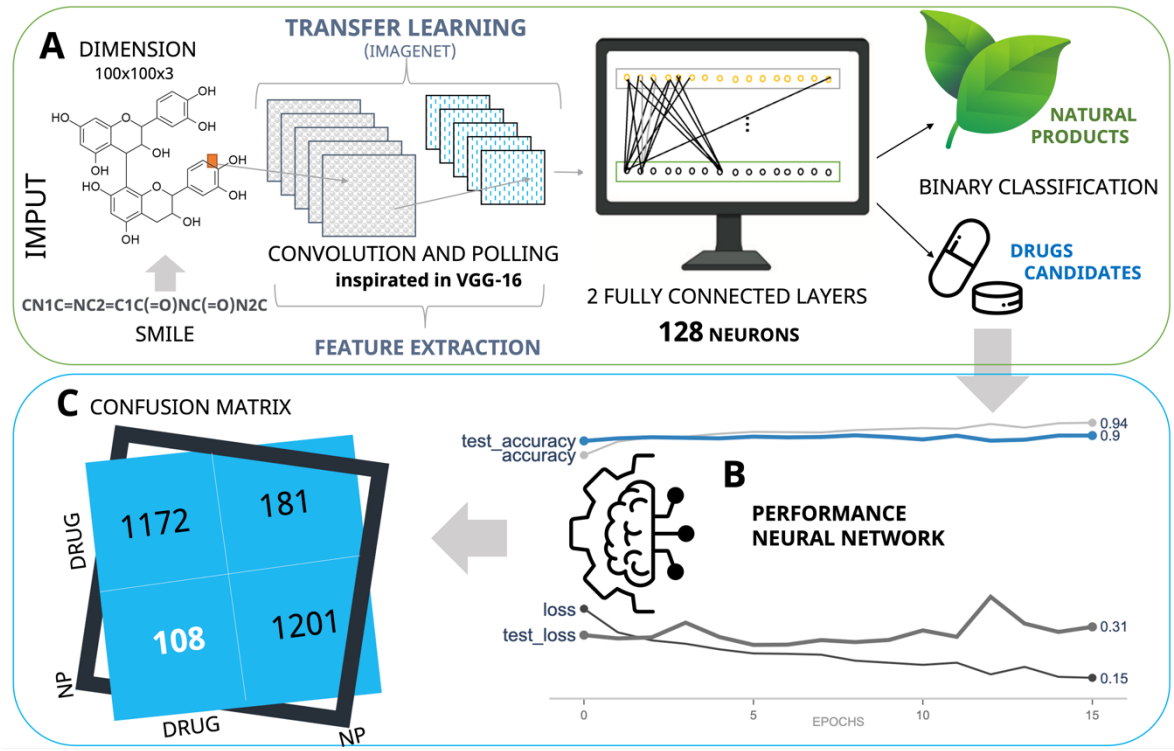
Inicialmente, os dados usados pelo algoritmo de ML tradicional foram utilizados nesta fase. A partir de SMILES, foi construído um algoritmo que converte automaticamente a string (SMILES) em uma imagem colorida, com dimensões de 100x100x3 pixels, criando assim duas coleções de figuras, uma rotulada como produtos naturais e outra como moléculas candidatas a fármacos.

Há casos em que o treinamento de dados usando redes neurais pode ser custoso e demorado. Assim, uma alternativa é usar algoritmos de alta performance construídos inicialmente para resolver problemas em outros domínios. Esse tipo de abordagem é conhecido como transferência de aprendizado (*transfer learning*). O algoritmo LUMIOS para redes neurais utiliza essa abordagem, sendo inspirado na clássica rede VGG-16 (TAMMINA, 2019) (com os pesos treinados para resolver o problema Imagenet (DENG et al., 2009), ligado a duas camadas totalmente conectadas com 128 neurônios, e terminando (por se tratar de uma classificação binária), com um neurônio contendo a função de ativação sigmoidal.

Inicialmente, a arquitetura VGG-16 utilizava imagens com dimensões de 224x224x3 pixels. No entanto, como foi construído um banco de imagens contendo 13.400 imagens moleculares, a quantidade de memória necessária para alocar imagens com tais dimensões estabelecidas por VGG-16 tornou-se inviável, pois seriam necessários mais de 16 Gigabytes (GB) apenas para acomodar essas informações na memória do computador. Assim, optou-se por utilizar dimensões menores, de 100x100x3 pixels, que exigiu 3,2 GB para armazenar as estruturas químicas na memória do computador, sinalizando uma redução de 80% no espaço de memória.

A Figura 5-A resume a arquitetura usada para extrair as características das moléculas e a rede neural totalmente conectada para classificação binária entre as duas classes propostas.

Figura 5 - Representação da arquitetura utilizada pela rede neural artificial para performance do modelo.



Fonte: Elaborado pelo autor (2023).

Além disso, considerando os pesos já treinados para resolver o problema Imagenet, o número de parâmetros necessários para a rede neural aprender os padrões de classificação é reduzido de 15.337.793 para apenas 623.105 parâmetros (apenas os provenientes da rede neural totalmente conectada). Essa abordagem é essencial, pois o custo computacional, a quantidade de memória necessária e o tempo estimado para obter as respostas da rede são reduzidos em mais de 96%. Na Figura 5-B, é possível visualizar o comportamento do treinamento e teste do modelo; com apenas 15 épocas, a arquitetura proposta apresenta uma perda (loss) de 0,16 e acurácia, nos dados de teste, de mais de 89% de assertividade. Isso é transmitido para a matriz de confusão, Figura 5-C, que permite visualizar que, das 1.353 moléculas pertencentes à classe "candidato a fármacos", o algoritmo reconheceu características moleculares assertivas em 1.172 estruturas, definindo o grupo Verdadeiro Negativo (True Negatives – TN); no entanto,

apresentou confusão na classificação das outras estruturas, 181, configurando, portanto, o grupo de falsos positivos (False Positives – FP).

Por outro lado, para a classe de produtos naturais, nota-se que o algoritmo apresentou mais dificuldade em classificar este grupo porque, das 1.309 moléculas de produtos naturais, indicou 108 como pertencentes a fármacos, configurando assim o grupo de Falso Negativo (False Negatives – FN). No entanto, reconheceu padrões de PN em mais de 1.201 imagens moleculares, acertando a classificação e gerando o grupo de Verdadeiro Positivo (True Positives – TP).

Esses dados possibilitam definir as métricas da estrutura desenvolvida para a classificação binária de moléculas, utilizando as mesmas equações empregadas no modelo convencional de aprendizado de máquina que já integra o núcleo de inteligência artificial do LUMIOS. Ao aplicar as equações predeterminadas, as métricas de avaliação dos modelos escolhidos foram adquiridas e contrastadas. Tais informações estão disponíveis na Tabela 3.

Tabela 3 - Métricas de avaliação dos modelos de Machine Learning e Deep Learning incorporados ao LUMIOS

	MACHINE LEARNING			DEEP LEARNING		
CLASSES	**PRECISÃO**	**SENSIB.**	**F1**	**PRECISÃO**	**SENSIB.**	**F1**
FÁRMACOS	0,95	0,90	0,93	0,92	0,87	0,89
NP	0,91	0,95	0,93	0,87	0,92	0,89

2.3 Docagem Molecular

As estruturas químicas anotadas pelo software são direcionadas pelos modelos ML/DL como possuindo padrões de moléculas que poderiam ser usadas em doenças respiratórias e são direcionadas para transformação 3D para testes em alvos biomacromoleculares de doenças respiratórias, como asma e SARS-CoV-2. Auxiliadas pela biblioteca OpenBabel, as moléculas são convertidas para o formato .pdbqt, e usando o algoritmo AutoDockVina a docagem é realizado em quatro proteínas obtidas da plataforma Protein Data Bank – PDB, a saber: 7P2G (SARS-

CoV-2) (ROSSETTI et al., 2022a), 4DD8 (asma) (HALL et al., 2012a), 1NC6 (asma) (COSTANZO et al., 2003a) e 6VVU (asma) (MAUN et al., 2020a).

Inicialmente, as proteínas foram pré-processadas usando o software Chimera (www.cgl.ucsf.edu/chimera), visando identificar as coordenadas dos sítios ativos das proteínas a fim de realizar um teste prático de docagem através do método de redocagem, em que se identifica o ligante originalmente co-cristalizado à proteína, define suas coordenadas e realiza a docagem novamente nesta mesma região, definindo agora um limiar para que possa realizar comparações com os demais ligantes que serão testados. Todas as anotações sinalizadas pelo modelo ML como tendo potencial para modular tais tipos de proteínas são automatizadas pelo LUMIOS. Assim, a melhor pose obtida está vinculada à anotação molecular.

2.4 Processamento dos dados e métricas de avaliação para os modelos de inteligência artificial.

No pré-processamento dos dados para o modelo de aprendizado de máquina, os dados foram divididos em um conjunto para treinamento (80%) e outro para testes (20%). Uma série de transformações foi aplicada para preencher dados ausentes utilizando a mediana e para padronizá-los usando o RobustScaler. Esse escalonador subtrai a mediana e ajusta os dados conforme o intervalo interquartil (IQR, do inglês 'Interquartile Range'). O IQR é a diferença entre o primeiro quartil (25º percentil) e o terceiro quartil (75º percentil). Ademais, para avaliação dos modelos, considerou-se métricas tais como: taxa de verdadeiros positivos (True Positive Rate – TPR)[*2]; taxa de falsos positivos (False Positive Rate – FPR); precisão, especificidade, F1-Score e acurácia.[3]

As fórmulas para cálculo das métricas de avaliação dos modelos são descritas no Quadro 1:

[3] Na classificação binária, a True Positive Rate (TPR), também é conhecida como Sensibilidade ou Recall, é calculada usando a equação 1. Essencialmente, a TPR indica a proporção de positivos reais que foram corretamente identificados como tais.

Quadro 1 - Equações associadas às métricas de avaliação dos modelos de inteligência artificial.

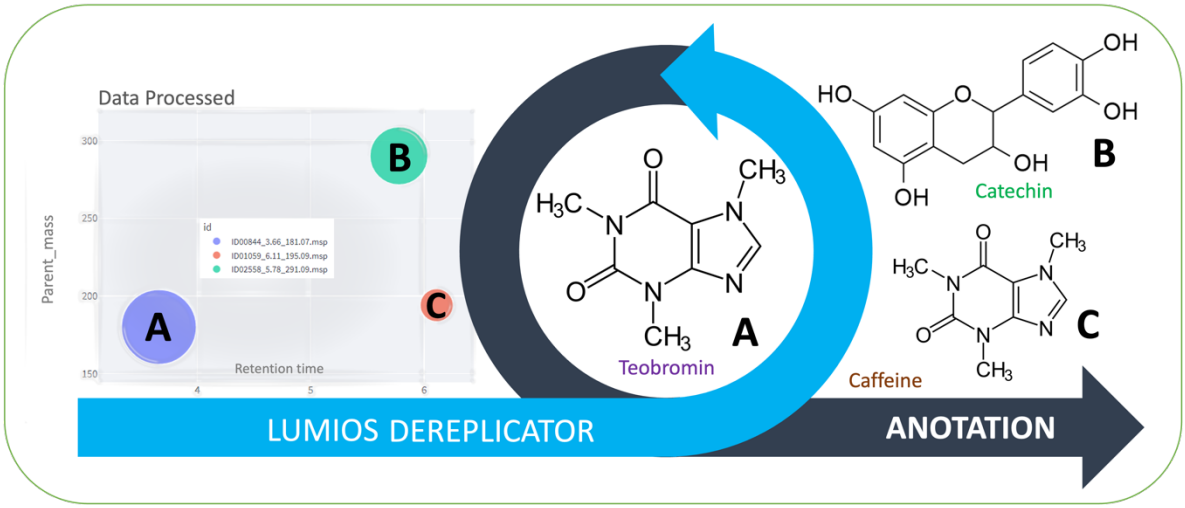
- $\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$ **(Equação 1)** – Taxa de Verdadeiros Positivos*
- $\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$ **(Equação 2)** – Taxa de Falsos Positivos
- $\text{PRECISÃO} = \frac{\text{TP}}{\text{TP+FP}}$ **(Equação 3)**
- $\text{ESPECIFICIDADE} = \frac{TN}{TN+FP}$ **(Equação 4)**
- $\text{ACURÁCIA} = \frac{TP+TN}{TP+TN+FP+FN}$ **(Equação 5)**
- $F1 = 2 \times \frac{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}}$ **(Equação 6)**

3 RESULTADOS E DISCUSSÃO

3.1 Operando o LUMIOS[4]

A Figura 6 sinaliza a disposição de três moléculas (A, B e C) usando o tempo de retenção e o "*parent mass*" após o processamento de dados. Neste exemplo, foram carregados três espectros para mostrar a intensidade de cada sinal. O gráfico é interativo e permite selecionar o alvo e verificar informações sobre o sinal.

Figura 6 - Representação esquemática do processamento e desreplicação de teobromina (A), catequina (B) e cafeína (C).



Fonte: Elaborado pelo autor (2023).

O gráfico disponibilizado pelo LUMIOS é interativo e permite ampliação e redução da imagem, selecionando, assim, o sinal que o usuário gostaria de visualizar com mais precisão. Para demonstrar o funcionamento completo do software, três espectros com alta intensidade de sinal foram selecionados, categorizados de acordo com seus valores de massa-carga (m/z), sendo 195,0881 m/z, 181,0723 m/z e 291,0867 m/z, que foram desreplicados (anotados) como teobromina, cafeína e catequina, respectivamente (Figura 6-A). O reconhecimento de padrões dos espectros de massa MS/MS para cada uma das moléculas

[4] Tutorial de instalação do LUMIOS disponível em: https://vieira-rafael.com/lumios-app/

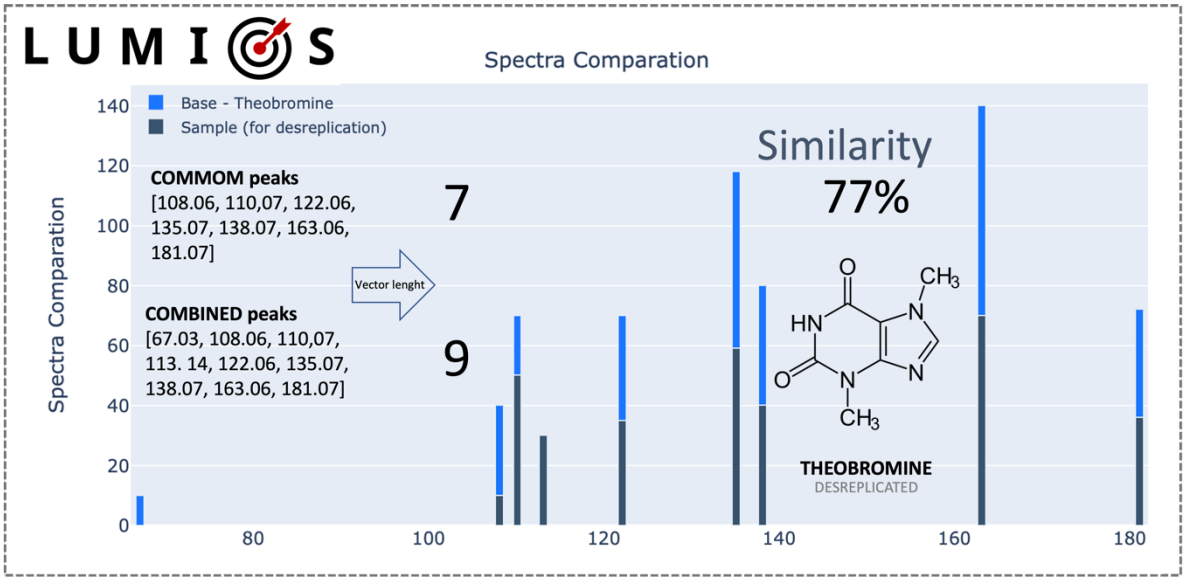
selecionadas foi baseado na combinação vetorial de cada sinal de amostra com dados da base espectral, como mostrado na Figura 7.

O repositório espectral do LUMIOS possui um acervo de mais de 1,2 milhão de amostras. Contudo, durante o procedimento de desreplicação molecular, o algoritmo executa uma filtragem, definindo um limiar de tolerância de 0,01 Dalton para a massa molecular, com o objetivo de diminuir a dimensionalidade do espaço químico investigado. Assim, somente são comparadas estruturas com massas moleculares próximas, que neste exemplo é de 180,0781 g/mol.

O critério de similaridade empregado pelo LUMIOS se fundamenta na conversão de picos MS2 em dois vetores distintos. O primeiro constitui um vetor onde os picos da amostra e da base coincidem, denominado de "Picos Comuns", enquanto o segundo vetor é formado por todos os picos distintos encontrados tanto na amostra quanto na base, sendo este chamado de "Picos Combinados". No caso da molécula com massa molecular de 180,0781 g/mol que passou pelo processo de desreplicação (conforme ilustrado na Figura 7), foram gerados vetores com os valores 7 e 9. Ao calcular a proporção entre esses dois vetores, constatou-se que a similaridade entre o espectro da amostra e o espectro armazenado no banco de dados foi de 77%. O LUMIOS estabelece como aceitável um percentual de similaridade superior a 50%. Assim, determinou-se que a razão massa-carga do pico 181,0781 m/z é análoga ao espectro de referência, que neste caso corresponde à molécula de teobromina. Dessa forma, procedeu-se com a desreplicação molecular do pico 181,0781 m/z, registrando-o como teobromina.

Análise idêntica foi realizada para as amostras com razões massa-carga de 195,0881 m/z e 291,0867 m/z, resultando em índices de similaridade de 100% e 75%, respectivamente. Isso permitiu a identificação e rotulação desses picos como cafeína e catequina.

Figura 7 **-** Representação esquemática para obtenção do grau de similaridade entre os espectros comparados



Fonte: Elaborado pelo autor (2023).

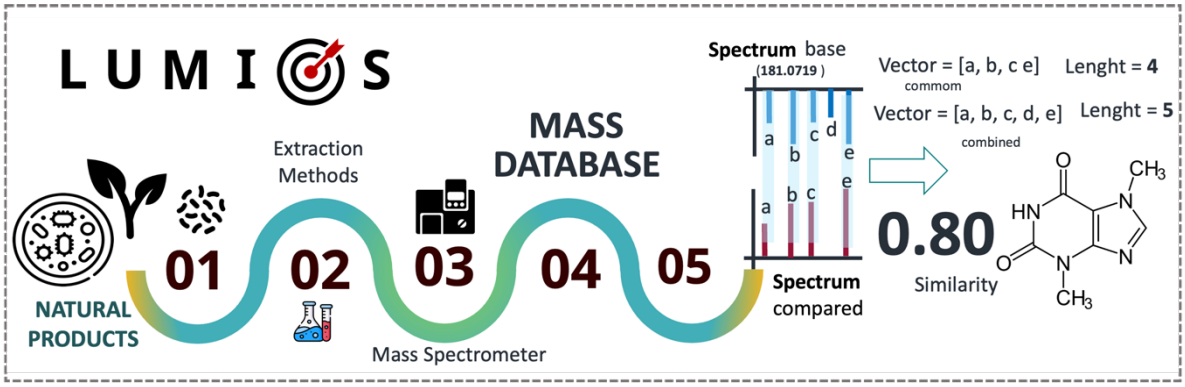
3.2 Desreplicação

O LUMIOS foi estrategicamente projetado para ser utilizado por colaboradores em áreas científicas que utilizam abordagens computacionais em suas rotinas de laboratório. Como resultado, uma interface amigável foi projetada para que os usuários possam utilizá-la de forma fluida e intuitiva e, assim, em poucos minutos, obter insights significativos em sinais de amostras (extratos brutos) de produtos naturais. Os testes foram realizados com dados obtidos por espectrometria de massa de alta resolução em modo positivo. Para validação do algoritmo, foram utilizados três marcadores principais do cacau: cafeína, teobromina e catequina, que geram precursores com uma razão massa/carga (m/z) de 195,0881 m/z, 181,0723 m/z e 291,0867 m/z.

Os dados foram carregados no formato .msp e efetuou-se o pré-processamento da amostra, no qual a intensidade do sinal foi normalizada para o intervalo de 0 a 1, aplicando-se um limiar de 0,2 Dalton para minimizar o ruído espectral nos dados. A Figura 8 sintetiza os passos necessários para alimentar o núcleo de desreplicação do LUMIOS, destacando a comparação dos espectros

MS/MS das amostras de teobromina e indicando uma similaridade de 80% entre os espectros MS/MS analisados.

Figura 8 - Representação dos passos de execução do algoritmo LUMIOS para desreplicação molécula.



Fonte: Elaborado pelo autor (2023).

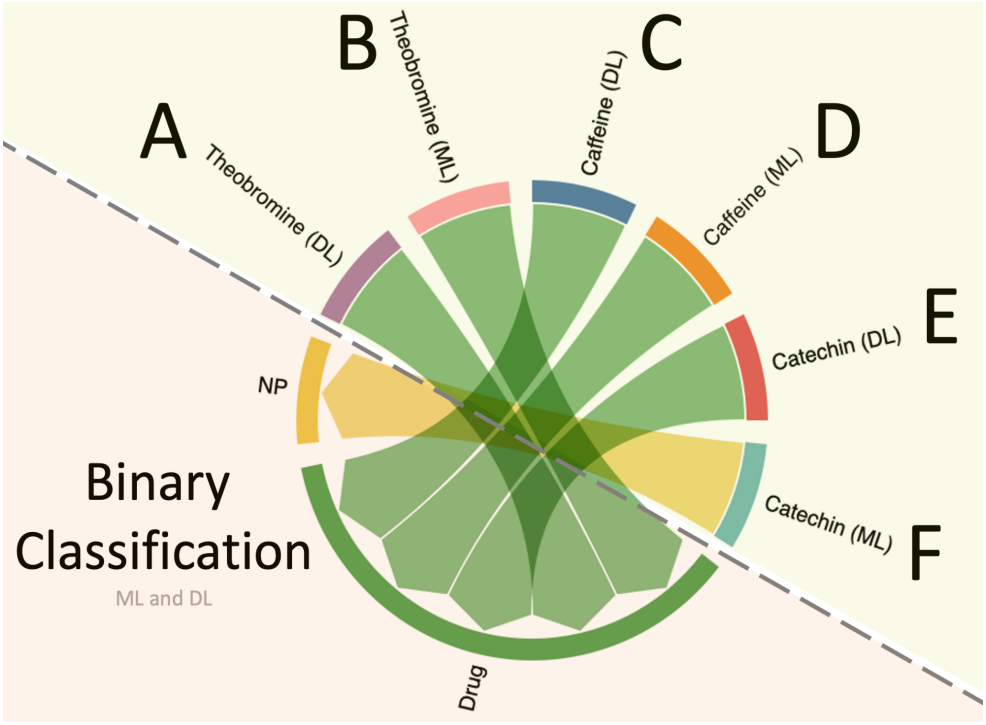
Após a conclusão do processo de desreplicação molecular e a identificação das moléculas candidatas para anotação, o modelo de aprendizado de máquina pode ser empregado para a identificação de padrões em estruturas que possam ser submetidas a testes em alvos biomacromoleculares de doenças tais como bronquite, asma e SARS-CoV-2.

3.3 Classificação de Anotações Usando Modelos de IA

As anotações moleculares podem ser processadas por modelos de aprendizado de máquina (ML) ou aprendizado profundo (DL) treinados para identificar padrões em estruturas químicas potencialmente aplicáveis no tratamento de doenças respiratórias. Trata-se de uma classificação binária onde a molécula é categorizada como produto natural ou candidata a fármaco. Para as três moléculas analisadas neste estudo, a partir de suas respectivas estruturas SMILES, foram computados 30 descritores químicos utilizando a biblioteca RDKIT (LANDRUM; OTHERS, 2013) (conforme Tabelas 1 e 2). Os modelos de IA foram então treinados e avaliados utilizando estes descritores como “features” (ML) e as imagens das

fórmulas estruturais (DL), resultando nas classificações moleculares exibidas na Figura 9.

Figura 9 – Classificação através de técnicas de inteligência artificial (Machine Learning e Deep Learning) para classificação das anotações moleculares em "fármacos" ou "produtos naturais".



Fonte: Elaborado pelo autor (2023).

• **Xantinas (Cafeína e Teobromina)** – Ilustradas na Figura 9, as anotações denotadas pelas letras A-D foram reconhecidas por ambos os algoritmos de inteligência artificial (ML e DL) como possuindo características moleculares análogas a estruturas já avançadas em estudos voltados para alvos respiratórios, sendo dois deles candidatos a medicamentos. Em 1985, Simons e colaboradores (SIMONS et al., 1985) já haviam documentado a atividade broncodilatadora da teobromina. Posteriormente, outros estudos corroboraram a eficácia das metilxantinas no tratamento de doenças respiratórias (JAIN et al., 2020), o que confirma a acurácia dos modelos de ML e DL na identificação desta classe de moléculas.

• **Flavonoide (Catequina)** – Apesar da literatura mais recente citar catequinas na luta contra a influenza e doenças do trato respiratório (UMEDA et al., 2021a), o modelo de ML não indicou a catequina como candidata a fármaco, atribuindo-lhe o rótulo de NP (Figura 6-F). No entanto, a CNN usada para seleção de recursos de imagens de estruturas moleculares reconheceu padrões diferentes daqueles indicados pelas técnicas de ML.

Enquanto o ML classificou as anotações apenas por descritores numéricos para reconhecer padrões moleculares, a CNN realiza uma varredura por diversos filtros chamados Kernels, que visam extrair ou realçar recursos de imagem e armazená-los em um vetor, que é inserido como dados de entrada para uma rede neural densa, classificando a catequina como possível candidata a fármaco (Figura 6-E). Além disso, as catequinas exibem atividade antiviral (em estudos *in vitro*) contra diversas doenças infecciosas agudas (FURUSHIMA et al., 2019a). Portanto, é essencial destacar que um produto natural tem potencial em apresentar atividades biológicas em vários alvos biomacromoleculares.

Os modelos de ML podem ser aplicados como ferramentas para prospectar moléculas com possível atividade biológica; no entanto, eles devem ser complementados com outras técnicas computacionais, como docking molecular e deep learning.

Sequencialmente as anotações podem ser submetidas ao núcleo de docking molecular do LUMIOS. Este núcleo automatiza a preparação das estruturas e realiza a docagem nas quatro proteínas descritas na Tabela 4.

Tabela 4 - Configuração e resultados de docagem para os ligantes padrão associados aos receptores disponíveis no LUMIOS

RECEPTOR	**COORDENADAS (X, Y, Z) -**	**TAMANHO GRADE**	**VALORES ($kcal.mol^{-1}$)**
7P2G (SARS-CoV-2)	8.62, -0.44, 20.33	11.97, 6.92, 9.60	-6.1
4DD8 (Asthma)	12.98, -6.18, -4.50	8.18, 11.35, 13.23	-6.4
1NC6 (Asthma)	43.74, 22.74, 50,02	7.64, 17.28, 11.06	-6.2
6VVU (Asthma)	23.30, 11.98, 65.28	8.32, 7.99, 10.26	-4.3

3.4 Docagem Molecular

As proteínas disponíveis no LUMIOS para estudos de docking molecular foram pré-processadas para identificar a grade tridimensional na qual as anotações moleculares serão inseridas. Assim, o ligante padrão foi removido da proteína e substituído na mesma região para criar uma linha de base comparativa para cada alvo biomacromolecular. Para inserção no núcleo de docking do LUMIOS, as proteínas foram convertidas para o formato .pdbqt usando o software AutoDockTools (HUEY; MORRIS; FORLI, 2012). O docking de cada anotação foi automatizado nas quatro proteínas usando algoritmos do AutoDock Vina. As poses que configuraram as melhores interações com a proteína foram vinculadas à anotação. Os resultados referentes ao estudo de docagem podem ser consultados na Tabela 5.

Tabela 5 - Resultado da docagem molecular efetuada pelo LUMIOS.

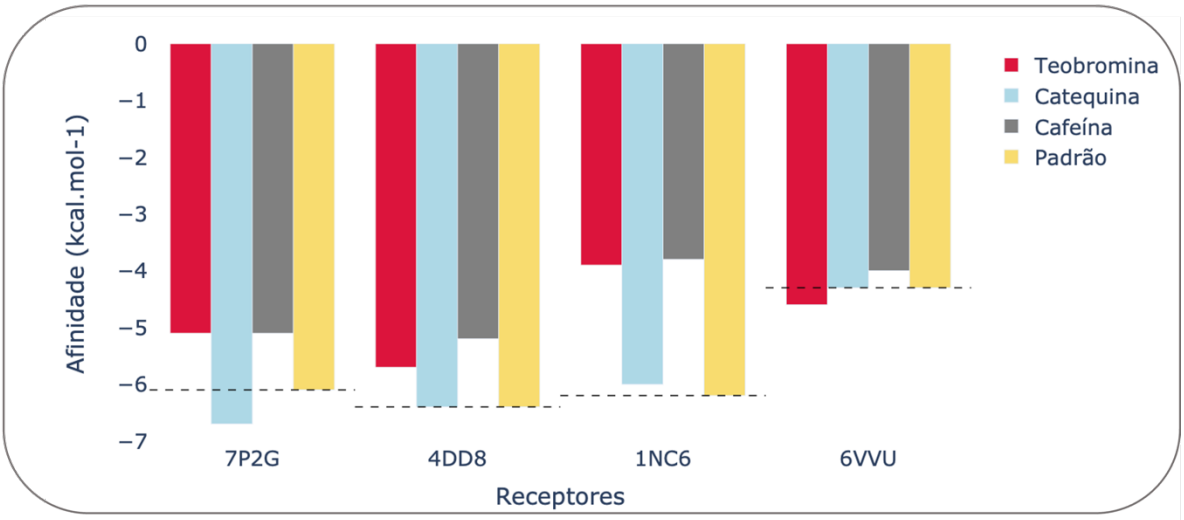
ANOTAÇÃO	**RECEPTORES (kcal.mol^{-1})**			
	7P2G	**4DD8**	**1NC6**	**6VVU**
TEOBROMINA	-5.1	-5.7	-3.9	-4.6
CATEQUINA	-6.7	-6.4	-6.0	-4.3
CAFEÍNA	-5.1	-5.2	-3.8	-4.0
PADRÃO	***-6.1***	***-6.4***	***-6.2***	***-4.3***

Os resultados de docagem molecular apontam nuances interessantes na afinidade dos compostos com os respectivos receptores. A catequina demonstrou alta afinidade para o receptor 7P2G, a principal protease da síndrome respiratória aguda grave SARS-CoV-2, essencial para o ciclo de vida viral, em linha com os relatos na literatura sobre a ação efetiva das catequinas em casos sintomáticos e assintomáticos afetados pelo SARS-CoV-2 (CHOURASIA et al., 2021a; HENSS et al., 2021a; MISHRA et al., 2021a). Os resultados para a catequina validam o sinal indicado pelo modelo CNN, superando o composto originalmente co-cristalizado com a proteína, além de ser um ligante melhor do que a teobromina e cafeína, que apresentaram valores semelhantes entre si.

No caso do receptor 4DD8, a catequina e o composto-padrão exibiram a mesma afinidade, enquanto a teobromina mostrou uma afinidade ligeiramente reduzida. O receptor 1NC6 mostrou maior afinidade pelo composto padrão, seguido de perto pela catequina, enquanto teobromina e cafeína permaneceram na extremidade inferior da escala de afinidade por este alvo. No contexto do receptor 6VVU, tanto a catequina quanto o composto padrão mantiveram sua superioridade em afinidade, coincidindo exatamente nos valores, enquanto cafeína e teobromina exibiram uma afinidade marginalmente menor. A cafeína, apesar de ter uma estrutura muito próxima à da teobromina, apresentou uma afinidade de apenas -4,0 kcal/mol. Tais resultados corroboram os relatados na literatura, nos quais a teobromina é indicada no tratamento da asma e outros problemas do trato respiratório, como a tosse, para os quais até agora não foi aplicada para medicina tradicional, mas o mesmo não é relatado para a cafeína (MARTÍNEZ-PINILLA; OÑATIBIA-ASTIBIA; FRANCO, 2015).

A Figura 10 sistematiza os resultados de docagem, onde é possível realizar uma análise comparativa mais clara entre os ligantes avaliados. De modo geral, a catequina emerge como um composto de significativa afinidade, especialmente para os receptores 7P2G, 4DD8 e 6VVU. A cafeína, embora demonstre afinidades consistentes, tende a apresentar valores inferiores quando comparada à catequina. A teobromina, por sua vez, exibe uma variação de afinidade mais pronunciada dependendo do receptor em questão.

Figura 10 - Representação gráfica dos resultados gerados pelas anotações em comparação com o ligante originalmente co-cristalizado aos alvos biomacromoleculares estudados.



Fonte: Elaborado pelo autor (2023).

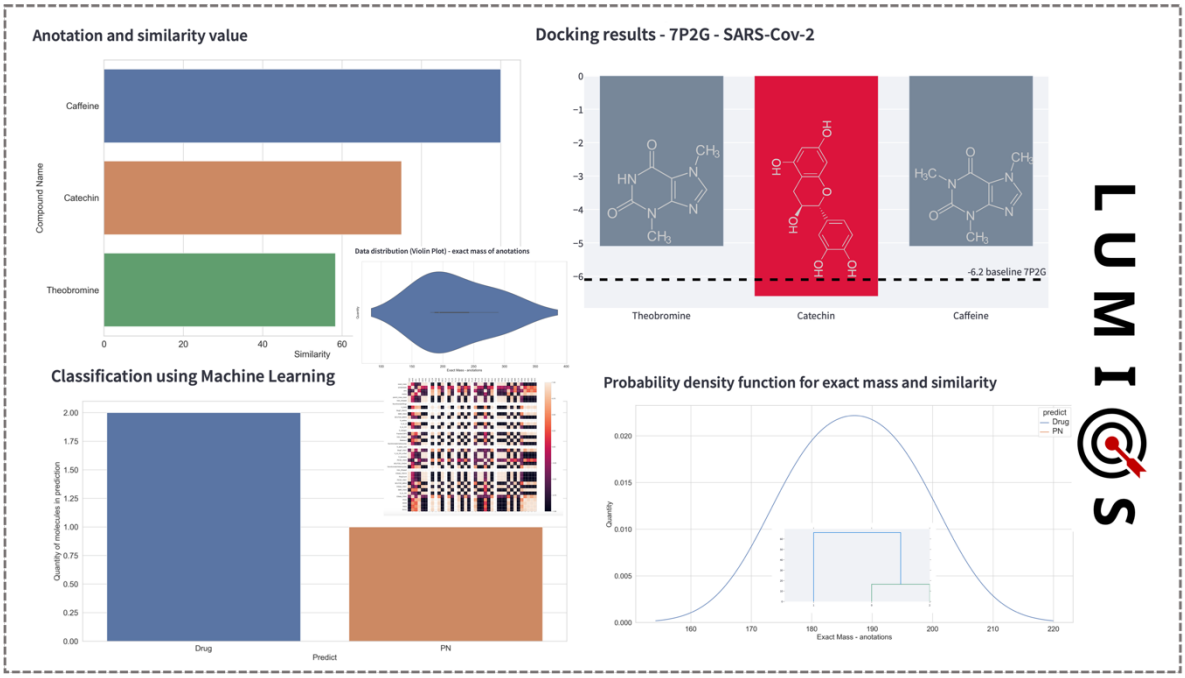
3.5 Storytelling: Contando histórias (a partir de dados moleculares) com LUMIOS

A fase de análise de dados representativos é fundamental e apresenta desafios únicos. Não só é necessário encontrar o método mais adequado e realizar a análise prática dos dados, mas também é vital apresentar os resultados de maneira clara e compreensível, o que requer tempo e esforço consideráveis (HUNTER, 2010). O storytelling desempenha um papel crucial nesse processo. Ao transformar dados brutos em uma narrativa coerente, o storytelling ajuda a destacar os insights mais importantes e facilita a compreensão e a interpretação dos resultados. Com a ajuda do storytelling central do LUMIOS, o usuário poderá não apenas automatizar a análise exploratória dos dados anotados, mas também obter insights significativos usando ferramentas gráficas e estatísticas, como Análise de Componentes Principais – PCA, Análise de Cluster Hierárquico – HCA, métricas de distribuição, tendências, visualização de resultados, comparações gráficas e notas das estruturas que podem modular as proteínas disponíveis no LUMIOS.

Além disso, o storytelling para dados é essencial para comunicar eficazmente os resultados da análise facilitando, assim, a tomada de decisões informadas e a

implementação de ações baseadas em dados. A Figura 11 sistematiza algumas das visualizações geradas pelo software.

Figura 11 – Exemplos de visualizações gráficas obtidas das análises exploratória dos dados anotados usando o LUMIOS.



Fonte: Elaborado pelo autor (2023).

4 CONSIDERAÇÕES SOBRE O LUMIOS

LUMIOS é uma API multitarefa que oferece cinco grandes blocos de processamento de dados que permitem insights significativos a respeito dos espectros de massa explorados, especialmente de amostras de produtos naturais. Portanto, é essencial enfatizar que o processo de desreplicação molecular é baseado em informações consolidadas em bancos de dados espectrais (GAUDÊNCIO et al., 2023a). LUMIOS está em processo de melhoria e implementação, mas algumas observações sobre o software devem ser destacadas:

- Em relação à padronização dos dados, existem vários espectros para a mesma molécula, obtidos em espectrômetros de marcas diferentes e com energias de colisão distintas. É necessário definir um limite para a intensidade dos sinais que serão considerados nas comparações para evitar a aquisição de um grande número de ruídos, que podem perturbar o poder comparativo do algoritmo;
- Como alguns dados são obtidos em plataformas públicas, isso pode gerar inconsistências nos metadados disponíveis, como erros no aduto e no modo de aquisição do espectro. Portanto, é necessário confiar na ciência de dados e realizar um pré-tratamento e seleção dos espectros mais adequados a serem inseridos no núcleo desreplicativo do software. Essa etapa de curadoria dos espectros é indubitavelmente a que demanda mais tempo no desenvolvimento dos códigos;
- LUMIOS ainda não opera com dados de massa obtidos no modo negativo (esforços estão sendo feitos para superar essa questão);
- As versões subsequentes do LUMIOS incorporarão o processamento de dados de espectros de baixa resolução;
- Em relação ao poder computacional, LUMIOS pode processar mais de 1,2 milhão de dados espectrais, tornando possível expandir a coleção de espectros para fins comparativos. No entanto, o tempo de processamento de dados durante a desreplicação terá que ser estendido (mas essa expansão

é possível e viável) e o custo computacional para alocar dados adicionais na memória;

- No treinamento de ML, as instâncias podem ser aumentadas. Novos alvos biomacromoleculares podem ser investigados e inseridos no núcleo de aprendizagem do LUMIOS (de acordo com as necessidades dos usuários, que podem solicitar essas inserções, bem como relatar inconsistências e sugestões de melhorias);
- LUMIOS é projetado para ampliar o filtro de produtos naturais. Por meio de abordagens de web scrapping, novas coleções de produtos naturais estão sendo incorporadas ao LUMIOS, com o objetivo de refinar a desreplicação, apontando apenas para estruturas químicas obtidas de produtos naturais, como plantas, fungos e bactérias;
- Disponibilidade do software em servidores mais eficientes para o processamento mais rápido e eficaz de dados do usuário.
- Embora as métricas de avaliação do modelo de ML tenham sido mais altas na indicação de moléculas candidatas a fármacos, observou-se que o poder classificatório do deep learning, mesmo com métricas ligeiramente menores do que o ML, fez indicações mais consistentes quando comparado aos resultados de docking molecular. Tanto o deep learning quanto o docking utilizaram a estrutura molecular para testes, enquanto o algoritmo de ML foi baseado em diferentes descritores moleculares.
- O ML usa 30 descritores químicos diferentes para propor a classificação, e o docking usa a estrutura em formato tridimensional para testar se um determinado metabólito pode interagir efetivamente com a proteína. Assim, diferentes técnicas *in silico*, como ML e docking molecular, devem ser hifenizadas. Dessa forma, foi possível obter uma ideia das potenciais aplicações biológicas e quais anotações são mais promissoras em alvos biomacromoleculares.

CAPÍTULO 2 – ANÁLISE EXPLORATÓRIA DAS MATRIZES COMPLEXAS DOS EXSUDATOS DE CACAU (*Theobroma cacao* L.) VISANDO O RECONHECIMENTO DE ESTRUTURAS COM AFINIDADE POR ALVOS DE DOENÇAS RESPIRATÓRIAS (ASMA E SARS-COV-2)

RESUMO

Esse capítulo destaca a relevância da integração de ferramentas quimiométricas e quimioinformáticas para conduzir a análise exploratória das matrizes complexas encontradas nos exsudatos de cacau, efetuando a desreplicação molecular de extratos brutos de tais matrizes (*Theobroma cacao* L.). As anotações moleculares geradas foram processadas pelo algoritmo de inteligência artificial do LUMIOS, que foi treinado e validado com uma acuracidade superior a 90%, utilizando informações moleculares oriundas de plataformas de bancos de dados de produtos naturais e de moléculas em fases avançadas de pesquisa biológica. O objetivo foi identificar padrões em moléculas de produtos naturais e entidades químicas que pudessem interagir com alvos biomacromoleculares associados a doenças respiratórias, como bronquite, asma e Covid-19. Das 13 anotações químicas extraídas dos extratos brutos, 10 delas (teobromina, catequina, trealose, procianidina, adenina, indol-3-acetamida, fenilalanina, tirosina, ácido ftálico e anidrido ftálico) demonstraram afinidade com proteínas relacionadas à asma e ao SARS-CoV-2.

Palavras-chave: SARS-CoV-2; Asma; docagem molecular; produtos naturais; inteligência artificial

1 INTRODUÇÃO

Produtos Naturais (PNs) obtidos de plantas, micro-organismos, animais e organismos marinhos têm sido utilizados para reestabelecer a saúde humana há milênios, seja para alívio ou tratamento de diversas doenças. Tais PNs têm sido usados como base de produtos farmacêuticos e como ingredientes de medicamentos tradicionais, desempenhando um papel essencial no desenvolvimento de novos medicamentos (HUANG; ZHANG, 2022; KOEHN; CARTER, 2005).

A relevância das medicinas tradicionais (MTs) é indiscutível e elas frequentemente recorrem ao uso de produtos naturais. Estas modalidades de medicina, incluindo a medicina tradicional chinesa (MTC), práticas indígenas, Ayurveda, Kampo, medicina tradicional coreana e Unani, incorporam produtos naturais em seus tratamentos e têm sido adotadas globalmente por centenas ou até milhares de anos (HEINRICH, 2003; YUAN et al., 2016). Embora possam apresentar limitações em suas diversas formas, ainda são um valioso repositório de conhecimento humano.

Os PNs representam fontes promissoras de estruturas moleculares para a descoberta de novos medicamentos para numerosas terapias (ZHANG et al., 2014). Somente no último século, os PNs possibilitaram a descoberta de moléculas importantes para a medicina, como a penicilina e a lovastatina (isoladas de fungos), captopril (obtido do veneno da serpente brasileira jararaca), tubocurarina, artemisinina e taxol (isolados de plantas) (VALLI; BOLZANI, 2019), e posteriormente, foram encontrados em extratos fúngicos, sinalizando novos esforços para prospectar metabólitos produzidos por fungos endofíticos, que são semelhantes ou iguais aos metabólitos sintetizados pelas plantas hospedeiras (ZHAO et al., 2011).

A versatilidade de usar PNs em diferentes áreas – como polímeros, suplementos alimentares, agricultura e cosméticos – tem impulsionado o crescimento de bancos de dados moleculares abertos e restritos (comerciais) (AHMED et al., 2010; CROTEAU et al., 2000; KULKARNI VISHAKHA; BUTTE

KISHOR; RATHOD SUDHA, 2012; SPARKS et al., 2019). Um banco de dados molecular contém uma coleção de moléculas que armazena vastas informações sobre compostos químicos (produtos naturais), seus descritores físico-químicos e diversos recursos biológicos (KOULOURIDI et al., 2019b).

Apesar dos avanços na química sintética e da biotecnologia, os produtos naturais ainda permanecem como uma valiosa fonte de moléculas ativas para o desenvolvimento de medicamentos (ATANASOV et al., 2021). Em um estudo de revisão abrangente realizado por Newman e colaboradores (NEWMAN; CRAGG, 2016) é apresentada uma análise dos novos medicamentos aprovados pela FDA (*Food and Drug Administration*) de 1981 a 2014, mostrando que a maioria das novas moléculas aprovadas durante esse período era baseada em compostos naturais.

A exploração racional de fontes naturais com o objetivo de reconhecer estruturas químicas de alto valor agregado gerou uma quantidade significativa de informações moleculares que estão sendo consolidadas e disponibilizadas em bancos de dados públicos (HASTINGS et al., 2012). Esses bancos de dados são utilizados como ferramentas para estudos envolvendo trabalhos de desreplicação molecular, que é uma estratégia utilizada em produtos naturais para reduzir esforços e tempo gastos na elucidação de moléculas inéditas, ou até mesmo para reconhecer estruturas de interesse comercial em extratos brutos (CORLEY; DURLEY, 1994).

Além disso, avanços na cristalografia de raios-X, aliados aos esforços garantidos por consórcios genômicos, permitiram que inúmeras proteínas de diferentes alvos terapêuticos fossem disponibilizadas em bancos de dados biológicos (SUSSMAN et al., 1998), abrindo caminho para estudos computacionais (*in silico*) analisando o comportamento de pequenas moléculas aplicadas em alvos biomacromoleculares diversos (KITCHEN et al., 2004).

Graças a esses esforços, abordagens computacionais têm se tornado presentes na química de produtos naturais (BAJORATH, 2002), afinal, desde a década de 1980, a metodologia de docking molecular tem sido utilizada (KUNTZ et al., 1982) e ganhou força com o aumento do desempenho computacional, permitindo o teste de estruturas moleculares pequenas em alvos

biomacromoleculares, possibilitando estudos mais direcionados para descoberta de medicamentos (GOHLKE; KLEBE, 2002).

Neste trabalho, explorou-se as matrizes complexas provenientes do processo de fermentação do cacau, utilizado como protagonista no chocolate, mas que também tem se mostrado uma fonte promissora de moléculas com ação broncodilatadora (DUKE, 2000) e em alvos biomacromoleculares de outras doenças respiratórias, como o SARS-CoV-2 (YAÑEZ et al., 2021a).

Neste contexto, em consonância com bancos de dados público amplamente acessível, que engloba tanto a quantidade quanto a qualidade dos dados, acompanhado pelo constante avanço do poder computacional, surgiram novas ferramentas baseadas na esfera do aprendizado de máquina (ARTRITH et al., 2021). Essas ferramentas foram integradas ao universo da química, e permitiram acessar espaços anteriormente desconhecidos (DREW et al., 2012; GROMSKI et al., 2019; LOWE, 2015).

A riqueza de informações em repositórios moleculares tornou possível a utilização de ferramentas de quimioinformática capazes de realizar cálculos de diversos descritores químicos, incluindo aspectos físico-químicos, bidimensionais (e tridimensionais) e quânticos (MORIWAKI et al., 2018). Esta abordagem facilita a identificação de padrões em entidades químicas encontradas em produtos naturais, possibilitando avaliações numéricas que transcendem a esfera estritamente molecular (NETTLES et al., 2006).

Este capítulo focou em métodos computacionais, especificamente em aprendizado de máquina, quimioinformática e química computacional (docking molecular), utilizando o software LUMIOS (descrito no capítulo 1), para desreplicar o extrato bruto obtido em diferentes etapas do processo de fermentação do cacau (*Theobroma cacao*) visando a identificação de padrões em entidades químicas com potencial terapêutico para alvos macromoleculares de doenças respiratórias, como asma e SARS-CoV-2.

2 METODOLOGIA[5]

2.1 Fermentação espontânea das sementes de cacau

Os grãos de cacau fermentados espontaneamente foram gentilmente fornecidos por uma cooperativa em Ji-Paraná, Rondônia, Brasil. Os grãos, provenientes de diferentes híbridos de *Theobroma cacao* L., foram oriundos da safra de 2019 (entre abril e maio).

A fermentação espontânea dos grãos foi conduzida nas instalações da cooperativa, em tanques de madeira, ao longo de um período de 168 horas. Para análise, amostras foram coletadas em momentos distintos do processo de fermentação: no início (0 horas), no meio (84 horas) e ao final (168 horas). Posteriormente, estas foram conservadas a 4°C a fim de preservar os metabólitos produzidos durante a fermentação dos grãos. Os momentos de coleta foram estrategicamente definidos para assegurar que as amostras fossem representativas das alterações metabólicas ocorridas ao longo de todas as fases da fermentação. É importante destacar que as amostras foram compostas, resultantes da mistura de quinze amostras simples, retiradas de diferentes tanques e profundidades.

2.2 Exsudatos da fermentação do cacau

Os exsudatos aderidos aos grãos, formados durante os diferentes estágios de fermentação, foram submetidos a macerações sequenciais a frio, sob diferentes misturas de extratores (1:3 m/v), de acordo com o Planejamento de Misturas Simplex-Lattice. Foram obtidos oitenta e quatro extratos brutos: 28 (devido a duplicata) a partir de amostras de grãos de cacau coletadas no início da fermentação (0 horas), 28 de amostras de grãos de cacau fermentados por 84 horas e 28 de amostras de grãos de cacau fermentados por 168 horas. Após a obtenção

[5] Ao longo desta tese, a menos que seja especificamente indicado de outra forma, todos os dados explorados nos capítulos subsequentes foram originados das mesmas matrizes complexas, utilizando os mesmos dados espectrais e empregando as metodologias detalhadas no Capítulo 2 (páginas 81-88).

dos extratos brutos, as amostras foram enviadas para análise por Cromatografia líquida acoplada à espectrometria de massas sequencial (HPLC-MS/MS).

2.3 Análises de HPLC-MS

Os oitenta e quatro extratos foram individualmente purificados por extração em fase sólida (SPE) em cartuchos de fase reversa C18. Inicialmente, os cartuchos foram ativados com 4,5 mL de MeOH e equilibrados com 4,5 mL de uma solução 95:5 (v/v) MeOH:H_2O. Então, 20 mg de cada extrato bruto foram solubilizados em 1,5 mL de MeOH:H_2O 95:5 (v/v) e submetidos a um banho ultrassônico por 2 minutos. Sequencialmente, a eluição foi realizada com 1,5 mL de MeOH:H_2O 95:5 (v/v). Após a secagem, 4 mg de cada extrato foram dissolvidos em 2 mL de MeOH, centrifugados a 6.000 rpm e filtrados através de uma membrana de nylon de 0,22 µm.

As análises de HPLC-MS/MS foram realizadas em um espectrômetro Agilent 6565 Q-TOF (Agilent Technologies Inc., Santa Clara, CA) equipado com uma fonte de ionização por electrospray. O software MassHunter® foi utilizado para aquisição de dados e processamento. A separação cromatográfica foi realizada utilizando uma coluna Agilent Zorbax SB-C18 (3.0 × 50 mm), usando um gradiente linear com solventes A: H_20/0,1% ácido fórmico (v/v) e B: MeOH/0,1% ácido fórmico (v/v) variando de 5% a 100% de B em 15 minutos para reequilibração da coluna, com injeção de 5 µL da amostra e fluxo de 0,3 $mL.min^{-1}$.

Os dados de massa foram adquiridos através do modo de ionização positiva [$(M+H)^+$] sob uma fonte de electrospray e analisador TOF-MS, aplicado no modo Auto-MS/MS para fragmentação. Os parâmetros de operação do analisador TOF-MS foram estabelecidos utilizando os seguintes critérios: voltagem do capilar de 2.400 V, voltagem do cone de 65 V e voltagem do fragmentador de 110 V. A pressão do gás nebulizador utilizada foi de 28 psi, com uma taxa de fluxo de ar seco de 10 L/min, temperatura do gás a 300°C, gás envolvente com um fluxo de 10 L/min e temperatura de 350°C, com aquisição de 3 espectros por segundo e resolução de

32.000. Os dados foram processados utilizando o banco de dados MassHunter®, levando em consideração a ionização ($[M+H]^+$) obtida do TOF-MS.

2.4 Processamento dos dados

Os dados do espectrômetro de massa foram convertidos para o formato .MZML usando o software ProteoWizard. Em seguida, foram inseridos no software MS-DIAL (versão 4.9.2), configurando os dados de MS1 e MS/MS como centroides, e o modo de ionização positiva aplicado para metabolômica. A tolerância para MS1 e MS2 foi de 0,01 e 0,02 Da, respectivamente, definindo a faixa de detecção de massa de 100 a 1000 Da para MS1 e MS2.

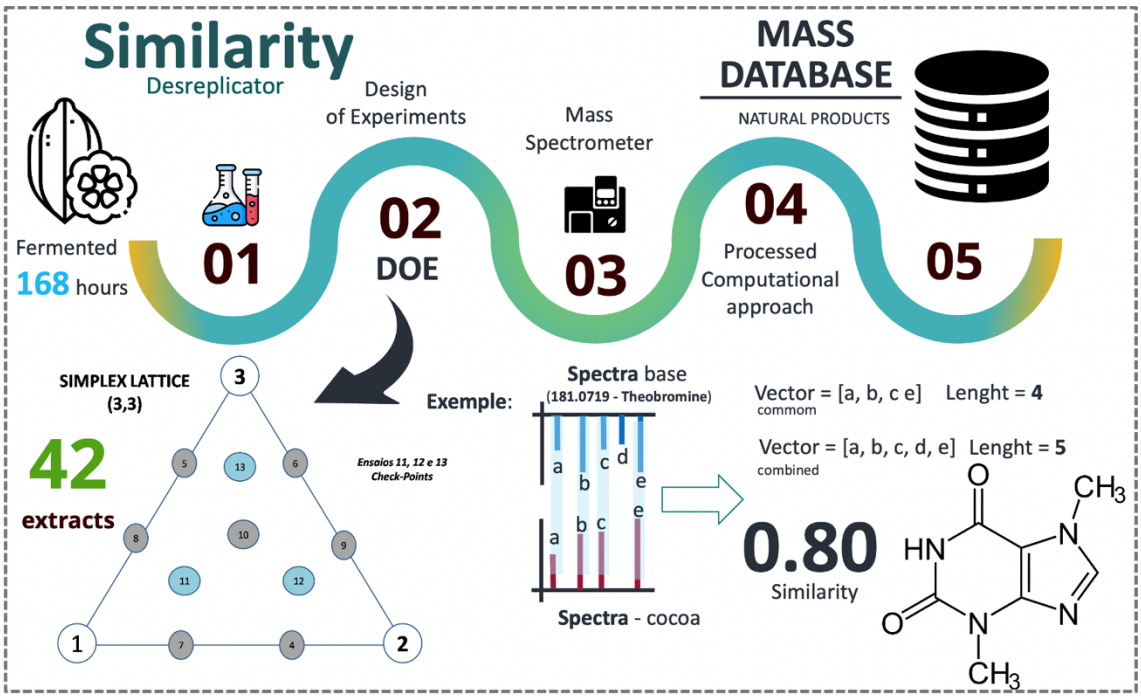
Em relação à amplitude do sinal, foi definido um limiar de 1000 e uma abundância de amplitude de 30 para MS/MS. Foi inserido filtro para remover o branco, e comparados os sinais moleculares dos oitenta e quatro extratos, representando as diferentes etapas do processo de fermentação. Os dados alinhados foram exportados no formato .MSP. Os arquivos foram inseridos na plataforma LUMIOS (Label Using Machine In Organic Samples) para realização da desreplicação de matrizes complexas e testagem das anotações moleculares por meio de técnicas *in silico*, como a docagem, além do auxílio de algoritmos de inteligência artificial. Os modelos de *machine* e *deep learning* reconheceram padrões em moléculas que puderam ser testadas em alvos biomacromoleculares de doenças respiratórias.

2.5 Desreplicação – Algoritmo de similaridade molecular

Os espectros analisados previamente no MS-Dial foram importados para a plataforma LUMIOS. Os dados MS2 foram normalizados, estabelecendo um limite de detecção e comparação de 0,01 Da. Os espectros para comparação foram adquiridos da MoNA - MassBank of North America (https://mona.fiehnlab.ucdavis.edu), totalizando 1.200.000 espectros.

Para cada íon precursor identificado no cacau, uma busca foi realizada no banco de dados espectral para encontrar íons precursores com uma diferença de massa de no máximo 0,01 Da. Após essa etapa inicial de filtragem, os dados MS-MS foram comparados. A métrica de similaridade espectral foi estabelecida com base no comprimento dos vetores gerados para cada amostra. Por exemplo, se um espectro de cacau contiver fragmentos com massas hipotéticas de 110, 138 e 195 Da, o comprimento do vetor correspondente a essa amostra seria 3. Se esse vetor fosse comparado com um espectro do banco de dados contendo fragmentos de massas 110, 138, 195 e 210 Da, o comprimento do vetor seria 4. Portanto, a similaridade espectral seria de 3/4, ou 75%. A Figura 12 ilustra a metodologia utilizada para gerar as anotações moleculares que foram posteriormente submetidas aos modelos de *machine* e *deep learning*.

Figura 12 - Representação esquemática da abordagem exploratória dos extratos brutos de cacau.



Fonte: Elaborado pelo autor (2023).

2.6 Modelos de inteligência artificial: *Machine Learning* e *Deep Learning*

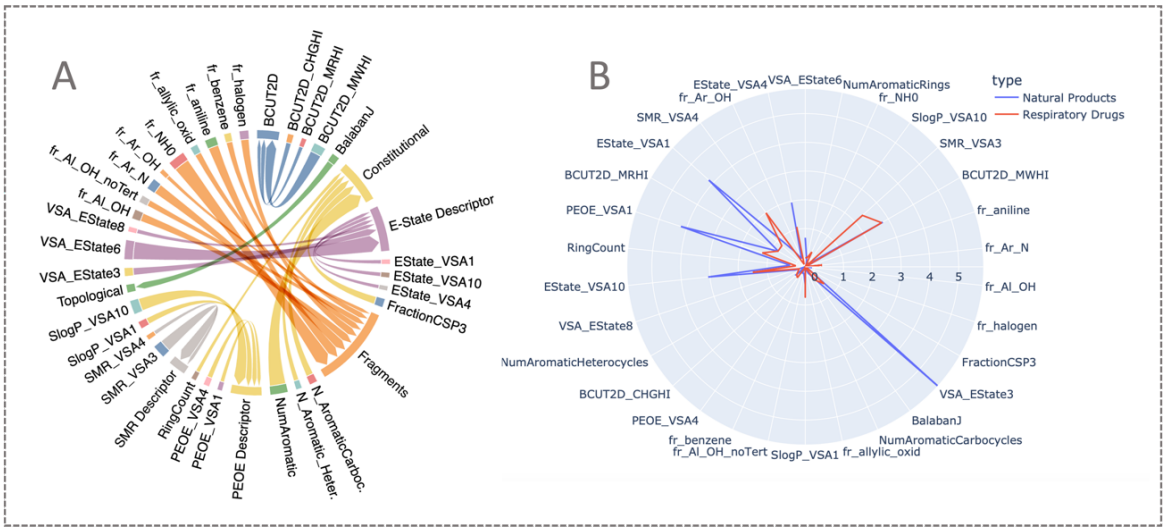
Foi determinado que as moléculas com mais de 50% de similaridade espectral seriam classificadas como potenciais candidatas para o sinal em análise, sendo classificadas como anotações moleculares. Para uma investigação mais detalhada, esse conjunto de anotações estruturais foi avaliado por dois diferentes modelos de inteligência artificial.

O primeiro modelo aplicado foi o Classificador *Light Gradient Boosting Machine* (LGBM Classifier). Este modelo foi treinado para categorizar as moléculas com base em suas características estruturais. Utilizou-se métodos de aprendizado de máquina para identificar padrões em um conjunto de treinamento e, posteriormente, classificar novas moléculas com base nesses padrões identificados. O propósito deste modelo foi fornecer uma classificação precisa das moléculas anotadas, identificando-as como Produtos Naturais ou Moléculas Fármaco-similares.

O segundo modelo foi desenvolvido a partir da fórmula estrutural das moléculas. Usando um algoritmo que converteu os SMILES (Simplified Molecular Input Line Entry System) em imagens de 100x100 pixels, foi possível obter informações visuais das moléculas. Estas imagens foram então analisadas por uma rede neural convolucional, um tipo de modelo de aprendizado profundo especializado em identificar padrões complexos em imagens. O propósito deste modelo foi identificar características moleculares nas imagens e classificá-las, da mesma forma que o modelo de aprendizado de máquina, como produto natural ou molécula-fármaco, com base nas características estruturais aprendidas durante o treinamento.

Para o processo de aprendizado de máquina, inicialmente foram calculados 208 descritores químicos diferentes para as moléculas. No entanto, através de técnicas de seleção de características, esse número foi reduzido para 30, como descrito no capítulo anterior sobre o software LUMIOS (VIEIRA; ALVES DE SOUSA; CASTRO-GAMBOA, 2023).

Figura 13 - Descritores selecionados como features dos modelos de *machine learning* e suas classificações (A). Comparação entre as médias de cada descritor comparados nas duas classes estudadas (B).



Fonte: Elaborado pelo autor (2023).

O modelo foi construído a partir de 13.400 estruturas moleculares, divididas igualmente entre produtos naturais (PNs) e moléculas utilizadas em estágios avançados no combate a doenças respiratórias. Esse conjunto de dados foi obtido compilando informações provenientes das seguintes bases de dados moleculares:

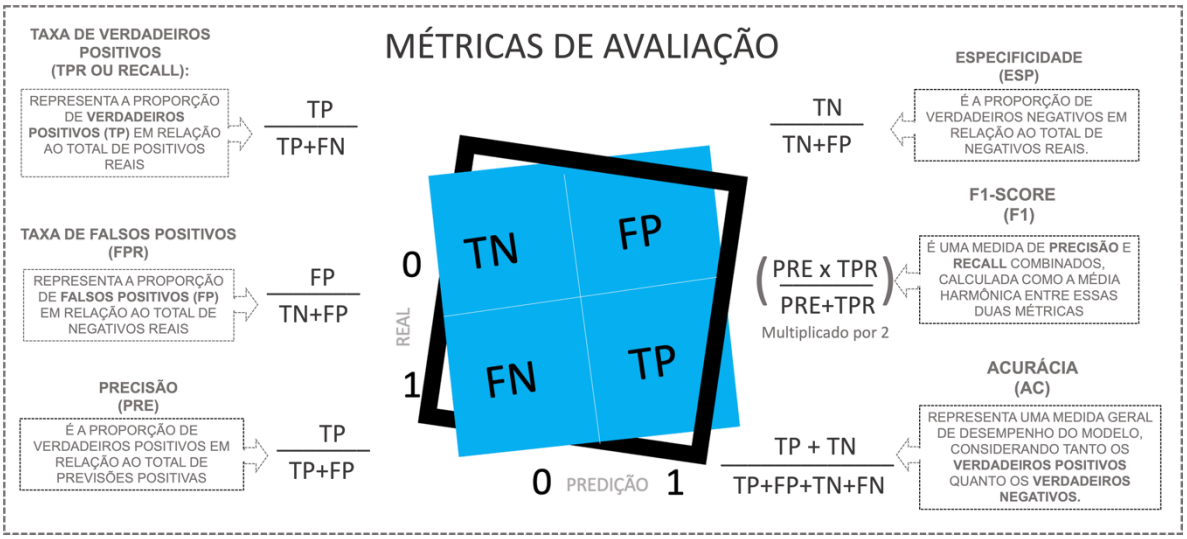
- Lotus (https://lotus.naturalproducts.net): Lotus é atualmente uma das maiores bases de dados moleculares disponíveis (RUTZ et al., 2022b).
- Cortellis (https://www.cortellis.com): Foram obtidas 6.700 estruturas moleculares a partir da plataforma Cortellis, mantida pela Clarivate, e essas moléculas estão sendo testadas para doenças respiratórias.

Vale ressaltar que a mesma molécula pode aparecer em diferentes bases de dados. Por esse motivo, os duplicados foram removidos, sendo atribuídos a apenas uma base de dados.

2.7 Métricas de avaliação

A plataforma LUMIOS utiliza, para avaliação, sete métricas: Taxa de Verdadeiros Positivos (TPR), Taxa de Falsos Positivos (FPR), Precisão, Recall, Especificidade, F1-Score e Acurácia Balanceada. Todas as métricas são descritas na Figura 14:

Figura 14 - Fórmulas utilizadas para cálculos das métricas de avaliação dos modelos



1 - TPR ou sensibilidade, recall, taxa de detecção. TP (Verdadeiros Positivos) representa o número de casos positivos corretamente classificados pelo modelo. FN (Falsos Negativos) representa o número de casos positivos erroneamente classificados como negativos pelo modelo.

2 - FPR – Taxa de Falsos Positivos. FP (Falsos Positivos) representa o número de casos negativos erroneamente classificados como positivos pelo modelo. TN (Verdadeiros Negativos) representa o número de casos negativos corretamente classificados pelo modelo.

3 - Precisão – TP (Verdadeiros Positivos) representa o número de casos positivos corretamente classificados pelo modelo. FP (Falsos Positivos) representa o número de casos negativos erroneamente classificados como positivos pelo modelo.

4 - TN (Verdadeiros Negativos) representa o número de casos negativos corretamente classificados pelo modelo. FP (Falsos Positivos) representa o número de casos negativos erroneamente classificados como positivos pelo modelo.

5- (Precisão) é a proporção de verdadeiros positivos em relação à soma dos verdadeiros positivos e falsos positivos. É uma medida de quão precisas são as predições positivas do modelo. Recall (Recall) é a proporção de verdadeiros positivos em relação à soma dos verdadeiros positivos e falsos negativos. É uma medida de quão completas são as predições positivas do modelo.

6 - Acurácia é uma medida que indica a proporção de predições corretas em relação ao total de predições feitas pelo modelo. Ela fornece uma visão geral do desempenho geral do modelo em todos os resultados, independentemente da classe.

Fonte: Elaborado pelo autor (2023).

2.8 Docagem molecular

As anotações moleculares, resultantes do processo fermentativo do cacau, foram avaliadas em quatro receptores disponíveis na plataforma LUMIOS, nomeadamente: 4DD8, 7P2G, 1NC6 e 6VVU. As proteínas 4DD8, 1NC6 e 6VVU estão associadas a doenças inflamatórias, como alergias e asma (COSTANZO et al., 2003a; HALL et al., 2012a; MAUN et al., 2020a), enquanto a 7P2G é uma protease M, crucial para o ciclo de vida do SARS-CoV-2 (ROSSETTI et al., 2022a).

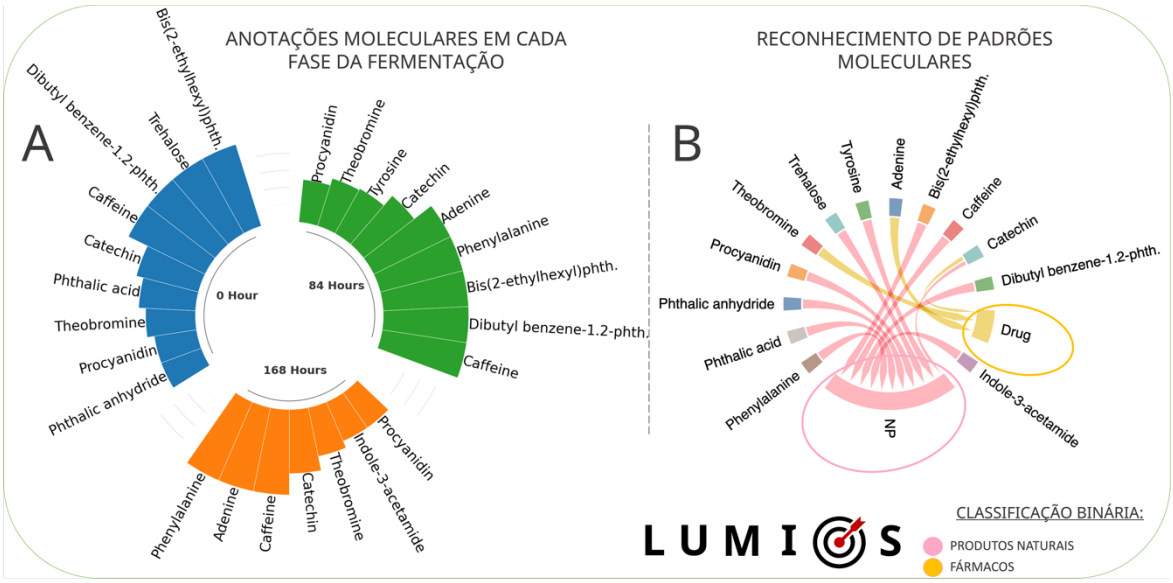
As proteínas foram manipuladas computacionalmente pelo software Chimera (versão 1.16) em um computador equipado com um processador AMD Ryzen 7 5800x, 80 gigabytes de memória RAM e acelerado pelo processamento paralelo da placa de vídeo NVIDIA 3080, operando no sistema Windows 11. Todos os ligantes (identificados pelo software LUMIOS como uma anotação) foram pré-processados e convertidos para o formato .pdbqt. As estruturas das proteínas alvo foram retiradas do banco de dados de proteínas RCSB (https://www.rcsb.org).

3 RESULTADOS E DISCUSSÃO

3.1 Anotações Moleculares e Combinação de ML e DL

No total, 13 moléculas foram sinalizadas como possíveis anotações durante a desreplicação, nas diferentes etapas da fermentação de grãos de cacau (0 horas, 84 horas e 168 horas) (Figura 15) (ao todo, será visualizado 25 moléculas, devido ao fato de algumas aparecem em mais de uma etapa). A plataforma LUMIOS, utilizada na abordagem computacional deste trabalho, disponibiliza um filtro capaz de anotar apenas as estruturas químicas registradas na plataforma LOTUS (https://lotus.naturalproducts.net), garantindo que apenas moléculas rotuladas como PN façam parte da coleção de anotações. Esse filtro compreende mais de 270.000 moléculas de NP, contendo referências bibliográficas e informações químicas relevantes para consultas e exploração de um vasto espaço químico.

Figura 15 – A) Distribuição das anotações em cada etapa da fermentação. B) Classificação de cada anotação pelos modelos de inteligência artificial do LUMIOS.



Fonte: Elaborado pelo autor (2023).

Adicionalmente, foi estabelecido um critério de seleção para as anotações, onde as estruturas químicas classificadas como anotações apresentassem um erro

inferior a cinco (permitindo ajustes) ppm (partes por milhão) quando comparadas às massas exatas determinadas pela espectrometria de massas. A aplicação destes critérios de seleção pode diminuir consideravelmente o universo de anotações químicas consideradas. No entanto, isso resulta em dados mais robustos, valiosos e realistas, em detrimento de sugestões de modificações estruturais ou estruturas sintéticas que não são coerentes com os extratos brutos obtidos pela fermentação dos grãos de cacau.

Portanto, todas as anotações que cumpriram os critérios estabelecidos pelos filtros seletivos de PN foram levadas em conta para as etapas subsequentes do processo de desreplicação. Todas as estruturas foram automaticamente convertidas para o formato SMILES – Simplified Molecular-Input Line-Entry System pelo LUMIOS, e 30 descritores foram calculados a partir de cada anotação molecular, servindo como dados de entrada para o modelo de ML. O algoritmo empregado para classificar essas estruturas foi o LGBM Classifier, que constrói árvores de decisão de maneira aleatória e, ao combinar os resultados, classifica os alvos (KE et al., 2017b).

Outro modelo de classificação molecular, disponível no núcleo de IA do LUMIOS, também foi aplicado às anotações. No entanto, utilizando uma Rede Neural Convolucional (CNN) para reconhecer padrões de moléculas de NP e moléculas com atributos similares a estruturas em estágios avançados de estudos contra alvos biomacromoleculares de doenças do trato respiratório. Nesse caso, a partir do SMILES, o algoritmo transforma uma sequência de caracteres em uma imagem (contendo a fórmula estrutural da molécula). Por fim, ele analisa a própria imagem para reconhecer características moleculares que permitem a separação entre as duas classes analisadas. Por exemplo, o modelo de ML e a CNN apresentaram alta precisão no processo de classificação, com 92% e 90%, respectivamente.

Considerando apenas as diferentes moléculas anotadas durante o processo de fermentação de grãos de cacau, 13 anotações moleculares atribuídas a sinais espectrais foram submetidas ao poder de classificação da CNN. Em resposta, houve indicação de que 11 (de acordo com ML) a 10 moléculas (DL) possuíam

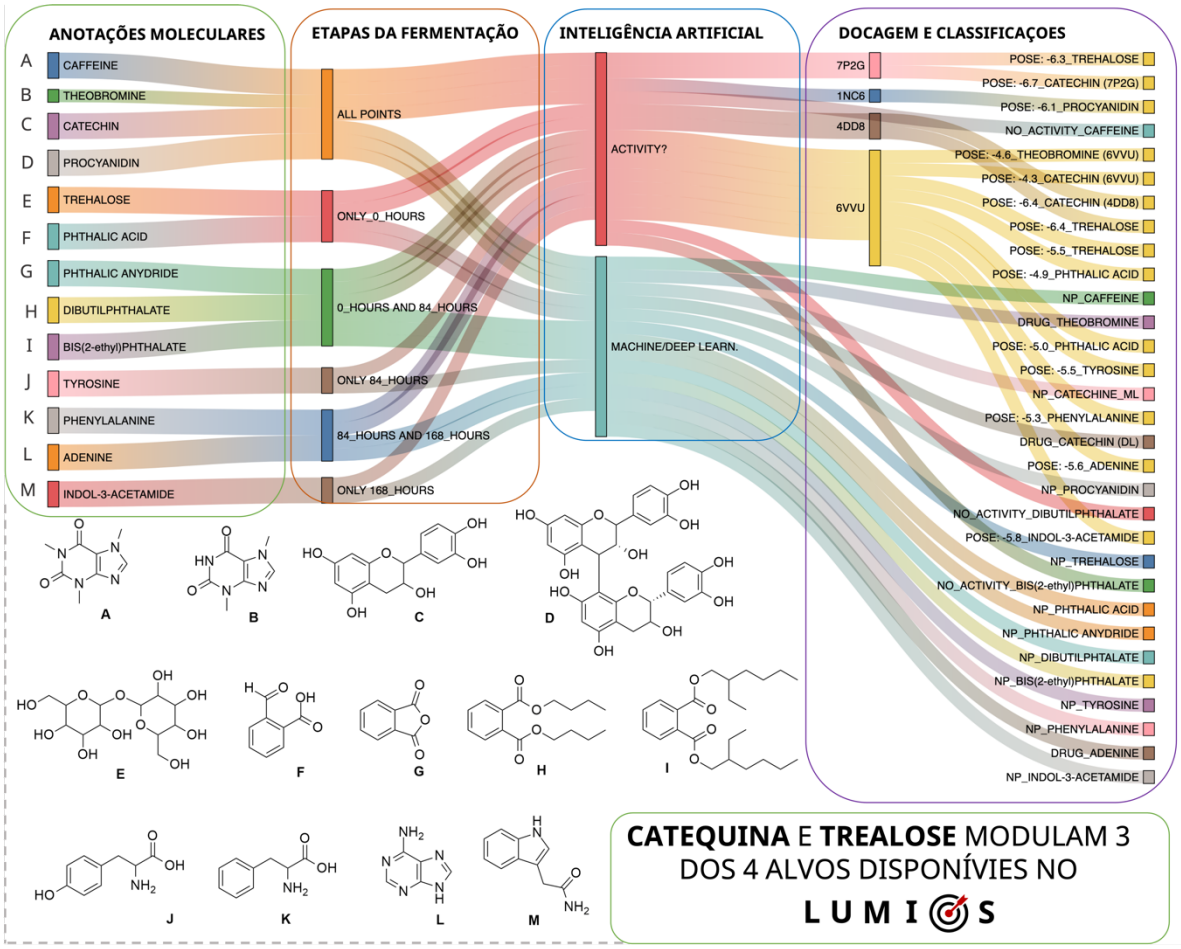
características exclusivas de PN, enquanto as demais apresentaram características de moléculas que atuam em doenças do trato respiratório. A Figura 15-B contempla a atribuição classificatória efetuada pelos dois modelos de IA.

3.2 Docagem molecular

A quantidade expressiva de dados gerados e a crescente demanda por processamento rápido e técnicas eficientes e inteligentes de mineração estimularam o surgimento de plataformas web, que proporcionam praticidade e versatilidade a estudos robustos em ambientes computacionais (DONG et al., 2007).

Dessa forma, explorando as funcionalidades da plataforma LUMIOS, estudou-se possíveis interações biológicas relacionadas às anotações por meio de avaliações *in silico* utilizando conceitos de docagem molecular para avaliar as modulações dos complexos formados entre proteínas e ligantes (DA SILVEIRA et al., 2019). O objetivo foi verificar a compatibilidade estrutural de todas as anotações em relação a quatro alvos biomacromoleculares relacionados a doenças do trato respiratório. Além dos resultados da docagem, indicações de padrões moleculares foram associadas às anotações obtidas. A compilação das indicações moleculares e dos resultados de docagem pode ser vista na Figura 16.

Figura 16 - Resultados das classificações das anotações moleculares por meio de ML, docagem e DL.



Fonte: Elaborado pelo autor

Neste ponto, pressupõe-se que as 13 estruturas descritas acima possam ser atribuídas aos sinais moleculares explorados, sendo consideradas, portanto, anotações moleculares. Quatro anotações podem causar divergências nesse processo de desreplicação, pois pertencem à classe dos ftalatos. A próxima subseção apresentará considerações sobre todas as classes de anotações, subdividindo-as em grupos químicos.

3.3 Trealose

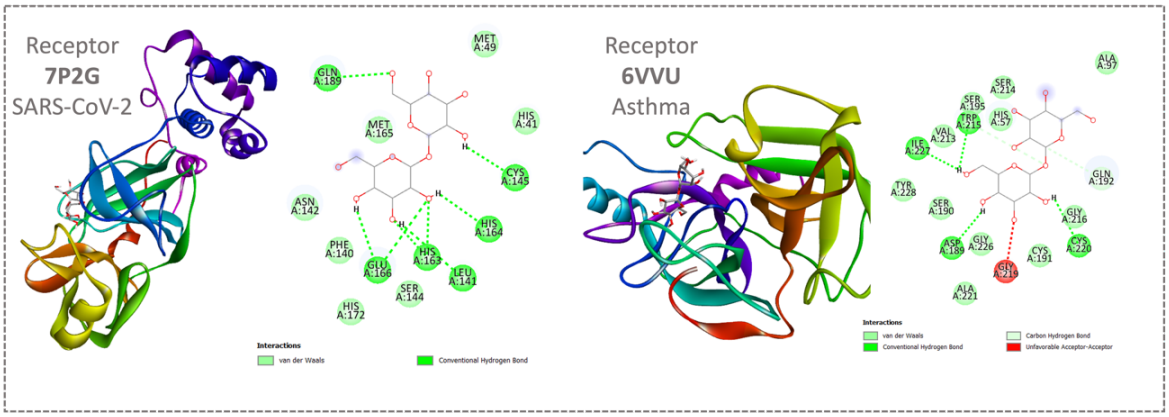
O algoritmo utilizado para a desreplicação molecular indicou que o sinal espectral da relação massa/carga (m/z) 365,1056, no modo de ionização positiva ($[M+Na]^+$), pode ser associado à molécula de Trealose. Essa identificação ocorre exclusivamente no início do processo de fermentação (0 horas). A trealose (2R,3S,4S,5R,6R)-2-(hydroxymethyl)-6-[(2R,3R,4S,5S,6R)-3,4,5-trihydroxy-6-(hydroxymethyl)oxan-2-yl]oxyoxane-3,4,5-triol), um dissacarídeo não redutor com dois resíduos de glicose na configuração alfa no carbono anomérico, ocorre naturalmente e está presente em plantas e algas (LUNN et al., 2014), fungos, leveduras e bactérias (FRANÇOIS; PARROU, 2001; SCHIRALDI; DI LERNIA; DE ROSA, 2002), insetos (WANG et al., 2021) e outros invertebrados (DMITRYJUK; ŁOPIEŃSKA-BIERNAT; FARJAN, 2009). Em micro-organismos e plantas superiores, está relacionada ao fato de proporcionar tolerância à seca e aumento de temperatura, induzindo a superexpressão de enzimas e reduzindo o estresse ambiental (GARG et al., 2002; SCHIRALDI; DI LERNIA; DE ROSA, 2002).

Lima e colaboradores (LIMA et al., 2011) relataram a ocorrência abundante de genes que catalisam a síntese de trealose durante a fermentação dos grãos de cacau. Os autores identificaram a presença dos genes (1-4)-α-$_D$-glicano 1-α-$_D$-glicosilmutase e 4-α-$_D$-{(1-4)- α-$_D$-glicano}trealose trealohidrolase, possivelmente relacionados à resposta ao estresse ambiental (osmótico e térmico) ao qual a cocobiota foi submetida no processo de fermentação das sementes de cacau.

Várias atividades biológicas são atribuídas à molécula de trealose, como antitumoral (MCLAUGHLIN et al., 1978; NOLIBE et al., 1986), ação adjuvante contra tuberculose (HOLTEN-ANDERSEN et al., 2004), antibacteriana (PARANT et al., 1977), capacidade de reduzir granulomas (YARKONI; BEKIERKUNST, 1976) e angiogênese (SAITA et al., 2000), além de estudos contra o SARS-CoV-2 (MARTINON et al., 2020a). Essas informações corroboram com os resultados obtidos no estudo de docagem molecular. A trealose apresentou resultados positivos para as proteínas relacionadas à asma 6VVU (com -5,5 kcal/mol) e 4DD8 (-6.4 kcal/mol) , além de resultados promissores com o alvo 7P2G (-6,3 kcal/mol),

uma proteína M-Pro do SARS-CoV-2. A comparação de afinidade é sempre estabelecida perante os padrões originalmente co-cristalizados com os alvos biomacromoleculares (Tabela 4 apresentada no Capítulo 1). Essa alta afinidade se deve ao número de átomos de oxigênio em sua estrutura, que permitem a formação de ligações de hidrogênio, com destaque principalmente para interações com glutamina-166, histidina-163-164, cisteína-145, leucina-141 e glicina-189 (Figura 17), conferindo o potencial de atuar em alvos biomacromoleculares. Além disso, modelos de IA indicaram que a estrutura da trealose apresenta padrões de moléculas classificadas como NP. Esses dados foram adicionados aos resultados da docagem molecular, indicando que esse NP pode ser um candidato a fármaco.

Figura 17 - Resultados oriundos das testagens de docagem molecular para a molécula de trealose, a qual apresentou afinidade pelo receptor 7P2G (A) e 6VVU (B).



Fonte: Elaborado pelo autor (2023).

3.4 Grupo indólico

A presença do íon precursor $[M+H]^+$ com m/z 175,0857 foi sinalizada como sendo a molécula de indol-3-acetamida. Esse sinal aparece exclusivamente na fase final do processo, após um período de 168 horas de fermentação. A molécula pertence à categoria de indol é substituída por um grupo 1H-indol-3-il na posição 2. Esta estrutura está relacionada à síntese de hormônios vegetais, entre os quais o ácido indol-3-acético (IAA) se destaca, desempenhando um papel essencial no crescimento e desenvolvimento de plantas, conforme detalhado por Tsavkelova et

al. (2012) a identificação dessa molécula já foi relatada em fungos (KULKARNI et al., 2013) e em bactérias (LIN; XU, 2013), e está fortemente relacionada ao sinergismo entre micro-organismos e plantas (DUCA et al., 2014). Tanto em bactérias quanto em plantas, o triptofano foi identificado como o principal precursor para a síntese de IAA, sendo a via metabólica para a síntese do intermediário indol-3-acetamida a mais bem caracterizada (DUCA; GLICK, 2020). Rottiers e colaboradores (2019) identificaram, no início do processo de fermentação das sementes de cacau, altas concentrações de triptofano, que serviriam como substrato para vias metabólicas secundárias para a síntese de compostos derivados do indol (AGYIRIFO et al., 2019; ZUMAETA et al., 2022).

Ao reconhecer padrões estabelecidos por modelos de IA, a molécula indol-3-acetamida apresenta características moleculares de PN e mostrou resultados positivos na modulação da proteína 6VVU relacionada à asma. A afinidade pelo alvo biomacromolecular foi de -5,8 kcal/mol, com interações do tipo ligação de hidrogênio entre o ligante e os aminoácidos asparagina-189 e serina-190-195 (Figura Suplementar 23)

3.5 Grupo dos aminoácidos e derivados purínicos

Os íons precursores $[M+H]^+$ em m/z 166,0852, 182,0803 e 136,0610 foram atribuídos a dois aminoácidos, fenilalanina e tirosina, e um derivado de purina, a adenina. No contexto dos derivados de purina, tais moléculas desempenham um papel intermediário na síntese de ácidos nucleicos e cofatores como nicotinamida adenina dinucleotídeo (NAD), fosfato de nicotinamida adenina dinucleotídeo (NADP), dinucleotídeo de flavina e adenina (FAD) e coenzima A. Além disso, desempenham um papel crucial na transdução de energia em processos metabólicos, conforme discutido por Chapman e Atkinson (1977).

Além disso, devido à complexidade das matrizes resultantes do processo de fermentação do cacau, relata-se que uma pequena quantidade de adenina pode ser convertida em teobromina, um dos principais compostos marcadores do cacau, por meio de 7-metilxantina ou 3-metilxantina (KOYAMA et al., 2003)

Por outro lado, os aminoácidos apontados como anotações moleculares podem estar relacionados com o aroma e sabor do chocolate, principal produto do cacau, como destacado por Lima e equipe (LIMA et al., 2011). Estudos indicam que, durante o primeiro dia de fermentação, há uma redução nos aminoácidos ácidos, mas há um aumento progressivo nos demais, especialmente fenilalanina, conforme Brunetto et al. (2020b).

Dentro das classificações moleculares realizadas pelos modelos de aprendizado de máquina, os dois aminoácidos foram identificados como produtos naturais, enquanto a adenina foi associada a características semelhantes às moléculas estudadas em pesquisas avançadas sobre doenças respiratórias. Além disso, considerando os resultados de docagem molecular, todas as estruturas deste conjunto demonstraram a capacidade de modular a proteína associada à asma, 6VVU, com afinidades (energia de ligação) de -5,3 kcal/mol (fenilalanina), -5,5 kcal/mol (tirosina) e -5,8 kcal/mol (adenina). As diversas interações entre o ligante e os aminoácidos da proteína 6VVU estão ilustradas nas Figura Suplementar 20 – Figura Suplementar 22, disponíveis no material suplementar A.

3.6 Grupo dos flavonoides

Os íons precursores $[M+H]^+$ com valores de m/z 579,1443 e 291,0871 mostraram fragmentação típica de substâncias pertencentes à classe dos flavonoides, sendo identificados como procianidina e catequina, respectivamente. Esses polifenóis são encontrados em todas as etapas do processo de fermentação do cacau, sendo compostos que contribuem significativamente para diversos setores industriais, como farmacêuticos e alimentícios.

Essas estruturas, juntamente com as xantinas, são marcadores moleculares de *Theobroma cacao*, conforme indicado por Gallego et al. (2021). A ingestão regular desses compostos tem potencial antioxidante, pode ajudar a reduzir a pressão arterial (KHAN et al., 2014a), aumentar a concentração (NEHLIG, 2013) e desempenhar um papel em ações anticâncer, anti-inflamatórias, antivirais,

imunossupressoras, além de combater doenças crônicas e distúrbios metabólicos (DASIMAN et al., 2022).

Além disso, compostos fenólicos têm sido objeto de investigações em pesquisas relacionadas a doenças do sistema respiratório. Evidências indicam que o consumo de catequina pode atenuar tanto o início quanto a intensidade dos sintomas respiratórios durante o período de inverno (OZATO et al., 2022).

A catequina foi identificada pelo modelo de aprendizado de máquina como tendo descritores moleculares semelhantes a produtos naturais, enquanto o modelo de rede neural artificial apontou padrões estruturais indicativos de moléculas candidatas a fármacos. Por outro lado, a procianidina foi classificada por ambos os modelos de inteligência artificial como tendo padrões de produto natural. Além disso, os resultados dos estudos de docking molecular indicaram que ambas as estruturas podem modular efetivamente alvos biomacromoleculares de doenças respiratórias.

A catequina exibiu afinidade entre o ligante e a proteína no alvo SARS-CoV-2 (7P2G) com uma energia de ligação de -6,7 kcal/mol, e nos alvos de asma (6VVU e 4DD8) com energias de ligação de -4,3 kcal/mol e -6,4 kcal/mol, respectivamente. Esses resultados para o alvo SARS-CoV-2 estão alinhados com as descobertas de MAJUMDER e MANDAL (2022), que conduziram estudos de docagem e dinâmica molecular com compostos polifenólicos e encontraram respostas promissoras de afinidade ligante-proteína. A procianidina demonstrou uma energia de ligação de -6,1 kcal/mol no alvo 1NC6 (asma), sendo a única estrutura identificada nas anotações capaz de modular esse receptor, apresentando resultados superiores ao ligante original cristalizado com a proteína. O material suplementar ilustra as interações intermoleculares entre os flavonoides e os receptores (Figura Suplementar 15 e Figura Suplementar 16).

3.7 Grupo das xantinas

A cafeína e a teobromina foram associadas a íons de moléculas protonadas com m/z 195,0888, 181,0707, respectivamente. As metilxantinas estão presentes

em todas as etapas da fermentação do cacau. Essa informação está de acordo com a revisão de Matissek (1997) sobre xantinas, enfatizando que a teobromina predomina nos extratos, corroborando com os achados de Cortez e equipe (2023).

A cafeína desperta interesse na indústria farmacêutica devido às suas propriedades estimulantes sobre o sistema nervoso central (CAMANDOLA; PLICK; MATTSON, 2019), aumentando as funções motoras, mas reduzindo o apetite e a fadiga (TARKA; CORNISH, 1982). Por outro lado, a molécula de teobromina tem sido alvo de investigação devido às suas propriedades antitussígenas e broncodilatadoras (SIMONS et al., 1985), bem como sua ação diurética (DORFMAN; JARVIK, 1970), e redução de tumores (FREDHOLM; SMIT, 2011), além de apresentar potencial para doenças inflamatórias (LEE; CHOI; HA, 2022).

Quando submetida ao poder classificatório dos algoritmos de IA, a cafeína é identificada como produto natural (PN) e a teobromina como fármaco-similar. Além disso, os resultados de docagem molecular mostraram que a cafeína não apresentou resultados significativos para afinidade de ligação às proteínas. Ao mesmo tempo, a teobromina foi capaz de modular a proteína da asma (6VVU), com uma afinidade de -4,6 kcal/mol (teobromina), corroborando com autores previamente citados, que apontam a teobromina como molécula aplicável em ações contra asma e outros problemas respiratórios, como tosse. As interações intermoleculares da teobromina (que apresentou afinidade maior que o ligante co-cristalizado) é retratada na Figura Suplementar 14.

3.8 Grupo dos ftalatos. Produtos Naturais ou contaminantes?

Ésteres do ácido ftálico (sigla em inglês: PAEs) são usados como plastificantes (EALES et al., 2022; NET et al., 2015) em diferentes materiais poliméricos, que, quando adicionados a produtos industriais, aumentam sua flexibilidade, durabilidade, longevidade e transparência. Estudos bem estabelecidos indicam que a exposição a PAEs desencadeia impactos relevantes na saúde humana, como comprometimento dos sistemas reprodutivo, cardiovascular,

respiratório e endócrino, além de distúrbios neurológicos e da tireoide (EALES et al., 2022).

Por muito tempo, acreditava-se que os ésteres do ácido ftálico, conhecidos como ftalatos, eram moléculas exclusivamente antropogênicas (ZHANG et al., 2018). Testes biológicos em animais usando moléculas da classe dos ftalatos, como Di-2-etilhexilftalato (DEHP), mostraram resultados positivos para disfunção do esperma, incluindo redução da contagem e motilidade de espermatozoides (PAN et al., 2006). No entanto, embora existam indicações nesse sentido, muitos outros estudos explorando as relações entre ftalatos e o sistema reprodutivo tiveram impacto limitado (LAMBROT et al., 2009) e persistem com inconsistências em testes de toxicidade humana. Além disso, estudos demonstraram a origem orgânica dos ftalatos naturais e descartaram a possibilidade de que tais moléculas estejam associadas à ubiquidade de contaminantes ambientais provenientes de ftalatos químicos (ROY, 2020a; ROY; SEN, 2013).

As amostras analisadas neste trabalho provêm de extratos da fermentação espontânea de sementes de cacau. A fermentação das sementes de cacau é um processo complexo que envolve a ação de uma microbiota diversificada, que inclui a sucessão de grupos microbianos de diferentes espécies de leveduras, bactérias láticas (LAB), bactérias acéticas (AAB) (DE VUYST; LEROY, 2020) e espécies do gênero *Bacillus* (FIGUEROA-HERNÁNDEZ et al., 2019).

No entanto, a presença de fungos filamentosos neste processo ainda não foi adequadamente esclarecida (DELGADO-OSPINA et al., 2021). Esta microdiversidade produz a maioria dos metabólitos encontrados em exsudatos de cacau fermentado. Como se trata de uma sucessão microbiana, alguns micro-organismos aparecem em diferentes etapas da fermentação, possibilitando estabelecer uma relação com os metabólitos obtidos no processo de dereplicação.

Em nossa exploração, o ácido ftálico (167,0342 m/z) e o anidrido ftálico (149,0239 m/z) foram anotados na fase inicial da fermentação do cacau. Na primeira etapa da fermentação, a atividade das leveduras é predominante, de diferentes espécies, que atuam na metabolização da glicose, frutose e sacarose, presentes na polpa do cacau, para a formação de etanol e dióxido de carbono, conforme descrito

por Agyirifo et al. (2019). *S. cerevisiae* é uma espécie de levedura frequentemente detectada nesta fase de fermentação, conforme indicado por Gutiérrez-Ríos et al. (2022). Ela exibe a capacidade de sintetizar éster dioctildecil-1,2-benzenodicarboxilato em um meio de cultura contendo glicose como fonte primária de carbono, conforme relatado por Abdel-Kareem, Rasmey e Zohri (2019). Assim, a presença dessas moléculas nos extratos pode estar relacionada à atividade microbiana desta espécie de levedura. Além do metabolismo das leveduras, o éster bis(2-metilpropil)-1,2-benzenodicarboxilato já foi sinalizado como um fitoconstituinte em extratos da casca do fruto do cacau, de acordo com Yahya, Ginting e Saidi (2021).

Após 84 horas do processo de fermentação, não há sinal de referência atribuído ao ácido ftálico, mantendo-se o anidrido ftálico e aparecendo duas novas estruturas: dibutilftalato (279,1612 m/z) e bis-(2-etilhexil)ftalato (391,2853 m/z). Esses compostos podem ser produzidos devido a uma reação de esterificação do ácido ftálico (presente no processo inicial de fermentação) com vários álcoois, conforme descrito por Ortiz e Sansinenea (2018). Adicionalmente, estudos relataram a capacidade de *Bacillus subtilis* e *Bacillus pumilus*, micro-organismos já detectados na cocobiota, de sintetizar bis-(2-etilhexil)ftalato, por exemplo (LOTFY et al., 2018; MOUSHUMI PRIYA; JAYACHANDRAN, 2012). Na fase final da fermentação, as técnicas de dereplicação propostas não identificaram a presença de ftalatos nos extratos. Esta descoberta pode ser atribuída à possibilidade de degradação microbiana de diferentes classes de ftalatos em condições óxicas e anóxicas, conforme apontado na mini revisão conduzida por Boll et al. (2020).

Quanto às potenciais ações dos ftalatos em atividades biológicas, Thiemman, em 2021 relata através de uma revisão de centenas de publicações envolvendo diferentes tipos de ftalatos isolados de PNs que não relacionam a ação antropogênica a tais compostos. Além disso, são relatadas informações sobre as atividades biológicas desses compostos, como atividade antitumoral, larvicida, antimicrobiana, anti-inflamatória, antiviral e antidiabética, entre outras (THIEMANN, 2021a).

Quando submetidas a modelos de ML e DL, todas as moléculas de ftalato mostraram padrões de PNs, e quando aplicadas em simulação computacional usando ferramentas de docagem molecular, apenas o ácido ftálico e o anidrido ftálico tiveram resultados positivos na modulação da proteína 6VVU, que está relacionada à asma, com -4,9 kcal/mol para ácido ftálico, e -5,0 kcal/mol para anidrido ftálico. Nas Figura Suplementar 18 e Figura Suplementar 19 são retratadas as interações intermoleculares responsáveis pelo potencial significativo dos ftalatos para aplicações além de plastificantes, merecendo uma exploração racional mais detalhada (DRICHE et al., 2015; HABIB; KARIM; OTHERS, 2012; HOANG; LI; KIM, 2008; QIAN; KANG; KIM, 2012; ROY, 2020b).

Neste capítulo, para cada anotação molecular examinada, buscou-se correlacionar as indicações de desreplicação realizadas pelo software LUMIOS com os mecanismos de fragmentação de cada espectro, com objetivo de reforçar as evidências de que os sinais moleculares dos extratos de cacau estão associados a essas moléculas. Os mecanismos específicos podem ser visualizados nas figuras suplementares (Figura Suplementar 1 a Figura Suplementar 13).

4 CONCLUSÃO

Neste trabalho, a utilização do algoritmo LUMIOS para desreplicação molecular revelou a identificação de 13 estruturas moleculares distintas, incluindo moléculas de ftalato, nas matrizes complexas de cacau. Estas moléculas, potencialmente associadas ao intrincado processo de fermentação do cacau, exibiram relevante afinidade por proteínas associadas a doenças respiratórias, quando avaliadas por modelos de aprendizado de máquina e docking molecular. Foi evidenciado que a maioria das moléculas anotadas (10 de 13) apresentaram interações significativas com proteínas inseridas no núcleo de docagem do LUMIOS. Notavelmente, algumas dessas estruturas moleculares destacam-se pela capacidade de modular proteínas relacionadas a condições de saúde relevantes, incluindo aquelas associadas à SARS-CoV-2 e desordens respiratórias. Este estudo, portanto, não só valida a eficácia do algoritmo LUMIOS na desreplicação molecular, mas também sublinha o potencial terapêutico de compostos derivados do cacau, abrindo caminhos promissores para o desenvolvimento de novos medicamentos, especialmente no contexto de desafios globais de saúde, como a recente pandemia de coronavírus.

CAPÍTULO 3 – CHEMISTIKA: FERRAMENTA PARA AUTOMATIZAÇÃO E APLICAÇÕES DE PLANEJAMENTO DE MISTURAS DO TIPO DE SIMPLEX-LATTICE ENVOLVENDO DADOS DE ESPECTROMETRIA DE MASSAS (LC-MS)[6]

RESUMO

Chemistika é uma plataforma inovadora para analisar produtos naturais (PNs) automatizando o planejamento de misturas do tipo Simplex-Lattice (Simplex-Lattice Design (em inglês – SLD). É uma plataforma desenvolvida para análise estatística, consolidada como uma interface de programação de aplicativos (API). Reconhece anotações moleculares em matrizes naturais complexas e atribui suas respectivas intensidades de sinal quando acopladas às saídas da API LUMIOS. Para demonstrar a funcionalidade da API Chemistika, foram utilizados extratos brutos de exsudatos de sementes de cacau não fermentados (*Theobroma cacao* L.). Modelos matemáticos obtidos pela análise estatística do SLD, que diagnosticam a composição ideal da mistura de solventes para a extração da anotação molecular de trealose, foram criados pela API Chemistika. Ao se acoplar à API LUMIOS, desenvolvida pelo mesmo grupo de pesquisa, a API Chemistika pode combinar informações sobre a presença desta anotação molecular nas amostras e atribuir sua respectiva intensidade de sinal. Os designs experimentais priorizaram a extração de trealose, que tem afinidade por alvos biomacromoleculares de doenças respiratórias como SARS-CoV-2 e asma. Ao contrário de softwares comerciais, limitados a análises estatísticas, a API Chemistika pode oferecer uma abordagem eficiente, rápida, precisa e de acesso aberto para investigar bioativos presentes em matrizes naturais complexas e automatizar esse processo através de modelos matemáticos.

Palavras-chave: anotações moleculares, design experimental, modelos estatísticos.

[6] O artigo *CHEMISTIKA: TOOL FOR AUTOMATION AND APPLICATIONS IN THE SIMPLEX-LATTICE DESIGN INVOLVING LC-MS DATA* referente a este capítulo foi submetido à Revista **EXPERT SYSTEMS WITH APPLICATIONS**, e está sob avaliação.

1 INTRODUÇÃO

A análise de produtos naturais (PNs) e a investigação de moléculas presentes em matrizes complexas têm sido campos de pesquisa essenciais para a descoberta de novos compostos bioativos com potenciais aplicações na área da saúde (ATANASOV et al., 2021). No entanto, esse processo de investigação pode ser trabalhoso e requer planejamento cuidadoso.

Neste cenário, os designs experimentais surgem como uma alternativa, constituindo ferramentas altamente eficientes para explorar matrizes naturais complexas. O uso do design experimental em química, especialmente ao explorar matrizes complexas de PNs, oferece inúmeras vantagens significativas, permitindo uma abordagem estruturada e robusta para coleta de dados em experimentos químicos. Também possibilita a construção de modelos capazes de otimizar o uso de recursos, como tempo, materiais e reagentes, permitindo assim uma experimentação mais rápida e econômica (AZCARATE; PINTO; GOICOECHEA, 2020; FREIESLEBEN; KEIM; GRUTSCH, 2020).

Além disso, permite a investigação de vários fatores simultaneamente e a avaliação de sua influência nos resultados químicos desejados (MONTGOMERY, 1992; MYERS et al., 2004). Os designs experimentais também permitem uma análise mais abrangente dos efeitos de diferentes variáveis, facilitando a detecção de possíveis interações entre os fatores estudados (SALMASO et al., 2022), visando entender, por meio de cálculos estatísticos, como as variáveis influenciam umas às outras e compreender, assim, as relações complexas presentes em matrizes de PNs (VIEIRA et al., 2021).

Adicionalmente, os designs experimentais podem ser aprimorados ao serem vinculados a abordagens computacionais, que permitem uma análise detalhada de dados químicos, além da exploração e entendimento adequado de dados espectrais para extrair informações significativas (ARBORETTI et al., 2022). Assim, a plataforma Chemistika foi desenvolvida, como uma solução de Interface de Programação de Aplicativos (*Application Programming Interface – API*), para automatizar o Design Simplex-Lattice (SLD) com a possibilidade de analisar

intensidades de sinal molecular obtidas de espectros de massa de matrizes naturais complexas, ao se acoplar à plataforma LUMIOS (VIEIRA; ALVES DE SOUSA; CASTRO-GAMBOA, 2023). Desta forma, a API Chemistika pode combinar informações sobre as anotações moleculares nas amostras e atribuir suas respectivas intensidades de sinal, possibilitando a construção de modelos estatísticos que fornecem dados valiosos sobre tais moléculas. A plataforma oferece uma abordagem eficiente, rápida e precisa para investigar bioativos em matrizes naturais complexas, automatizando esse processo.

A construção de APIs foi facilitada pelo aumento da qualidade e velocidade da internet e abriu portas para a comunicação e integração entre diferentes sistemas e aplicações (HESHMATISAFA; SEPPÄNEN, 2023). Estas plataformas são geridas por interfaces padronizadas que permitem que diferentes sistemas troquem informações e interajam eficientemente, possibilitando o compartilhamento de dados (EKINS; CLARK; WILLIAMS, 2012), criando aplicações e permitindo reutilização de código (SWAINSTON et al., 2016). As APIs também promovem a consolidação de uma comunidade de usuários que podem explorar cooperativamente um determinado ecossistema computacional (ONG et al., 2015), o que impulsiona a inovação e permite aos desenvolvedores criar novas soluções e serviços com base em funcionalidades existentes e úteis. Ao fornecer acesso a recursos e dados específicos, as APIs incentivam a criação de novas ideias e o desenvolvimento de produtos de maneira rápida e simples.

Com base em uma abordagem inovadora, este artigo visa demonstrar o potencial e a funcionalidade da API Chemistika. Para esse fim, utilizou-se dados prospectados do estudo metabólico de extratos brutos de exsudatos de sementes de cacau não fermentados (*Theobroma cacao* L.). Um modelo que diagnosticou a composição da mistura mais adequada para a extração da anotação molecular de trealose foi criado atribuindo as intensidades relativas aos sinais desta anotação. O modelo criado por esta abordagem forneceu informações valiosas sobre as características químicas da trealose, presente na polpa do cacau, que posteriormente seria submetida ao processo de fermentação espontânea.

Atualmente, aplicações comerciais como Statistica (HILBE, 2007) e Minitab (ALIN, 2010) podem tratar estatisticamente dados obtidos pelo design simplex-lattice. No entanto, a API Chemistika, ao contrário das opções comerciais, é capaz de integrar automaticamente dados espectrais e realizar análises químicas exploratórias de extratos de amostras naturais complexas para obter insights significativos.

2 METODOLOGIA

2.1 Design de Lattice-Simplex (SLD)

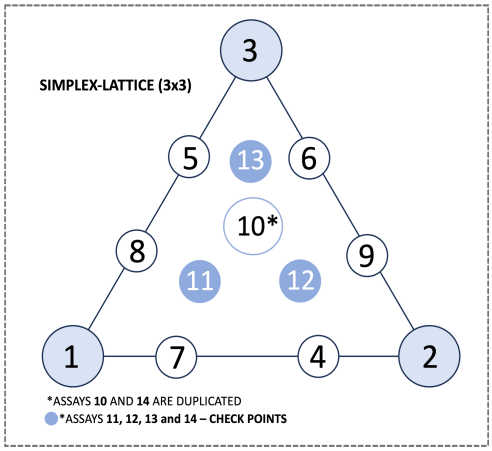
Para extrair as moléculas presentes na polpa das sementes de cacau não fermentadas, foram utilizadas misturas padrão de solventes estabelecidas pelo SLD 3x3 para três variáveis dependentes, quando a proporção total dos três componentes era um. Os procedimentos de extração envolveram macerações a frio e sequenciais sob diferentes misturas de extração (1:3 m/v). Os componentes da mistura eram: hexano (A), acetato de etila (B) e etanol (C). As variáveis de resposta do planejamento experimental foram as intensidades relativas das anotações moleculares geradas pelo processo de dereplicação da plataforma LUMIOS. O número de execuções experimentais e as proporções dos três fatores (A, B e C) foram descritas em termos de variáveis codificadas (0 e +1 para os valores mínimo e máximo, respectivamente) e não codificadas (0 a 45 mL de misturas de solvente), conforme a Tabela 6. O planejamento da mistura utiliza experimentos previamente selecionados para obter os modelos matemáticos. Os dez primeiros experimentos formaram a base do SLD, e os quatro testes restantes (11 a 14) são experimentos de verificação e avaliação do modelo (Figura 18).

Tabela 6 - Layout do planejamento Simplex-Lattice para diferentes misturas de solventes

EXP	VARIÁVEIS CODIFICADAS				VARIÁVEIS DESCODIFICADAS (mL)			
	COMPONENTES			Proporção Total	COMPONENTES			Proporção Total
	A: Hexano	B: Ac. Etila	C: Etanol		A: Hexano	B: Ac. Etila	C: Etanol	
1	1.00	0.00	0.00	1.00	45	0	0	45
2	0.00	1.00	0.00	1.00	0	45	0	45
3	0.00	0.00	1.00	1.00	0	0	45	45
4	0.33	0.67	0.00	1.00	15	30	0	45
5	0.33	0.00	0.67	1.00	15	0	30	45
6	0.00	0.33	0.67	1.00	0	15	30	45
7	0.67	0.33	0.00	1.00	30	15	0	45
8	0.67	0.00	0.33	1.00	30	0	15	45
9	0.00	0.67	0.33	1.00	0	30	15	45

10	0.33	0.33	0.33	1.00	15	15	15	45
11	0.67	0.17	0.17	1.00	30	7.5	7.5	45
12	0.17	0.67	0.17	1.00	7.5	30	7.5	45
13	0.17	0.17	0.67	1.00	7.5	7.5	30	45
14	0.33	0.33	0.33	1.00	15	15	15	45

Figura 18 - Visualização gráfica da disposição experimental criada através do planejamento de misturas do tipo Simplex-Lattice (3x3).



Fonte: Elaborado pelo autor (2023).

2.2 Pré-processamento de dados espectrais dos extratos de polpa de sementes de cacau não fermentados[7]

Os dados brutos foram importados e processados usando o aplicativo MS-DIAL (versão 4.9.2). O pré-processamento dos dados foi aplicado para otimizar a análise de espectros de massa, que consistiu nas seguintes etapas:

a) **Modo positivo:** Os espectros de massa foram adquiridos em modo positivo para identificar moléculas com uma massa molecular de até 1000 Dalton (Da);

b) **Intervalo de tempo de retenção:** O tempo de retenção foi definido entre 0 e 15 minutos, permitindo a detecção de moléculas dentro dessa faixa;

[7] A menos que seja especificamente indicado de outra forma, todos os dados explorados nos capítulos subsequentes foram pré-processados utilizando os mesmos parâmetros.

c) **Tolerância de processamento:** para o processamento de informações espectrais em MS1, foi definida uma tolerância de 0,01 Dalton, e para MS2, uma tolerância de 0,02 Dalton;

d) **Carga máxima permitida:** A carga máxima permitida para moléculas foi definida para uma protonação;

e) **Detecção de pico:** Foi estabelecido um critério para detecção de picos oriundos da espectrometria de massas, com uma largura mínima de 5 unidades e uma intensidade mínima de 1000;

f) **Identificação de adutos:** O aplicativo permitiu a identificação de adutos de sódio (Na+) e potássio (K+);

g) **Alinhamento de espectros:** Os espectros foram alinhados com base em um espectro em branco, permitindo um desvio de 0,05 min no tempo de retenção e 0,01 em MS1. Os sinais do branco foram removidos de cada espectro;

h) **Extração e exportação de dados espectrais:** Após o pré-processamento, os espectros alinhados que continham dados MS/MS foram extraídos no formato .txt. Esses arquivos continham as informações necessárias para alimentar a API Chemistika.

2.3 DESENVOLVIMENTO DO APP CHEMISTIKA

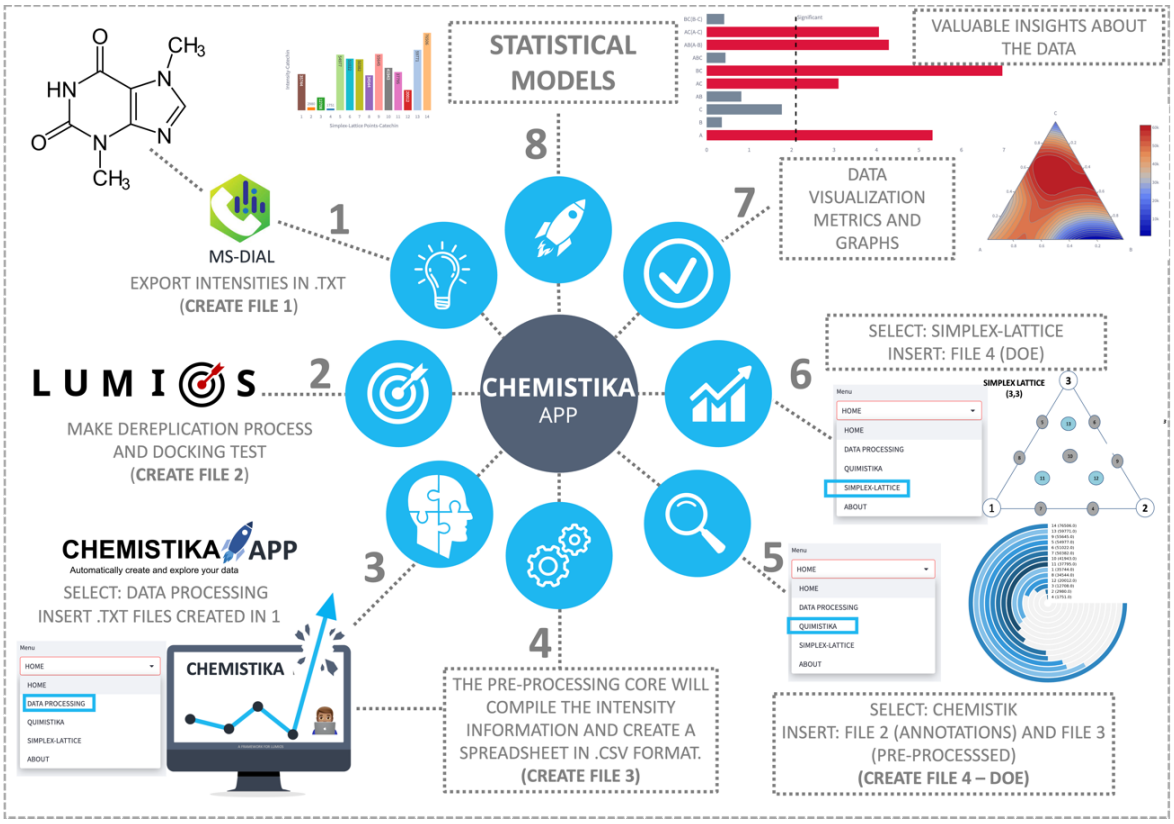
A criação da API Chemistika envolveu o desenvolvimento de um sistema totalmente programado na linguagem Python. A plataforma foi planejada para ser uma solução sem código com uma interface simples e intuitiva que permite aos usuários introduzir dados através da função "arrastar e soltar". Além de trabalhar com software de análise estatística, o API Chemistika pode incorporar funcionalidades LUMIOS, alinhando informações químicas de espectrometria de massas, anotações moleculares, desreplicação, docagem molecular e identificando PNs de interesse farmacológico para essas análises. Dessa forma, o API Chemistika pode avaliar diferentes matrizes naturais complexas ou funcionar como software de análise estatística. A API Chemistika segue o seguinte pipeline:

- **Preparação de dados espectrais (núcleo de pré-processamento):** O usuário faz upload de arquivos contendo informações espectrais, como exemplo, utilizou-se amostras de extratos de polpa de sementes de cacau não fermentada. Os dados são submetidos a uma etapa de pré-processamento, onde ocorre o alinhamento espectral, disponibilizando instantaneamente as informações no formato .csv. Durante esse processo, é criado um conjunto de dados contendo informações relevantes para o usuário. Este conjunto de dados pode ser baixado se o usuário precisar realizar análises diversificadas ou sistematizar tais informações.
- **Núcleo de processamento do API Chemistika:** O segundo núcleo de processamento é ativado quando o usuário insere informações contendo anotações moleculares, que foram desreplicadas pela aplicação LUMIOS, e o conjunto de dados produzido na primeira etapa de processamento da API Chemistika. Neste núcleo de processamento, as anotações moleculares que mostraram afinidade por alvos biomacromoleculares de doenças respiratórias (já integrados e indicados pelo LUMIOS) com suas respectivas intensidades relativas foram acopladas. A partir deste ponto, são geradas visualizações sobre a intensidade relativa média de todos os extratos brutos (28 execuções devido ao design Simplex-Lattice 3x3 com pontos adicionais, considerando as duplicatas).
- **Download de informações químicas e planejamento de experimentos:** É possível baixar informações químicas que serão incorporadas à funcionalidade final da API Chemistika, que é projetar e analisar experimentos. Nesta etapa, a análise estatística é automatizada e os modelos estarão disponíveis para consulta juntamente com uma análise estatística completa, como cálculos de coeficientes e tabelas de análise de variância (ANOVA), bem como gráficos de Pareto e curvas de contorno. O aplicativo tem diferentes etapas, desde a preparação de dados espectrais até a geração de visualização e projetos de experimentos. Esta abordagem automatizada visa facilitar a análise de informações químicas e fornecer aos

usuários uma plataforma intuitiva para explorar e obter insights valiosos a partir de dados espectrais voltados para anotações moleculares.

A Figura 19 resume o ecossistema computacional do API Chemistika, mostrando suas integrações com outras aplicações e indicando o fluxo de processamento de dados a partir de matrizes naturais complexas.

Figura 19 - Ecossistema de funcionamento do APP CHEMISTIKA.

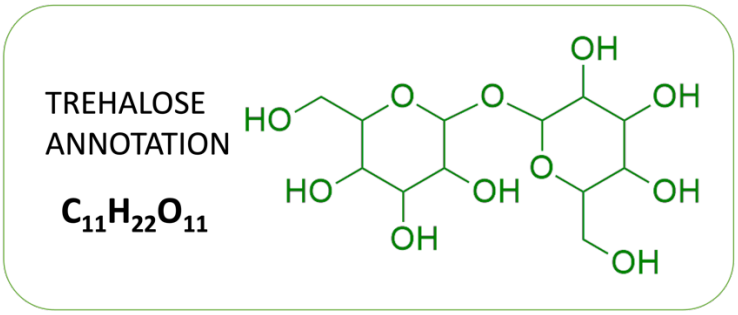


Fonte: Elaborado pelo autor (2023).

3 RESULTADOS E DISCUSSÃO

Os métodos estatísticos oferecidos pela plataforma Chemistika foram testados e avaliados com dados espectrais obtidos de moléculas presentes na polpa de sementes de cacau não fermentados. Com o acoplamento ao LUMIOS, treze anotações moleculares foram identificadas, e, dentre elas, dez mostraram resultados de afinidade promissores com quatro proteínas relacionadas às doenças respiratórias, como SARS-CoV-2 e asma. Com base neste filtro, decidiu-se demonstrar a funcionalidade do API Chemistika na análise estatística do sinal molecular da trealose ($C_{11}H_{22}O_{11}$) (Figura 20) como variável de resposta de um SLD 3x3.

Figura 20 - Fórmula molecular e estrutural da anotação trealose, utilizada como exemplificação e testagem do APP CHEMISTIKA.



Fonte: Elaborado pelo autor (2023).

O objetivo do SLD 3x3 foi o de encontrar um modelo que otimizasse o processo de extração deste metabólito. A Tabela 7 representa a planilha gerada automaticamente pelo aplicativo Chemistika. Os valores das variáveis independentes (componentes das misturas de solventes) estão codificados.

Tabela 7 - Planejamento SLD (3x3) utilizando a intensidade relativa da trealose como resposta.

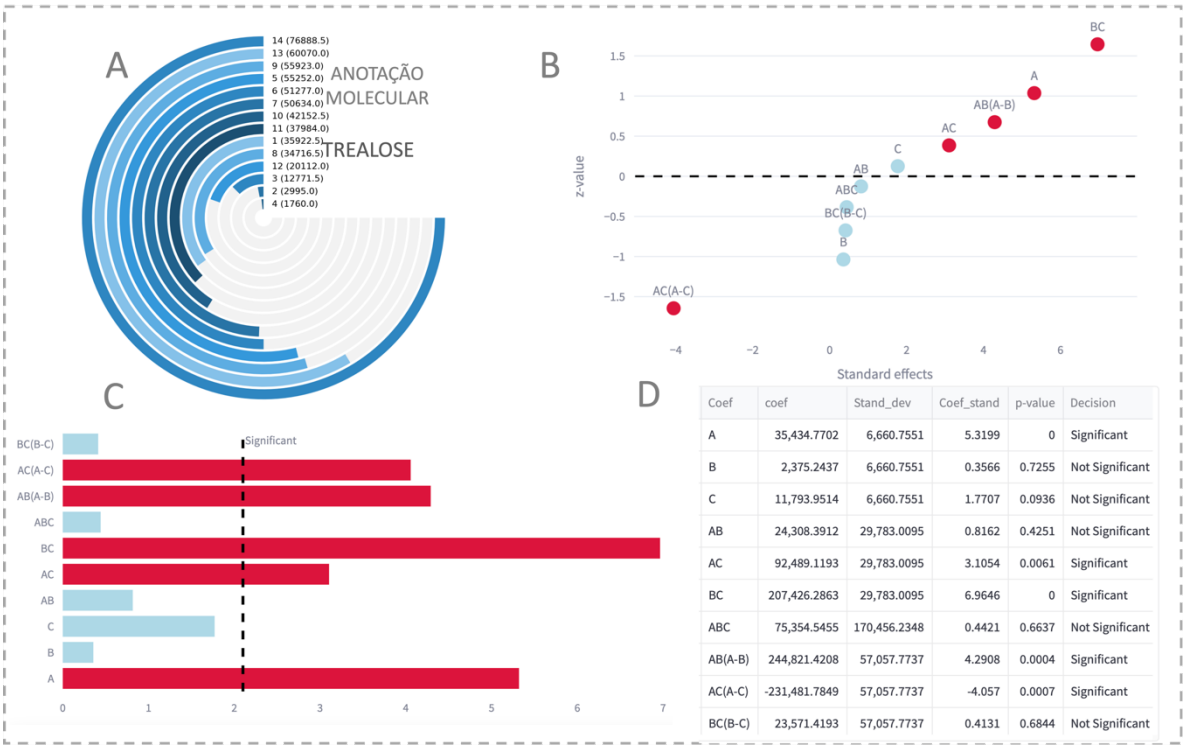
EXP.[1]	CONDIÇÕES EXPERIMENTAIS			MÉDIA DA INTENSIDADE DA TREALOSE
	COMPONENTES			
	A: Hexano	B: Ac. Etila	C: Etanol	
1	1,00	0,00	0,00	35.922,5
2	0,00	1,00	0,00	2.995,0
3	0,00	0,00	1,00	12.771,5
4	0,33	0,67	0,00	1.760,0
5	0,33	0,00	0,67	55.252,0
6	0,00	0,33	0,67	51.277,0
7	0,67	0,33	0,00	50.634,0
8	0,67	0,00	0,33	34.716,5
9	0,00	0,67	0,33	55.923,0
10	0,33	0,33	0,33	42.152,05
11	0,67	0,17	0,17	37.984,0
12	0,17	0,67	0,17	20.112,0
13	0,17	0,17	0,67	60.070,0
14	0,33	0,33	0,33	76.888,5

A trealose, um dissacarídeo composto por duas moléculas de glicose, recebeu atenção significativa devido às suas diversas atividades biológicas. Iturriaga e colaboradores (ITURRIAGA; SUÁREZ; NOVA-FRANCO, 2009) destacaram o papel multifacetado da trealose, enfatizando sua importância como molécula osmoprotetora e de sinalização, atuando como um agente protetor contra vários estresses ambientais e desempenho na sobrevivência celular. A segurança e a tolerância da trealose em humanos foram amplamente discutidas no estudo de Ohtake & Wang (OHTAKE; WANG, 2011). Em uma revisão abrangente, Richards e equipe (RICHARDS et al., 2002) confirmam o perfil favorável de segurança da trealose e seu potencial para aplicações generalizadas.

Além disso, a molécula de trealose foi identificada pelo LUMIOS como candidata a atuar em alvos biomacromoleculares de doenças respiratórias em simulações computacionais, como docagem e dinâmica molecular, corroborando com estudos de Martinon et al. (MARTINON et al., 2020a) que demonstraram o potencial da trealose contra SARS-CoV-2.

A intensidade do sinal da anotação molecular da trealose nos experimentos variou de 1.760,0 a 76.888,5 (Figura 21-A). A maior intensidade dos sinais foi detectada no experimento 14 (76.888,5) quando a composição da mistura extratora era ternária: 0,33% de hexano, 0,33% de acetato de etila e 0,33% de etanol. A menor detecção de sinais da anotação molecular da trealose ocorreu no experimento 4 (1.760,0), realizado com uma mistura extratora binária composta por 0,33% de hexano e 0,67% de acetato de etila.

Figura 21 - Resultados gerados pelo API Chemistika em análise à anotação molecular trealose.



Coef	coef	Stand_dev	Coef_stand	p-value	Decision
A	35,434.7702	6,660.7551	5.3199	0	Significant
B	2,375.2437	6,660.7551	0.3566	0.7255	Not Significant
C	11,793.9514	6,660.7551	1.7707	0.0936	Not Significant
AB	24,308.3912	29,783.0095	0.8162	0.4251	Not Significant
AC	92,489.1193	29,783.0095	3.1054	0.0061	Significant
BC	207,426.2863	29,783.0095	6.9646	0	Significant
ABC	75,354.5455	170,456.2348	0.4421	0.6637	Not Significant
AB(A-B)	244,821.4208	57,057.7737	4.2908	0.0004	Significant
AC(A-C)	-231,481.7849	57,057.7737	-4.057	0.0007	Significant
BC(B-C)	23,571.4193	57,057.7737	0.4131	0.6844	Not Significant

Fonte: Elaborado pelo autor (2023).

Os efeitos que impactaram a intensidade dos sinais da anotação molecular da trealose foram estatisticamente avaliados usando a distribuição normal de efeitos padronizados e o diagrama de Pareto.

Todos os fatores e interações não significativos na intensidade do sinal da anotação molecular da trealose estão próximos da linha zero no eixo das abscissas com uma probabilidade de 50% (considerando a distribuição normal dos dados).

Em contraste, aqueles considerados significativos estão afastados desse limiar (Figura 21-B).

O efeito principal A (hexano) e os termos de interação AC (hexano.etanol), BC (acetato de etila.etanol), AB(A-B) (hexano.acetato de etila.[hexano-acetato de etila]) e AC(A-C) (hexano.etanol.[hexano-etanol]) foram significativos na intensidade do sinal da anotação molecular da trealose. Os sinais positivos do efeito principal A e dos termos de interação AC, BC e AB(A-B) aumentam a intensidade dos sinais na anotação molecular da trealose (sinergismo). Em contraste, a interação do termo AC(A-C) indica uma redução na intensidade do sinal (antagonismo).

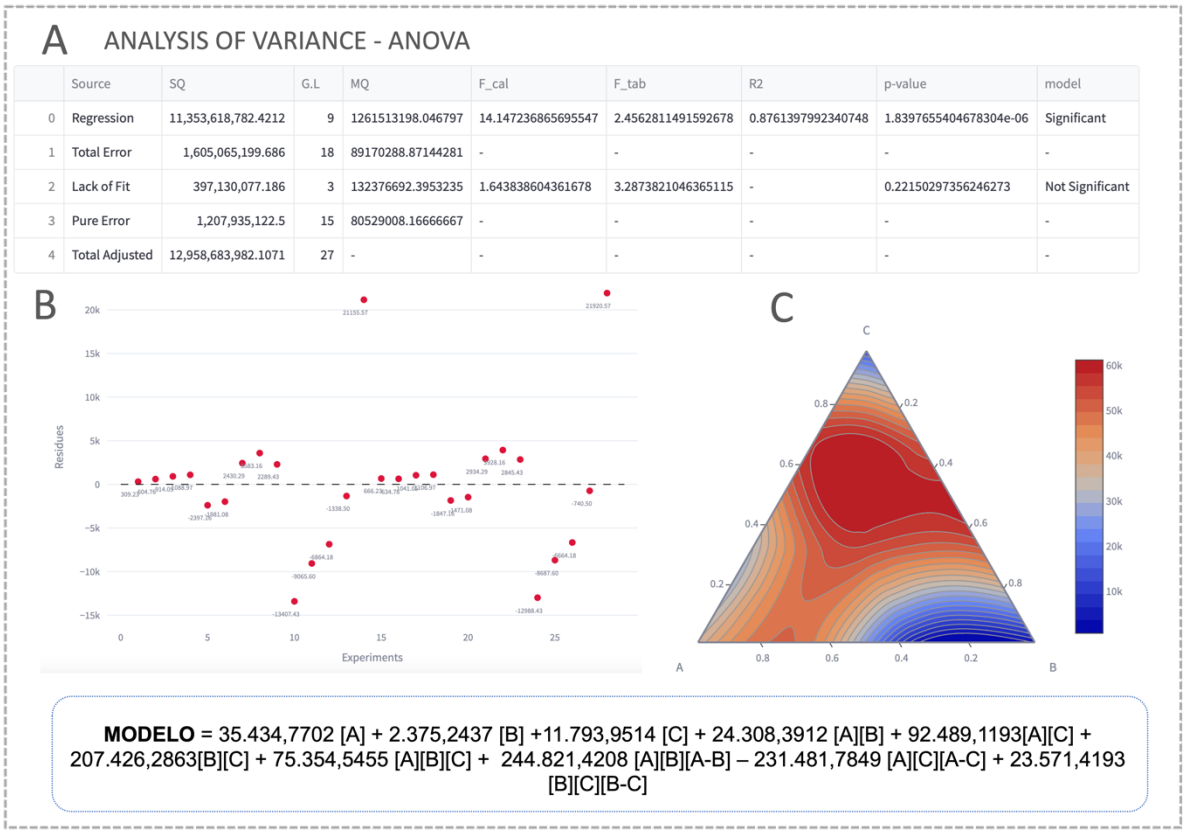
Em valores absolutos, os efeitos padronizados também foram plotados nos gráficos de Pareto (Figura 21-C). Efeitos maiores que 2,101 (p-valor = 0,05), que cruzam a linha divisória, são considerados significativos. Os valores absolutos padronizados dos efeitos principais e interações correspondem ao Teste T de Student calculado para cada componente da mistura.

O modelo matemático mais adequado para representação dos dados deste planejamento de misturas foi o cúbico completo (Figura 21-D). Assim, considerando todos os fatores envolvidos na intensidade dos sinais da anotação molecular da trealose, o modelo foi construído (Figura 22 – Modelo).

Subsequentemente, o modelo foi validado pela análise de variância (ANOVA) (Figura 23-A) e utilizando o teste F, comparando o valor F do modelo cúbico completo com o valor F tabelado. Nas condições investigadas, o teste F de regressão (14,14) foi superior ao valor tabelado (2,45). Note-se que o valor *p* da tabela ANOVA para o modelo de regressão é $1{,}83 \times 10^{-6}$. Quando comparado a um nível de significância de 0,05, verifica-se que, com um valor *p* menor que o nível de significância, a hipótese nula de que não há relação significativa entre as variáveis independentes e dependentes pode ser rejeitada. Isso indica que os coeficientes estimados no modelo são estatisticamente diferentes de zero. Quanto à falta de ajuste, o valor *p* de 0,22 indica evidência insuficiente para rejeitar a hipótese nula de que não há falta de ajuste no modelo. Isto sugere que o modelo de regressão capta adequadamente a relação entre as variáveis independentes e dependentes,

e a falta de ajuste não é estatisticamente significativa. Além disso, o teste F de falta de ajuste (1,64) foi menor que o valor F tabelado (3,28).

Figura 23 **-** Resultados do modelo estatístico gerado pelo APP CHEMISTIKA em análise à intensidade relativa de sinal da trealose. Em (A): tabela da análise de variância (ANOVA) do modelo. (B): Gráfico de dispersão para os resíduos do modelo e (C): mapa de contorno



	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	11,353,618,782.4212	9	1261513198.046797	14.147236865695547	2.4562811491592678	0.8761397992340748	1.8397655404678304e-06	Significant
1	Total Error	1,605,065,199.686	18	89170288.87144281	-	-	-	-	-
2	Lack of Fit	397,130,077.186	3	132376692.3953235	1.643838604361678	3.2873821046365115	-	0.22150297356246273	Not Significant
3	Pure Error	1,207,935,122.5	15	80529008.16666667	-	-	-	-	-
4	Total Adjusted	12,958,683,982.1071	27	-	-	-	-	-	-

Fonte: Elaborado pelo autor (2023).

No entanto, o modelo apresentou um coeficiente de determinação (R^2) de 0,87, indicando que as variáveis independentes incluídas no modelo podem explicar aproximadamente 87% da variação da variável dependente. Isso sugere que o modelo capta uma variabilidade substancial dos dados e pode estimar razoavelmente a relação entre as variáveis.

Embora apenas cinco coeficientes tenham sido considerados significativos, é importante destacar que a significância estatística nem sempre reflete a relevância prática, pois outros coeficientes não significativos podem desempenhar um papel

menor, mas ainda assim contribuem para a precisão global do modelo. Uma avaliação cuidadosa da interpretação e validade do modelo é essencial, considerando o contexto específico do problema e realizando análises adicionais, se necessário.

As suposições da ANOVA também foram verificadas através do comportamento dos resíduos; desta forma, um gráfico de resíduos contra a sequência de execução dos experimentos foi gerado. Observou-se que não havia correlação entre as variáveis independentes ou a sequência temporal dos experimentos, visto que os resíduos estão distribuídos ao redor de uma banda horizontal. Assim, é provável que os resíduos do modelo tenham as mesmas propriedades que o erro experimental (Figura 23-B).

Quanto à projeção gráfica do contorno (Figura 23-C), as maiores intensidades de sinal da anotação molecular da trealose estão localizadas na região central do gráfico (cores quentes), onde os experimentos foram conduzidos com misturas extratoras compostas pelos três solventes, em diferentes proporções (experimentos 13 e 14). Em contraste, as menores intensidades desta molécula estão localizadas no canto inferior direito do gráfico, que corresponde ao experimento 2, conduzido exclusivamente com acetato de etila (100%).

Estes resultados contribuem para o desenvolvimento de estratégias de extração mais eficientes e direcionadas para compostos específicos de interesse. No entanto, são necessários estudos adicionais para uma compreensão mais profunda dos mecanismos envolvidos na extração destas moléculas e na otimização dos protocolos de extração.

4 CONSIDERAÇÕES

Com base nas análises estatísticas apresentadas e na peculiaridade do problema abordado (análise de matrizes naturais complexas), o aplicativo Chemistika oferece os seguintes pontos positivos:

- **Análise rápida e robusta:** O aplicativo Chemistika permite uma análise eficiente dos dados, fornecendo resultados rápidos e confiáveis. Este aplicativo é útil quando se lida com matrizes de dados complexas, garantindo uma abordagem eficaz para análises espectrais.
- **Design experimental:** O Chemistika suporta design experimental do tipo misturas, permitindo que os usuários estruturem suas análises de forma planejada. Essa funcionalidade é valiosa para garantir que os experimentos sejam conduzidos eficientemente, economizando tempo e recursos.
- **Comparação com outros aplicativos:** O Chemistika foi comparado a aplicativos comerciais consolidados como o Statistica e o Minitab, e obteve resultados idênticos em termos de coeficientes, ANOVA e cálculos estatísticos. Isso indica que o Chemistika é uma opção específica e confiável para análise estatística, oferecendo resultados consistentes e comparáveis.
- **Plataforma gratuita:** O Chemistika é uma plataforma gratuita, tornando-a acessível a um amplo público de usuários. Esta disponibilidade gratuita é uma grande vantagem, permitindo que pesquisadores, estudantes e profissionais acessem uma ferramenta poderosa sem investimentos financeiros adicionais.
- **Interface intuitiva e sem necessidade de programação:** O Chemistika possui uma interface amigável que não exige conhecimento avançado de programação. Isso facilita o carregamento de dados e a visualização de informações, tornando o processo mais simples e claro para os usuários.
- **Insights mais profundos e conhecimento:** O Chemistika combina design experimental com análises estatísticas apropriadas, fornecendo insights valiosos e informações mais profundas sobre matrizes complexas de produtos naturais. Auxilia na identificação de padrões, relações e

características químicas importantes, ajudando a entender os melhores métodos de extração para um metabólito específico.

Além disso, o aplicativo Chemistika oferece várias vantagens e pontos positivos que o tornam uma ferramenta útil e flexível para usuários em diferentes dispositivos e sistemas operacionais, incluindo:

- **Disponibilidade em múltiplas plataformas:** O Chemistika é uma Interface de Programação Aplicada (API), o que significa que pode ser acessado e utilizado em diversos dispositivos, incluindo smartphones Android e iOS, bem como sistemas operacionais como Linux, Windows e Macintosh. Assim, os usuários podem acessar e utilizar o aplicativo independentemente do seu dispositivo ou sistema operacional.
- **Flexibilidade de acesso:** O Chemistika é uma API baseada na web, por isso os usuários podem acessar suas funcionalidades e recursos através de um navegador. Isso elimina a necessidade de instalação adicional do aplicativo nos dispositivos, simplificando o acesso e tornando-o mais conveniente.
- **Portabilidade:** Usar o Chemistika em diferentes dispositivos e sistemas operacionais oferece maior portabilidade ao aplicativo. Os usuários podem acessar e utilizar a API em qualquer lugar, a qualquer momento, com uma conexão à internet e um dispositivo compatível.
- **Integração com outros aplicativos:** O Chemistika pode ser facilmente integrado a outros aplicativos e sistemas, permitindo uma troca eficiente de dados e informações. Tal funcionalidade possibilita incorporar as funcionalidades do Chemistika em fluxos de trabalho existentes e utilizar dados gerados por outras ferramentas ou aplicativos.
- **Atualizações e melhorias contínuas:** Como o Chemistika é uma API baseada na web, os desenvolvedores podem atualizá-la e aprimorá-la regularmente. Assim, isso significa que os usuários podem se beneficiar de novos recursos, correções de bugs e aprimoramentos contínuos do aplicativo ao longo do tempo.

Desta forma, o aplicativo Chemistika se destaca por sua flexibilidade, portabilidade e capacidade de integração, permitindo que os usuários acessem suas funcionalidades em uma ampla gama de dispositivos e sistemas operacionais. Essas características contribuem para uma experiência de usuário conveniente e adaptável, oferecendo análise estatística rápida, robusta e confiável para matrizes de dados complexas, focando nas funcionalidades de design de mistura Simplex-Lattice. Sua disponibilidade gratuita, interface intuitiva e comparação favorável com outros aplicativos comerciais o tornam uma ferramenta específica para análise química e espectral.

4.1 Novas versões do aplicativo Chemistika:

- **Maior diversidade de tipos de planejamento:** Além de designs fatoriais simples, considera-se incorporar designs fatoriais complexos. Assim, permitirá uma abordagem mais abrangente e sofisticada não apenas para planejamentos de misturas, podem explorar e acessar informações mais complexas.
- **Integração com bancos de dados:** Integrar o aplicativo com bancos de dados químicos e biológicos, como PubChem e Lotus, para acesso rápido e automático a informações relevantes sobre moléculas identificadas. Essa interação permite uma contextualização completa e uma análise aprofundada das características e propriedades das substâncias.
- **Recursos de aprendizado de máquina:** No futuro, será implementado recursos associados às ciências de dados para melhorar suas capacidades de previsão e modelagem. Assim, pode ajudar a identificar tendências, correlações e padrões ocultos nos dados espectrais e fazer previsões mais precisas sobre a atividade biológica das moléculas.
- **Interação com outras técnicas analíticas:** Integrar o Chemistika com outras técnicas analíticas, como cromatografia gasosa atrelada à espectrometria de massas (GC-MS), para fornecer uma análise mais abrangente e complementar dos extratos naturais. Assim, permitirá uma

análise mais completa das moléculas presentes e uma melhor compreensão das características químicas dos extratos.

- **Interface de usuário melhorada:** Melhorar a interface do usuário do Chemistika para torná-la mais intuitiva, fácil de usar e personalizável, incluindo recursos como arrastar e soltar, visualizações interativas, gráficos personalizados e a capacidade de personalizar relatórios e resultados de acordo com as necessidades do usuário.
- **Integração com outras ferramentas de análise estatística:** Integrar o Chemistika com outras ferramentas de análise estatística, como as bibliotecas R ou Python, para expandir as opções de análise e fornecer capacidades avançadas de modelagem estatística e visualização de dados.
- **Ferramentas de anotação e identificação de compostos:** Integrar o Chemistika diretamente com o LUMIOS, que contém ferramentas de anotação e identificação de compostos integrados ao aplicativo. Tal ferramenta facilitará a correta atribuição de anotações moleculares em espectros de massa.
- **Colaboração e compartilhamento de resultados:** Incorporar recursos de colaboração e compartilhamento de resultados, permitindo que os usuários compartilhem seus dados, análises e modelos com outros pesquisadores, possibilitando a troca de conhecimentos e colaboração entre diferentes grupos de pesquisa.

Essas melhorias sugeridas têm como objetivo expandir as funcionalidades do aplicativo Chemistika, tornando-o mais poderoso, versátil e adaptado às necessidades dos usuários.

Figura 24 - Selo/logotipo do app Chemistika.

Fonte: Elaborado pelo autor (2023).

Aplicativo hospedado em: **https://chemistika.streamlit.app**

VEJA O VÍDEO DEMONSTRATIVO EM: **https://www.youtube.com/watch?v=xPZ7Btv3DrQ**

CAPÍTULO 4 – EXPLORANDO MATRIZES COMPLEXAS DO CACAU: ANÁLISE DE BIOATIVOS UTILIZANDO PLANEJAMENTO DE MISTURAS ATRAVÉS DA PLATAFORMA CHEMISTIKA E MODELOS DE AFINIDADE PARA ALVOS BIOMACROMOLECULARES ASSOCIADOS A DOENÇAS RESPIRATÓRIAS

RESUMO

Este capítulo aborda uma análise exploratória das matrizes naturais complexas provenientes do cacau submetido a um processo fermentativo de 168 horas. Utilizando o aplicativo LUMIOS, foram identificadas 13 estruturas, chamadas de anotações, nos extratos brutos tanto da polpa do fruto, como durante o processo fermentativo. Dentre tais anotações, 10 apresentaram potencial capacidade de modular alvos biomacromoleculares de doenças respiratórias, como SARS-CoV-2 e asma. Testes de docagem molecular foram realizados para analisar a interação dessas moléculas com os bioreceptores associados a essas doenças. O aplicativo Chemistika foi utilizado para automatizar o planejamento de mistura do tipo Simplex-Lattice, permitindo a construção de modelos estatísticos que visaram otimizar as intensidades relativas de cada anotação, identificando as melhores etapas para a extração dessas moléculas. Os resultados obtidos forneceram, em sua maioria, modelos com altos coeficientes de determinação, sugerindo um direcionamento para futuras pesquisas no desenvolvimento de produtos farmacológicos para doenças respiratórias, explorando o potencial terapêutico do cacau e suas moléculas durante o processo fermentativo. Esse estudo tem como impacto o avanço da terapêutica respiratória e abre novas perspectivas para o uso de compostos derivados do cacau como agentes terapêuticos potenciais.

Palavras-chave: Simplex-Lattice; automatização de análises, software, tratamento de dados, modelos estatísticos, anotações moleculares.

1 INTRODUÇÃO

O cacau (*Theobroma cacao*) tem sido amplamente estudado devido à sua importância na produção de alimentos e ao seu potencial para o desenvolvimento de produtos farmacológicos. O cacau contém uma variedade de compostos bioativos, incluindo polifenóis, alcaloides e ácidos graxos, que apresentam propriedades farmacológicas promissoras (ELLAM; WILLIAMSON, 2013). Esses compostos têm demonstrado efeitos antioxidantes, anti-inflamatórios e neuroprotetores (KATZ; DOUGHTY; ALI, 2011). Além disso, os flavonoides encontrados no cacau, como catequina e a procianidina, têm sido associados à redução do risco de doenças cardiovasculares, como mencionado em pesquisas (BUITRAGO-LOPEZ et al., 2011; RIED et al., 2012), e estão presentes em estudos mais recentes em tratamento de doenças respiratórias, como SARS-CoV-2 (CHOURASIA et al., 2021b; HENSS et al., 2021b; MISHRA et al., 2021b). Essas descobertas têm despertado o interesse da comunidade científica e da indústria farmacêutica para explorar o potencial do cacau não só através do chocolate, mas como fonte promissora de moléculas que possam atuar na prevenção e tratamento de doenças (BERRY et al., 2010; CHEN et al., 2618; RIED et al., 2012).

A fermentação tem sido utilizada como um processo-chave na obtenção de produtos do cacau (*Theobroma cacao* L.), que influencia diretamente suas propriedades sensoriais e químicas (DE BRITO et al., 2001; DE VUYST; WECKX, 2016b) devido à ação de um consórcio microbiano bastante complexo (SARBU; CSUTAK, 2019). Durante esse processo, ocorrem diversas transformações nas matrizes naturais complexas do cacau, resultando em uma variedade de compostos bioativos (MOREIRA et al., 2013). A compreensão dessas transformações durante a fermentação e a identificação de moléculas promissoras com potencial terapêutico são de grande interesse científico, pois é a fase crucial para exploração da variabilidade metabólica do cacau (DE VUYST; WECKX, 2016b; JOHN et al., 2020).

O processo de fermentação do cacau geralmente dura de 5 a 7 dias (RAHARDJO et al., 2022). Durante esse período, as sementes de cacau são colocadas em caixotes ou montes, onde ocorre a fermentação natural. Durante o

processo de fermentação, ocorrem reações bioquímicas que resultam na transformação dos compostos presentes nas sementes de cacau, incluindo a redução da acidez e a formação de compostos voláteis responsáveis pelos aromas desejados (KONGOR et al., 2016b). O tempo de fermentação pode variar dependendo das condições locais, das práticas de fermentação utilizadas e do tipo de cacau (AFOAKWA et al., 2013). É importante ressaltar que o tempo de fermentação adequado pode influenciar diretamente na produção metabólica dos produtos (FEBRIANTO; ZHU, 2022).

Neste estudo, realizou-se uma análise exploratória das matrizes naturais complexas provenientes do cacau submetido a um processo fermentativo de 168 horas. Utilizando o aplicativo LUMIOS (VIEIRA; ALVES DE SOUSA; CASTRO-GAMBOA, 2023), uma ferramenta avançada de desreplicação molecular, identificou-se 13 diferentes estruturas, chamadas de anotações (BACH; SCHYMANSKI; ROUSU, 2022), no extrato bruto da polpa do fruto, antes da fermentação e em duas etapas do processo fermentativo, sendo 84 horas e 168 horas. Essas anotações representam moléculas com características distintas e potencial atividade biológica em alvos biomacromoleculares de doenças respiratórias, como SARS-CoV-2 (ROSSETTI et al., 2022b)e asma (COSTANZO et al., 2003c; HALL et al., 2012c; MAUN et al., 2020b), buscando investigar, assim, a afinidade das anotações identificadas com bioreceptores associados a essas doenças. Utilizou-se testes de docagem molecular para analisar a interação das moléculas com tais bioreceptores.

Entre as anotações, o anidrido ftálico, ácido ftálico, teobromina, catequina, trealose e procianidina foram sinalizadas antes da fermentação (0 hora); adenina, anidrido ftálico, teofilina, teobromina, tirosina, catequina e procianidina (84 horas de processo fermentativo) e, por fim, adenina, indol-3-acetamida, teobromina, procianidina, tirosina, fenilalanina e catequina em 168 horas de fermentação. Tais anotações demonstraram afinidade com proteínas relacionadas a doenças respiratórias, representando ligantes em potenciais para o desenvolvimento de terapias ou intervenções relacionadas a essas condições. A compreensão de suas propriedades e mecanismos de interação com os bioreceptores pode fornecer

insights importantes para a concepção de novos fármacos ou abordagens terapêuticas.

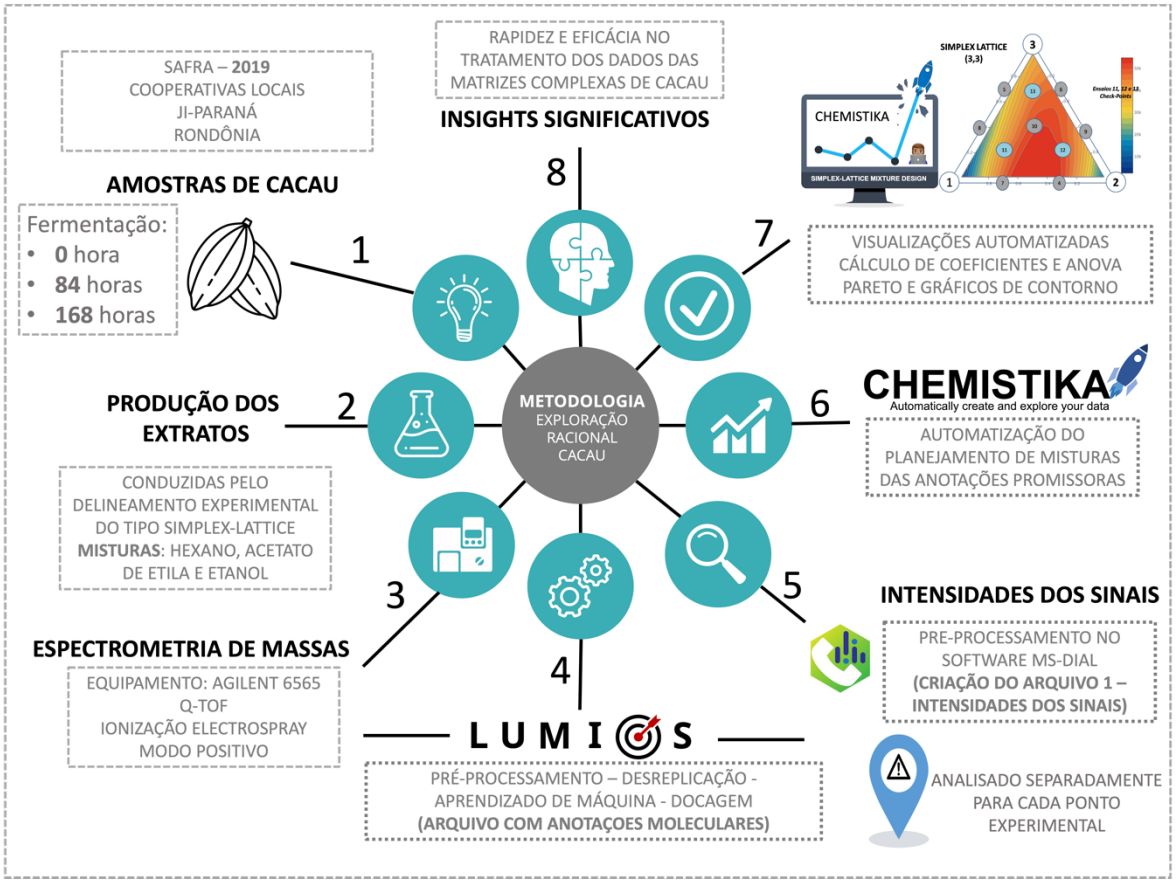
Neste estudo, o aplicativo Chemistika (https://chemistika.streamlit.app) foi utilizado para analisar as sinalizações moleculares com afinidade por bioreceptores relacionados a doenças respiratórias. Através da automatização do planejamento de mistura do tipo Simplex-Lattice, foram construídos modelos estatísticos que visaram identificar e otimizar as respostas relacionadas às intensidades relativas de cada anotação, identificando as melhores etapas para a obtenção das moléculas de interesse. Os resultados forneceram direcionamento para futuras pesquisas no desenvolvimento de produtos farmacológicos para doenças respiratórias, explorando as moléculas encontradas durante o processo fermentativo do cacau, dando um indicativo da etapa mais promissora que ela pode ser identificada/extraída para estudos futuros.

Assim, o objetivo deste estudo foi analisar as matrizes complexas do cacau, desenvolvendo modelos estatísticos (usando a plataforma Chemistika) que previram a intensidade relativa de anotações moleculares com afinidade para bioreceptores associados a doenças respiratórias, obtidas em diferentes etapas da fermentação. A compreensão de tais interações sinalizam impactos potenciais em avançar no tratamento de doenças respiratórias e abrir novas possibilidades para o uso de compostos derivados do cacau como potenciais agentes terapêuticos.

2 METODOLOGIA

O pipeline utilizado para exploração das misturas complexas de cacau, visando a construção de modelos estatísticos para previsão das intensidades relativas de cada anotação está sistematizados na Figura 25.

Figura 25 - Resumo da metodologia utilizada para exploração das matrizes complexas do cacau a partir do planejamento de misturas Simplex-Lattice.



Fonte: Elaborado pelo autor (2023).

2.1 Aplicativo Chemistika para automatização das análises Simplex-Lattice.

A plataforma Chemistika, detalhadamente descrita no Capítulo 3, foi empregada em associação com os dados disponibilizados pelo LUMIOS para determinar as intensidades relativas das anotações que demonstraram afinidade

por proteínas associadas a doenças respiratórias. Essas intensidades foram utilizadas como variáveis dependentes no desenho experimental. Ademais, todas as análises estatísticas, incluindo a determinação dos coeficientes, a execução da ANOVA, e a geração automática de gráficos de Pareto e mapas de contorno, foram exclusivamente organizadas e sistematizadas pela plataforma Chemistika.

3 RESULTADOS E DISCUSSÃO

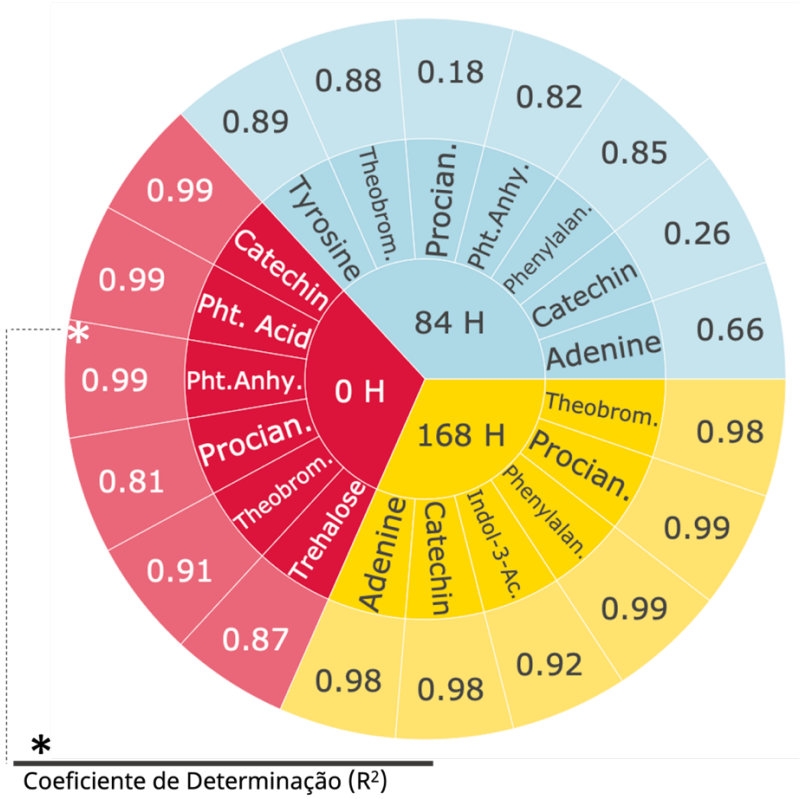
3.1 Ponto inicial (sementes de cacau sem fermentação) – 0 hora:

Para a fase inicial do processo fermentativo foram desreplicadas seis anotações moleculares, sendo: catequina, ácido ftálico, anidrido ftálico, procianidina, teobromina e trealose, em consonância com os dados descritos no Capítulo 2. Quando utilizado em conjunto com o LUMIOS, o Chemistika foi capaz de reconhecer as anotações que apresentaram desempenho notável em simulações de docagem molecular, inserindo as intensidades relativas de cada uma dessas estruturas.

Dessa forma, modelos estatísticos foram desenvolvidos utilizando a intensidade relativa de cada anotação como variável resposta. Empregou-se um planejamento de misturas do tipo Simplex-Lattice (3x3), composto por 14 pontos experimentais (realizados em duplicata), para construir modelos capazes de prever, com alta precisão, a intensidade do sinal de cada anotação. De modo geral, os modelos estatísticos alcançaram um coeficiente de determinação (R^2) elevado, indicando uma robusta capacidade preditiva.

A Figura 25 ilustra a distribuição dessas anotações nas diversas etapas de fermentação e inclui os resultados dos coeficientes de determinação de cada modelo.

Figura 26 - Apontamentos das anotações moleculares desreplicadas pelo aplicativo LUMIOS em cada ponto experimental (0 hora, 84 horas e 168 horas).



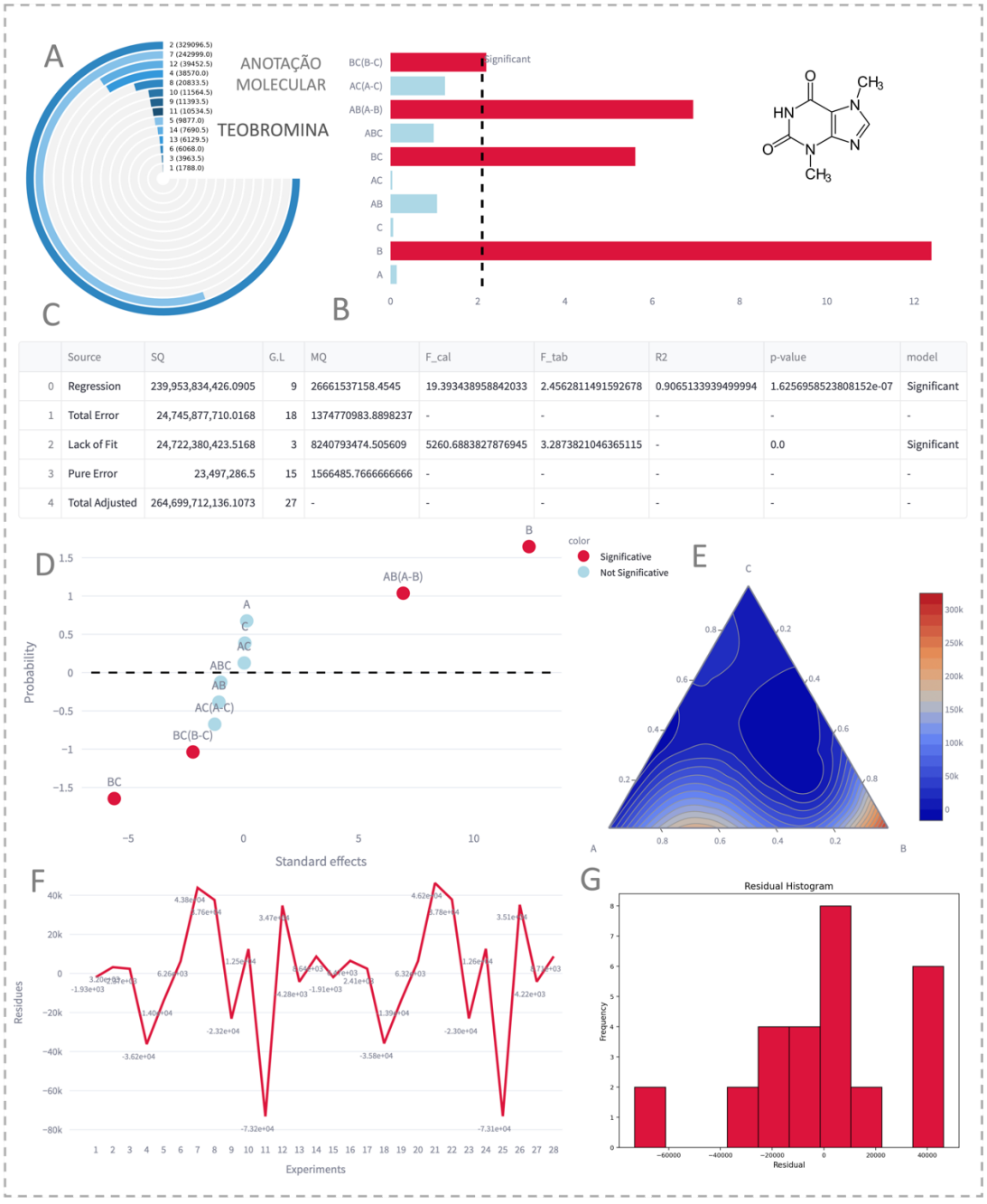
Fonte: Elaborado pelo autor (2023).

Para o ponto inicial, de 0 hora, destacou-se, em particular, a molécula de trealose, um dissacarídeo que apresentou afinidade com 3 das 4 proteínas disponíveis para docagem no software LUMIOS, tendo alta afinidade por alvos de SARS-CoV-2 e asma. Essa molécula tem sido objeto de estudos como uma promissora candidata para o desenvolvimento de terapias direcionadas a doenças respiratórias (MARTINON et al., 2020b; OHTAKE; WANG, 2011) e foi utilizada como exemplo na demonstração da plataforma Chemistika, no capítulo 3, portanto, será omitida nesta etapa.

Destaca-se também nos extratos não fermentados, a molécula de teobromina, que é o principal marcador molecular do cacau (CÁDIZ-GURREA et al., 2014). Ao realizar a modelagem estatística para descrever as intensidades relativas desse metabólito, foram obtidos resultados robustos, que podem ser visualizados na Figura 27. O coeficiente de determinação (R^2) encontrado foi de 0,91, indicando

que o modelo proposto consegue explicar 91% da variação nas intensidades relativas da teobromina. Embora esse valor seja considerado alto, sugere que o modelo pode não capturar completamente todas as variações observadas.

Figura 27 - Análises (Teobromina). A – Distribuição da intensidade dos sinais. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G - Histograma residual.



	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	239,953,834,426.0905	9	26661537158.4545	19.393438958842033	2.4562811491592678	0.9065133939499994	1.6256958523808152e-07	Significant
1	Total Error	24,745,877,710.0168	18	1374770983.8898237	-	-	-	-	-
2	Lack of Fit	24,722,380,423.5168	3	8240793474.505609	5260.6883827876945	3.2873821046365115	-	0.0	Significant
3	Pure Error	23,497,286.5	15	1566485.7666666666	-	-	-	-	-
4	Total Adjusted	264,699,712,136.1073	27	-	-	-	-	-	-

Fonte: Elaborado pelo autor (2023).

Além disso, dos 10 coeficientes disponíveis para a construção do modelo, apenas 4 deles foram considerados estatisticamente significativos. Isso significa que apenas uma parte das variáveis independentes incluídas no modelo demonstraram uma relação estatisticamente relevante com as intensidades da teobromina. Os outros coeficientes podem não estar contribuindo significativamente para a explicação das variações observadas. Dentre os coeficientes importantes para o modelo, dois deles apresentam valores negativos: BC (-5,61) e BC(B-C) (-2,19), que quando analisados perante o gráfico de distribuição normal, seus valores z indicam que estão 1,64 e 1,04 desvios-padrão abaixo da média dos demais efeitos, destacando que valores maiores deste coeficiente causam uma redução na intensidade dos sinais da teobromina. Adicionalmente, como existe um coeficiente modulado pela diferença entre B e C, quanto maior a diferença entre eles, maior a redução das intensidades deste metabólito. Por outro lado, o coeficiente B também se afasta 1,64 desvios-padrão da média, mas de maneira positiva, (coeficiente padronizado de valor 12,40) sinalizando que quanto maior o valor de B (relacionado ao solvente de acetato de etila), mais intenso seria o sinal atribuído a esta xantina.

Em análise aos resíduos gerados pelo modelo proposto para a molécula de teobromina, nota-se que são bem distribuídos, tanto na parte positiva quanto negativa, e que não apresentam correlações ou tendências, porém, é evidenciado pelo histograma que os resíduos não se distribuem igualmente, tendo uma tendência não normal, com desvios significativos da simetria, formando uma cauda mais acentuada do lado direito, isso pode indicar a presença de outliers ou uma inadequação do modelo em capturar a estrutura dos dados, o que vai de encontrado à falta de ajuste apontada na ANOVA.

As demais anotações moleculares assinaladas no ponto 0 do experimento são atribuídas a dois flavonoides: catequina e procianidina, e dois ftalatos (ácido e anidrido, que apresentam afinidade por alvos de doenças respiratórias). Com exceção da anotação associada à molécula de procianidina, que apresentou um R^2 de 0,81, as anotações moleculares estudadas apresentaram coeficiente de determinação (R^2) acima de 0,99 para o modelo cúbico completo.

Ao analisar especificamente a molécula de catequina, que apresentou alta afinidade por alvos biomoleculares de doenças respiratórias, validando dados da literatura sobre seu uso no reestabelecimento da saúde contra SARS-CoV-2 e outras doenças do trato respiratório (CHOURASIA et al., 2021b; FURUSHIMA et al., 2019b; HENSS et al., 2021b), observou-se que apenas cinco coeficientes foram considerados significativos na construção do modelo.

Um ponto de destaque é que os coeficientes relacionados aos solventes hexano (A) e etanol (C) não foram significativos individualmente para a construção do modelo. No entanto, apenas o coeficiente B (relacionado ao solvente de acetato de etila) tem valor positivo (59,55) e está 1,64 desvios-padrão da média do que os demais efeitos podem causar no modelo. Os demais coeficientes causam influência negativa na intensidade da resposta caso eles aumentem.

Esse resultado sugere que a interação entre os solventes desempenha um papel crucial na eficiência da extração dessa molécula, sinalizando que quanto menores os valores das interações entre os solventes, mais expressivo é a intensidade do sinal referente a esta anotação. Em análise aos resíduos proporcionados pelo modelo, apesar de haver boa distribuição e não sinalizar tendências ou padrões ao longo dos experimentos, no histograma há uma leve distorção para esquerda, o que corrobora com as informações apontadas pela ANOVA, a qual sinaliza falta de ajuste deste modelo. A análise dos dados da catequina pode ser visualizada na Figura Suplementar 25.

Para a anotação molecular vinculada à procianidina, o modelo cúbico completo que prevê as intensidades relativas dessa molécula em matrizes complexas de cacau indicou um coeficiente de determinação (R^2) de 0,81. Isso sugere que 19% da variância não é explicada pelo modelo.

A procianidina tem sido objeto de investigações avançadas para o tratamento de doenças respiratórias (RAJASEKARAN; RAJASEKAR; SIVANANTHAM, 2021), incluindo SARS-CoV-2 (GHOSH, 2021), alinhando-se com as indicações *in silico* deste estudo. Em relação ao modelo estatístico para a previsão das intensidades ligadas a este metabólito, com o objetivo de identificar a melhor combinação de misturas para sua extração das matrizes complexas de cacau, o modelo revelou

apenas três coeficientes significativos. Estes estão associados aos coeficientes que indicam relação, nomeadamente AB(A-B), AB e ABC, com valores padronizados de 5,50, 4,87 e 2,94, respectivamente. Os dois primeiros são coeficientes positivos e destacam-se da média dos outros por 1,64 e 1,04 desvios-padrão, respectivamente. Isso sugere que, quanto maior a diferença entre A e B, mais notável é a intensidade do sinal da molécula. Contudo, a combinação de todos os solventes (ABC) leva a uma redução na intensidade do sinal, apresentando, inclusive, 1,64 desvios-padrão inferior à média dos outros coeficientes.

Ao analisar o modelo cúbico completo, os resíduos desempenharam um papel crucial em avaliar a adequação do modelo aos dados. São os desvios entre os valores observados e os valores previstos pelo modelo, além da sua distribuição, que podem fornecer insights valiosos sobre a validade do modelo para as melhores composições de misturas para extração da procianidina. Dessa forma, com base na análise dos resíduos, não foram observadas tendências, pois estão bem distribuídos ao longo dos experimentos, de maneira uniforme.

No entanto, ao examinar o histograma dos resíduos Figura Suplementar 26-G, percebeu-se uma cauda prolongada para a direita, indicando uma distribuição assimétrica dos resíduos. Esse tipo de assimetria, muitas vezes referida como assimetria positiva, sugere a presença de outliers ou de valores extremos no lado direito da distribuição. A presença de uma cauda à direita pode ser problemática, pois indica que o modelo pode estar subestimando as respostas em determinadas situações. Essa falta de ajuste do modelo pode ser preocupante, pois pode comprometer a precisão das previsões, porém, por se tratar de uma mistura complexa, um modelo com 81% de coeficiente de determinação, pode ser um bom indicativo para obtenção deste metabólito. A molécula de procianidina está presente em todas as etapas do processo fermentativo do cacau, e com base neste indicativo, uma análise comparativa será efetuada nos próximos pontos experimentais (84 horas e 168 horas). As análises referentes à procianidina (em 0 hora de fermentação) podem ser consultadas na Figura Suplementar 26.

Sobre a identificação de moléculas de ftalatos nos extratos de cacau, no estudo conduzido por Thiemann, em 2021, é sinalizado que várias estruturas

derivadas deste tipo de molécula apresentam atividade biológica em alvos diversos, como antitumoral, anti-larvicida, anti-inflamatório, antiviral (THIEMANN, 2021b), mas não há uma ênfase específica em seu uso no tratamento de doenças respiratórias. Quanto ao ácido ftálico, observou-se que seis coeficientes tiveram relevância na elaboração do modelo, o qual demonstrou um coeficiente de determinação (R^2) de 0,99. O coeficiente associado ao solvente hexano (A) se destacou (com valor 61,25), que quando relacionado com valor de z é possível analisar que este coeficiente se distancia dos efeitos médios por 1,64 desvios-padrão, aproximadamente, enquanto os demais coeficientes primários (B) e (C) não apresentaram significância individualmente.

No entanto, as interações entre esses coeficientes causam variações na resposta estudada. Os coeficientes A e ABC apresentam efeitos positivos para o modelo, os demais coeficientes fazem o papel contrário, quanto maior a interação entre eles, menores as intensidades relacionadas à molécula deste ftalato. O processo de extração de um determinado metabólito em matriz complexa não é trivial. Esta complexidade pode, em muitos casos, refletir-se nos resultados modelados, especialmente nos resíduos. Ao examinar o histograma residual do modelo (Figura Suplementar 28-G), fica evidente que uma quantidade significativa dos dados está deslocada para a direita. Esta distribuição assimétrica sugeriu que o modelo pode não ter se ajustado perfeitamente aos dados observados, no sentido de buscar um ponto ótimo de obtenção desta molécula, especialmente considerando-se as nuances intrínsecas à matriz complexa e as particularidades do processo de extração, ou ainda, que outros planejamentos (utilizando outras variáveis) poderiam ser mais adequados para este propósito.

No caso do anidrido ftálico, observou-se que apenas os coeficientes associados ao solvente etanol (C) e às interações entre os coeficientes (B) e (C) (BC(B-C)) não foram significativos para a construção do modelo cúbico completo. Isso indica que o solvente etanol e as interações específicas entre os coeficientes (B) e (C) não são determinantes na compreensão das propriedades desse ftalato. Além disso, constatou-se que as interações entre os solventes AB e AC, juntamente com as interações moduladas pelas diferenças entre eles, AB(A-B) e AC(A-C),

causam uma diminuição na intensidade relativa dessa anotação, pois esses coeficientes apresentam valores negativos no modelo. A única interação significativa que causa um aumento na intensidade do sinal do anidrido ftálico é a interação cúbica ABC (que se distancia da média dos demais efeitos por aproximadamente 1 desvio-padrão), e que contribui positivamente para o modelo. Esse deslocamento dos resíduos para a direita (em ambos os ftalatos) pode, de fato, ser uma das razões para a falta de ajuste do modelo, podendo ser um indicativo de que o modelo não está capturando totalmente a variabilidade inerente aos dados, porém, no contexto de uma matriz complexa, onde o processo de extração não é simples, este modelo pode indicar uma composição de misturas adequada para extração mais eficiente destes compostos.

3.2 Ponto intermediário do processo fermentativo de cacau – 84 horas

Posteriormente, nos extratos obtidos após 84 horas de fermentação espontânea do cacau, foram identificadas sete estruturas químicas promissoras para atuarem em alvos biomacromoleculares de doenças respiratórias por meio de técnicas de desreplicação automatizadas e sistematizadas pelo aplicativo LUMIOS. Essas estruturas foram identificadas como tirosina, teobromina, anidrido ftálico, procianidina, catequina, adenina e fenilalanina (Figura 26). As moléculas de teobromina, catequina, procianidina e anidrido ftálico continuaram a serem sinalizadas nesta etapa, e surgiram sinais moleculares que foram atribuídos à adenina, tirosina e fenilalanina.

Estudos demonstraram que a teobromina é a estrutura-base na formação de moléculas como a cafeína (SUZUKI; ASHIHARA; WALLER~, 1992). Vale ressaltar, que a molécula de cafeína foi determinada no processo de desreplicação molecular, porém, não é contemplada nesta discussão pois não apresentou afinidade pelos receptores destacados. A teobromina é um composto encontrado naturalmente no cacau, sendo amplamente presente e responsável pelo sabor amargo (JAIN et al., 2020).

Ao comparar os extratos do ponto experimental inicial, constatou-se que a média da intensidade da teobromina era de $5,29x10^4$, e após 84 horas de fermentação, a intensidade média calculada aumentou aproximadamente uma ordem de grandeza, atingindo $3,93x10^5$. No entanto, o modelo proposto para a intensidade relativa desse metabolito nesse ponto experimental contempla 5 coeficientes (dos 10 possíveis para o modelo cúbico completo), com destaque para os coeficientes individuais A, B e C, que apresentam significância positiva para o modelo. Quando comparados com o valor z, o coeficiente A está 1,64 desvio-padrão da média dos efeitos do modelo, enquanto B se distancia apenas 1,03 desvio-padrão e o C, está mais próximo da média dos demais efeitos, deslocando-se positivamente apenas 0,67 desvio-padrão.

No entanto, quando as interações ocorrem, sejam elas interações quadráticas (AC) ou cúbicas (AB(A-B)), seus valores são negativos, indicando que quanto maiores forem os valores atribuídos às interações entre os solventes, menor será a intensidade observada para a teobromina nessa fase da fermentação do cacau. Sobretudo, a interação AB(A-B), que está abaixo da média dos demais efeitos por 1,64 desvios-padrão, pode causar a diminuição da intensidade da teobromina quando a diferença entre A e B foram maiores. Apesar de ser possível visualizar um resíduo atípico referente ao ensaio 10, que poderia ser um outlier, o histograma criado para análise residual apresenta um comportamento próximo do normal, sinalizando que esses cinco coeficientes apresentados permitiram a construção de um modelo de regressão cúbico completo, sem falta de ajuste, com um coeficiente de determinação de 0,88. As análises da teobromina em 84 horas de fermentação podem ser consultadas na Figura Suplementar 29.

Os flavonoides anotados permearam todas as fases do processo de fermentação. Quando analisada especificamente a procianidina, observa-se que a intensidade média se mantém estável desde o início até a fase intermediária da fermentação, apresentando uma ordem de grandeza de 10^3. Contudo, ao tentar desenvolver um modelo que pudesse prever a intensidade relativa deste metabólito ao longo dos diferentes ensaios, modelos com apenas um coeficiente foram sugeridos. O modelo mais promissor apresentou um coeficiente de determinação

(R^2) de apenas 0,18. Tal valor é considerado insuficiente para fornecer previsões confiáveis e, portanto, esse modelo não foi mais explorado no estudo.

Por outro lado, para a molécula de catequina, verificou-se uma redução na média das intensidades relativas do composto quando comparado com o ponto inicial. Na primeira desreplicação da catequina, foi apontado média das intensidades relativas dos 14 ensaios que agruparam o planejamento de mistura do tipo Simplex-Lattice era da ordem de 10^5, e no segundo ponto amostral, caiu uma ordem de grandeza, para 10^4. Porém, nesta etapa da fermentação, as intensidades relativas utilizadas como resposta não permitiram a construção de um modelo adequado para fazer a previsão das respostas, obtendo apenas um modelo cúbico especial, em que apenas o coeficiente relacionado ao solvente acetato de etila (B) teve importância, com valor de 2,29 e deslocado positivamente 1,46 desvios-padrão da média dos demais efeitos, mas para relevância das análises de regressão, não; obtendo coeficiente de determinação (R^2) de apenas 0,26, impossibilitando de criar um modelo que previsse a intensidade da catequina na fase intermediária do processo de fermentação do cacau. A análise estatística da catequina nesta fase do processo de fermentação está agrupada na Figura Suplementar 30.

Durante a etapa intermediária da fermentação, o sinal associado ao ácido ftálico desapareceu, permanecendo em evidência apenas a molécula de anidrido ftálico, conforme identificado pela plataforma LUMIOS. Nota-se que a intensidade relativa média dessa molécula se manteve constante, com a mesma ordem de magnitude. No entanto, o modelo cúbico completo para esta etapa apresentou um coeficiente de determinação (R^2) de 0,82. Esse valor, embora alto, é significativamente menor se comparado à fase inicial, que apresentava um coeficiente de 0,99.

O modelo cúbico completo proposto para o anidrido ftálico, para esta fase, foi constituído por somente três coeficientes significativos. Entre estes, a interação cúbica modulada pela diferença entre os coeficientes A e B (AB(A-B)) se destacou. Essa interação indica que, à medida que a diferença entre A e B aumenta, a intensidade dos sinais do anidrido ftálico também cresce. Contrapondo-se a isso, a interação ABC merece atenção, pois quando os componentes A, B e C estão

associados, resulta na diminuição da intensidade do anidrido ftálico. O modelo cúbico completo sugerido foi capaz de explicar 82% da variação nos dados desta anotação molecular presente nas matrizes de cacau.

No entanto, ao analisar os resíduos, verificou-se que o histograma mostra uma cauda estendida para a direita, indicando potencial assimetria nos resíduos. Essa assimetria nos resíduos sugere que o modelo pode não estar completamente ajustado aos dados e que a falta de ajuste pode ser um indicativo de que existem variáveis não consideradas ou interações não capturadas pelo modelo atual, mas que dá indícios de que a fase intermediária do processo fermentativo gera mais complexidade nas matrizes de cacau, exigindo necessidade de uma avaliação mais aprofundada para otimizar a resposta nesse nesta etapa do processo de fermentação.

Complementarmente, estudos científicos, como os de Deus (DEUS et al., 2021), têm identificado e analisado os compostos presentes no cacau fermentado, incluindo a identificação da presença de aminoácidos e aminas bioativas, como adenina e tirosina. A adenina é uma purina encontrada naturalmente em várias fontes, incluindo plantas, e tem sido detectada em cacau fermentado. Estudos relatam a presença de adenina em concentrações variáveis nas sementes de cacau durante a fermentação (KIEFER et al., 1983; ZHENG et al., 2004), além de considerem-na como um metabólito em potencial para efeito anti-inflamatório e imunomodulador (ZHU et al., 2012).

Nesta análise, a adenina apresenta média de intensidade em $2{,}50 \times 10^4$, porém, vale sinalizar que o modelo estatístico criado para previsão de tais respostas apresentou coeficiente de determinação mediano, de 0,66, além de dar indícios de falta de ajuste do modelo, o qual tem apenas a contribuição do coeficiente B (relacionado ao solvente acetato de etila) como significativo, justificando a baixa precisão, e não sendo, portanto, um modelo cúbico, pois as interações deste tipo de modelo não são significativas. Pela análise do gráfico de contorno, nota-se que a região dos ensaios 2 e 12 (na extremidade direita) são regiões que apresentam intensidades mais relevantes, sinalizando que além do solvente de acetato de etila, as combinações de solventes relacionadas ao décimo segundo ensaio (com 66,67%

de acetato de etila) é a mais expressiva para extração deste metabólito. Durante a análise dos resíduos gerados pelo modelo, não há uma distribuição homogênea dos dados, além da existência de alguns pontos atípicos e algumas correlações (ensaios 4 a 8, Figura Suplementar 32 – F e G), dando indícios de que padrões residuais como esses podem indicar uma relação não capturada pelo modelo ou a presença de variáveis omitidas.

Da mesma forma, a identificação de fenilalanina e tirosina (um aminoácido não essencial), também foi relatada em processos fermentativos de cacau. A conversão de fenilalanina em tirosina é chamada de "hidroxilação da fenilalanina". Esta reação é catalisada pela enzima fenilalanina hidroxilase (GARIBOTTO et al., 2002). A tirosina desempenha um papel importante na síntese de compostos fenólicos, como as catequinas, que contribuem para as características sensoriais em produtos do cacau fermentado (FEDURAEV et al., 2020). Além disso, à tirosina, é um aminoácido precursor de neurotransmissores importantes, como a dopamina e a noradrenalina, que desempenham um papel crucial na função respiratória (FERNSTROM; FERNSTROM, 2007). Estudos indicam que a tirosina pode ter efeitos benéficos na regulação da respiração e no alívio de sintomas respiratórios, especialmente em condições como a apneia do sono (SVATIKOVA et al., 2004) e a doença de Parkinson (DIFRANCISCO-DONOGHUE et al., 2014), nas quais há disfunção respiratória associada.

É importante destacar que a tirosina foi observada apenas nesta fase do estudo, não sendo encontrada nos pontos iniciais ou finais, já a fenilalanina, aparece na fase intermediária e é evidenciada ainda no processo final, sinalizando que durante a fase intermediária que ocorre a maior parte das conversões de uma estrutura em outra.

Para a molécula de fenilalanina, foi possível estabelecer um modelo cúbico completo que conseguiu elucidar 85% da variação nos dados coletados. Este modelo é composto por sete coeficientes significativos. Dentre eles, merece destaque o coeficiente associado ao solvente hexano (A). Esse coeficiente se destaca positivamente, estando 1,64 desvios-padrão acima da média, indicando que à medida que o valor deste coeficiente aumenta, o sinal associado a este

metabolito torna-se mais intenso. Em contrapartida, a relação entre os coeficientes A e B, representada pelo coeficiente AB, sugere uma relação inversa: quanto maior o valor deste coeficiente, menor a intensidade do sinal para a fenilalanina.

No entanto, apesar da capacidade do modelo cúbico completo em explicar uma grande parte da variância, há evidências de falta de ajuste. Uma análise mais aprofundada do histograma residual revelou uma assimetria com uma cauda estendida para a esquerda. Esta cauda sugere que o modelo pode estar superestimando os valores em determinadas condições ou que pode haver variáveis ou interações não consideradas no modelo atual. A presença dessa cauda no histograma residual é um indicativo de que o modelo pode não estar completamente ajustado aos dados. A análise estatística da fenilalanina está contida na Figura Suplementar 33.

Para a tirosina, o modelo gerado é bastante similar ao da fenilalanina, porém, com nove coeficientes significativos. Além disso, ao analisar o gráfico de contorno gerado pelo modelo cúbico completo para prever as intensidades de tirosina, observamos que os experimentos 13 e 14, localizados no centro, apresentaram maiores valores de intensidade. Embora o modelo não tenha um ajuste perfeito, foi possível realizar uma regressão com um coeficiente de determinação (R^2) de 0,89, no qual 9 dos 10 coeficientes são significativos na construção do modelo de previsão.

Apenas o coeficiente AC(A-C) não demonstrou relevância para o modelo associado à tirosina. Os efeitos dos coeficientes A, B, C, AC, BC e ABC têm valores positivos no modelo. Com destaque para o efeito C (7,05) que se distancia positivamente 1,64 desvios-padrão da média dos demais efeitos, indicando que a intensidade da tirosina tende a aumentar caso esses coeficientes também se elevem em conjunto. Por outro lado, os demais coeficientes têm um efeito oposto, com destaque para AB, que é a mais negativa, distanciando-se da média dos efeitos na mesma proporção que o coeficiente C, além de AB(A-B) e BC(B-C), que quando aumentam, a intensidade da tirosina diminui. Porém, com a associação conjunta desses 9 coeficientes, gera-se uma distribuição residual equilibrada e um

histograma capaz de agrupar os resíduos praticamente como uma distribuição normal.

3.3 Ponto final do processo fermentativo de cacau – 168 horas

Finalmente, após 168 horas de fermentação, observou-se uma redução no número de anotações, com apenas seis estruturas apresentando resultados promissores nos testes de docagem, sugerindo que tais anotações possam ser candidatas potenciais a interagirem com receptores relacionados a doenças respiratórias. Dentre tais estruturas, somente a anotação molecular referente ao indol-3-acetamida foi apontada exclusivamente na fase final da fermentação.

Quanto à molécula de adenina, o processo de desreplicação continuou sinalizando esta anotação na fase final da fermentação, apresentando média de intensidades praticamente igual à fase anterior, considerando os 28 ensaios analisados. No entanto, desta vez, mesmo com a análise de variância indicando falta de ajuste, foi possível construir um modelo estatístico mais preciso, com um coeficiente de determinação de 0,98, porém, apresentando uma distribuição residual com algumas tendências (como os ensaios de 6 até 10), mas, com um histograma melhor, com mais proximidade a uma distribuição normal dos dados, mesmo apresentando-se um pouco mais deslocado para a cauda esquerda da curva.

Nesse modelo cúbico completo, seis dos dez coeficientes foram considerados estatisticamente significativos, em contraste com o modelo sugerido na fase anterior, que tinha apenas um coeficiente relevante estatisticamente e um coeficiente de determinação de apenas 0,66. Dos coeficientes presentes no modelo de previsão da intensidade relativa da molécula de adenina, três coeficientes significativos apresentaram valores negativos (AB, BC e AB(A-B)), o que indica que maiores valores associados a essas interações resultam em menor intensidade da anotação de adenina, com destaque para o coeficiente BC (-9,81) e que distancia-se negativamente da média dos efeitos por 1,69 desvios-padrão. Pelo lado oposto, na mesma proporção, aponta-se o coeficiente B (25,52), contribuindo de maneira positiva para a previsão da intensidade desta anotação. A adenina permaneceu com

a mesma ordem de grandeza na fase final da fermentação em comparação com a etapa anterior, mas houve um leve aumento na intensidade, que agora é de $3,66x10^4$ (anteriormente $2,50x10^4$). O resumo da análise estatística referente ao modelo cúbico completo para predição das intensidades relativas da adenina está apontado na Figura Suplementar 35.

A presença da indol-3-acetamida na fase final do processo de fermentação do cacau é um achado molecular interessante, especialmente considerando sua afinidade por alvos relacionados a doenças respiratórias efetuadas no software LUMIOS. Embora compostos derivados da indol-3-acetamida tenham sido estudados e aplicados como agentes anti-hiperglicêmicos e antioxidantes (KANWAL et al., 2021), ainda não há relatos sobre seu uso específico no contexto de doenças respiratórias. A construção de um modelo estatístico baseado nas intensidades relacionadas a esse metabólito revelou uma falta de ajuste, o que indica que outros fatores podem influenciar sua extração e intensidade durante a fermentação do cacau.

No entanto, o coeficiente de determinação relativamente alto de 0,92 indica uma boa capacidade do modelo em explicar as variações nas intensidades observadas, uma vez que trata-se de um modelo com 6 coeficientes significativos em sua predição. Isso sugere que existem fatores específicos que afetam a presença e a concentração da indol-3-acetamida durante a fermentação, além de indicar que o acetato de etila se mostra mais eficaz na extração desse metabólito em comparação com hexano e etanol puros, que não tem significância para a modelagem estatística, porém, apresenta coeficientes de interação quadrática e cúbica, que são importantes para o modelo, têm valores negativos (quando não estão padronizados para a criação do gráfico de Pareto), a exemplo das interações BC, ABC, AB(A-B) e BC(B-C), que sinaliza que caso essas interações tenham valores mais expressivos, vão influenciar negativamente a intensidade desta anotação nos extratos finais do cacau, sobretudo o coeficiente modulado AB(A-B), que apresenta grande influência para o modelo, pois seu efeito está 1,64 desvios-padrão abaixo da média que os demais efeitos causam no modelo.

Em contraste, na mesma proporção de distanciamento da média, encontra-se o coeficiente B, associado ao solvente acetato de etila, que contribui de maneira mais expressiva no aumento da intensidade do composto indólico anotado. Porém, os resíduos gerados pelo modelo apresentam algumas tendências, correlações e pontos atípicos, como os resíduos produzidos pelo ensaio 12. Ademais, o histograma que apresenta a distribuição residual está mais deslocado para a direita, gerando uma cauda no gráfico de distribuição normal, indicando a presença de outliers ou uma inadequação do modelo em capturar a estrutura dos dados, corroborando com pontuações da análise de variância que sinaliza uma falta de ajuste do modelo preditivo deste composto.

Ao observar o mapa de contorno construído a partir do modelo, fica evidente que as maiores intensidades da indol-3-acetamida estão associadas principalmente aos ensaios 4 e 2 dos experimentos Simplex-Lattice. Isso sugere que as condições experimentais nesses ensaios foram mais propícias para a produção ou preservação desse metabólito durante a fermentação do cacau. No entanto, é importante destacar que essa descoberta da presença de indol-3-acetamida na fase final da fermentação do cacau e a construção do modelo estatístico fornecem uma base para investigações futuras. Estudos mais aprofundados são necessários para entender completamente o papel e o potencial terapêutico desse metabólito em relação a doenças respiratórias, uma vez que nosso estudo sinaliza uma análise em *in silico,* analisando com 4 proteínas associadas a esse tipo de doença.

As moléculas de teobromina e catequina foram anotações sinalizadas em todas as etapas do processo fermentativo. Para a xantina, a intensidade média associada ao final do processo fermentativo apresenta a mesma ordem de grandeza da etapa anterior, porém, para este caso, foi possível construir um modelo com 7 (dos 10 coeficientes possíveis) com um R^2 de aproximadamente 0,98, mesmo sinalizado pela análise de variância como tendo falta de ajuste e uma distribuição residual mais tendenciosa para o lado direito da curva gaussiana.

Para as intensidades da teobromina na fase final do processo fermentativo, os três coeficientes principais, A, B, C, são representativos para o modelo, com destaque para o B (associado ao acetato de etila), que se distancia positivamente

da média dos demais efeitos por 1,64 desvio-padrão. Há presença de coeficientes que representam interações cúbicas associadas, como AB(A-B), que apresenta a mesma distância proporcional ao coeficiente B, mas que apresentam valores negativos, dando indícios de que quanto maior a diferença dos valores entre A e B, menores as intensidades relacionadas a esse metabólito. Do mesmo modo, quanto maior os valores obtidos para a interação BC, a intensidade relativa desta anotação pode sofrer variações negativas significativas.

O perfil das intensidades relativas da molécula de catequina também apresentou excelente coeficiente de determinação, 0,98, utilizando para a consolidação do modelo 4 dos 10 coeficientes possíveis para o cúbico completo, destacando a interação principal significativa (B), com valor de 23,65 e um valor de z de 1,64, relacionada ao acetato de etila, ou seja, quanto maior o valor deste efeito, mais expressiva é a intensidade da catequina na fase final da fermentação do cacau. Porém, os outros 3 coeficientes que apresentam significância para predição do modelo apresentam valores negativos, principalmente as interações cúbicas moduladas pela diferença entre os fatores, tanto para AB(A-B) (que é o efeito negativo mais afastado da média dos coeficientes) quanto para BC, que quanto maiores as diferenças entre os fatores, mais negativos serão o conjunto dessas interações, e afetarão a resposta do modelo de modo negativo, reduzindo a intensidade da catequina nos extratos. As figuras suplementares Figura Suplementar 38 e Figura Suplementar 39 apresentam os resultados da teobromina e catequina, respectivamente.

Para o modelo cúbico completo relacionado à procianidina, o coeficiente de determinação foi de 0,99, ressaltando sete coeficientes significativos, com destaque para o B, com valor de 83,82, que se distancia positivamente 1,64 desvios-padrão da média dos demais coeficientes, sendo único coeficiente simples e significativo do modelo, indicando que o solvente de acetato de etila é mais importante para extração da procianidina nas misturas complexas de cacau após 168 horas de fermentação. Destaca-se também os coeficientes quadráticos BC, AB, BC(B-C) e AB(A-B), que são negativos, mostrando que quanto mais interações existem, menores são as intensidades dos sinais dessa molécula nos extratos.

Por outro lado, quanto maior a diferença entre A e C, mais expressiva é a intensidade do sinal (coeficiente AC(A-C)), assim como ABC, que também aumenta a intensidade relativa desta molécula. Apesar do alto valor de coeficiente de determinação, o modelo apresenta uma falta de ajuste, conforme evidenciado pelo histograma residual, que não segue uma distribuição normal e apresenta um deslocamento para a esquerda (Figura Suplementar 40-G). Isso sugere que o modelo pode não ser completamente adequado para descrever o comportamento das variáveis envolvidas na extração de procianidina, mas por se tratar de uma matriz complexa, pode ser utilizado este modelo para extrair essa determinada molécula, mas há um indicativo de que outros fatores não considerados podem estar influenciando o processo.

Por fim, para a molécula de fenilalanina, a média da intensidade dos sinais apresentou a mesma ordem de grande da fase anterior (10^4), porém, em comparação com o modelo de 84 horas, neste, 9 coeficientes foram significativos, com destaque para o coeficiente B, sinalizando que o solvente de acetato de etila é mais eficiente para extração desta molécula nas matrizes complexas do cacau no final da fermentação, obtendo um coeficiente de determinação de 0,99. A análise de variância indica a falta de ajuste do modelo, que pode ser traduzia na distribuição dos resíduos deixados pelo modelo, que pode ser consultada no histograma da Figura Suplementar 36-G.

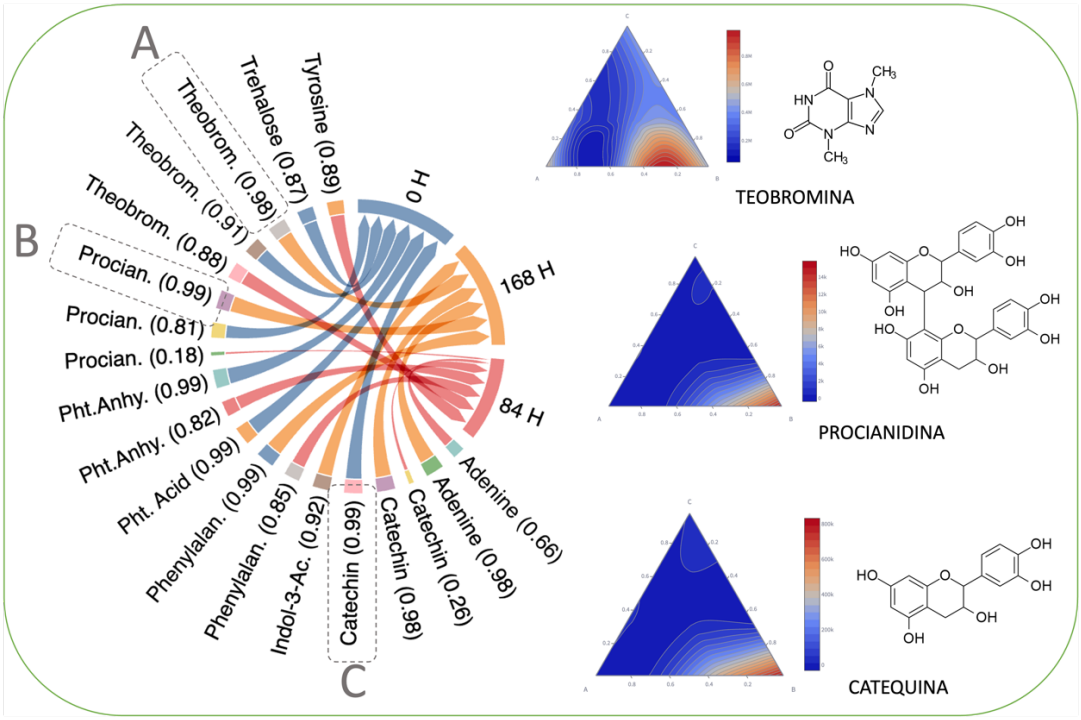
Durante o complexo processo fermentativo, diversas mudanças químicas ocorrem, conduzidas, sobretudo, por ação microbiana. As moléculas de catequina, procianidina e teobromina são consideradas marcadores do cacau, pois estão presentes em todas as etapas do processo fermentativo. A análise dessas três moléculas marcadoras ao longo da fermentação revela diferenças interessantes na adequação dos modelos utilizados para descrever suas concentrações ao longo do tempo. Para a catequina, o modelo mais preciso é obtido no início do processo fermentativo, ou seja, em 0 hora, quando as sementes de cacau ainda não foram submetidas à fermentação. Isso sugere que a concentração de catequina pode ser mais facilmente prevista no início do processo.

Por outro lado, para as moléculas de procianidina e teobromina, os modelos mais robustos são obtidos no final do processo fermentativo. Isso pode ser devido à sua maior estabilidade ou à sua formação ao longo do processo de fermentação, tornando-as mais previsíveis nesse estágio.

Embora o modelo mais preciso para a catequina seja obtido no início do processo, também se observa um modelo robusto para essa molécula no final da fermentação (com R^2 de 0,98). Isso sugere que, embora a concentração de catequina possa ser mais facilmente prevista no início do processo, ainda é possível obter um modelo robusto no final da fermentação.

A presença e a concentração de moléculas marcadoras, como catequina, procianidina e teobromina, são vitais para a qualidade final do cacau e dos produtos derivados, bem como fonte de moléculas de interesse farmacológico. A capacidade de modelar as concentrações dessas moléculas em diferentes estágios da fermentação é, portanto, de grande importância. Os resultados indicam que, enquanto a concentração de catequina é mais bem modelada no início e no final do processo fermentativo, as concentrações de procianidina e teobromina são mais bem previstas no final da fermentação. A Figura 28 faz a representação de todas as moléculas desreplicadas ao longo do processo fermentativo, com ênfase aos três marcadores moleculares.

Figura 28 - Indicações das anotações que apresentaram os melhores modelos nos diferentes pontos experimentais.



Fonte: Elaborado pelo autor (2023).

4 CONSIDERAÇÕES

Algumas anotações apresentaram alto coeficiente de determinação, indicando que o modelo estatístico utilizado no planejamento do experimento está bem ajustado aos dados observados, como é o caso da trealose nos extratos não fermentado e a teobromina, em 84 horas. No entanto, constatou-se falta de ajuste nos demais modelos, além das análises dos resíduos contribuírem para avaliar a qualidade deste ajuste, identificar possíveis violações das suposições do modelo e fornecer insights sobre melhorias ou ajustes necessários, é importante considerar esses aspectos para garantir a validade e robustez das conclusões obtidas a partir do modelo.

Isso sugere que tais modelagens não está capturando completamente a variabilidade dos dados ou não é apropriado para descrever o fenômeno em questão, uma vez que tratam-se como resposta a intensidade relativa dos compostos em fases bastante complexas da fermentação natural do cacau, isso pode ocorrer em situações em que o modelo estatístico captura bem a variação global dos dados, mas não é capaz de explicar todos os detalhes e variações locais, sobretudo dos efeitos antagonistas e de sinergismo que pode ocorrer entre os solventes. Dessa forma, mesmo com um R^2 alto, pode haver discrepâncias entre os valores observados e os previstos pelo modelo para determinados pontos ou regiões do espaço de mistura.

A falta de ajuste do modelo pode ser resultado de fatores não considerados no modelo estatístico, como interações complexas entre os componentes da mistura ou influências de variáveis não controladas. Além disso, pode haver limitações inerentes ao próprio modelo escolhido ou em relação ao delineamento experimental, porém, o planejamento de mistura Simplex-Lattice é uma abordagem peculiar e possui características próprias, permitindo uma exploração mais ampla do espaço de mistura, pois utiliza uma grade triangular (lattice) que é replicada e deslocada de maneira sistemática (SQUEO et al., 2021).

Esta abordagem de delineamento experimental foi selecionada neste trabalho porque possibilita uma distribuição equilibrada dos pontos experimentais

ao longo do espaço de mistura, garantindo uma cobertura adequada e representativa das regiões de interesse.

É importante ressaltar que a falta de ajuste não invalida necessariamente os modelos, mas indica que existem nuances e complexidades matemáticas não capturadas por ele. Essas informações podem ser úteis para orientar investigações adicionais e refinamentos no modelo, a fim de melhorar sua capacidade de prever e explicar os resultados observados.

Neste trabalho, o planejamento Simplex-Lattice foi utilizado como um método para otimizar a combinação de diferentes componentes ou fatores em uma mistura extremamente complexa, como no caso do cacau, onde diferentes solventes podem ser usados para geração dos extratos brutos obtidos da fermentação. O planejamento utilizado buscou explorar diferentes proporções e combinações desses solventes para obter o melhor resultado possível na exploração de dados químicos provenientes da fermentação natural do cacau.

O planejamento de misturas do tipo Simplex-Lattice, assim como em outros tipos de planejamentos, pode apresentar algumas limitações (CHASALOW; BRAND, 1995; PIEPEL; CORNELL, 1987), sobretudo, quando aplicado a misturas naturais complexas, sobretudo oriundas de processos de fermentações naturais. Essas limitações podem ser consultadas na Figura 29.

Figura 29 - Descrição de possíveis limitações do delineamento experimental do tipo Simplex-Lattice.

Fonte: Elaborado pelo autor (2023).

Para superar essas limitações, abordagens mais avançadas e flexíveis, como planejamentos de misturas mais complexos, modelagem não linear e métodos de análise de dados multivariados (PIEPEL; CORNELL, 1994), podem ser necessários para investigar adequadamente as misturas naturais complexas.

Como conseguinte, os resultados aqui descritos permitiram destacar a importância das interações entre os solventes e os coeficientes moleculares na extração e no comportamento das moléculas estudadas na complexa matriz fermentada do cacau. Além disso, ressaltaram a necessidade de abordagens mais refinadas e específicas para uma compreensão completa das características dessas substâncias. Futuros estudos poderão explorar a influência de outros solventes e interações moleculares para obter insights adicionais sobre essas moléculas e sua aplicabilidade em diferentes contextos.

Reforça-se que esses resultados demonstraram a variação das anotações moleculares ao longo do processo de fermentação do cacau e forneceram

informações promissoras para o desenvolvimento de terapias, perspectivas e oportunidades direcionadas ao campo da pesquisa da saúde respiratória. No entanto, é importante ressaltar que as pesquisa aplicadas os metabólitos aqui sinalizados estão em estágios preliminares e ainda são necessários estudos adicionais, incluindo ensaios clínicos, para confirmar a eficácia e a segurança delas no tratamento de doenças respiratórias.

5 CONCLUSÃO

Esses resultados ressaltam a importância da análise de matrizes naturais complexas, como o cacau, na busca por moléculas bioativas com potencial terapêutico. Ademais, o uso de abordagens estatísticas robustas, como o planejamento de misturas e a modelagem dos dados, oferecem uma ferramenta eficaz para a identificação de moléculas promissoras e a compreensão das interações moleculares relevantes em doenças respiratórias. A fermentação tem sido conduzida como principal forma do produto do cacau, chocolate, porém, metabolitos de interesse comercial podem ser obtidos diretamente do fruto, como é o caso de teobromina, catequina e procianidina.

Os resultados obtidos neste estudo demonstraram a pertinência do aplicativo Chemistika (https://chemistika.streamlit.app/) na análise das sinalizações moleculares com afinidade por bioreceptores relacionados a doenças respiratórias. Através do cálculo automatizado do planejamento de mistura do tipo Simplex-Lattice, com 14 pontos experimentais (realizados em duplicata), foi possível construir modelos estatísticos que permitiram a otimização e identificação das melhores condições de fermentação para obter as intensidades relativas mais favoráveis para cada molécula de interesse.

Esses modelos estatísticos forneceram informações valiosas para a seleção da etapa mais promissora no processo de fermentação, indicando em qual momento é mais provável identificar e extrair as moléculas de interesse com maior eficiência. Essa abordagem de planejamento de mistura permitiu uma análise sistemática das variáveis envolvidas no processo fermentativo do cacau, levando em consideração a interação entre os solventes utilizados (hexano, acetato de etila e etanol) e as sinalizações moleculares de interesse.

Esses resultados constituem uma base sólida para futuros estudos, proporcionando direcionamento para a investigação aprofundada dessas moléculas e seu potencial uso em aplicações farmacológicas relacionadas a doenças respiratórias. Entre as anotações moleculares exploradas, apenas as análises estatísticas associadas à molécula de adenina, catequina e procianidina na fase de

84 horas dos experimentos não permitiram a construção de um modelo estatístico adequado, porém, foi possível obter modelos com alto coeficiente de determinação na fase final do processo fermentativo tanto para adenina quanto para catequina.

Em suma, o aplicativo Chemistika, associado ao LUMIOS, mostrou-se uma ferramenta valiosa na análise das sinalizações moleculares com afinidade por bioreceptores, permitindo o planejamento de mistura e a construção de modelos estatísticos que contribuíram para a otimização e identificação das melhores condições de fermentação do cacau. Esses resultados abrem caminho para investigações futuras no desenvolvimento de produtos farmacológicos relacionados a doenças respiratórias, aproveitando as moléculas identificadas durante o processo fermentativo do cacau.

CAPÍTULO 5 – ANÁLISE DA VARIABILIDADE METABÓLICA DAS MATRIZES COMPLEXAS DE CACAU UTILIZANDO O SOFTWARE CHEMISTIKA.

RESUMO

Neste capítulo, utilizou-se o software Chemistika para analisar planejamentos de misturas, visando explorar a variabilidade metabólica das matrizes complexas de cacau fermentado em três diferentes momentos: 0 horas, 84 horas e 168 horas. Empregando o método Design Simplex-Lattice 3x3 (SLD 3x3), foram projetadas misturas dos solventes hexano (A), acetato de etila (B) e etanol (C), as quais foram utilizadas para obter os extratos de cacau. A quantidade de sinais de cada extrato, obtidos por meio da técnica de espectrometria de massas, foi utilizada como resposta para explorar o perfil químico de cada ponto experimental. Observou-se que, inicialmente, no ponto de 0 horas, havia quantidade mais acentuada de sinais moleculares, indicando uma composição química mais complexa. Conforme o processo de fermentação avançou, houve alterações progressivas no número de picos, evidenciando alterações químicas ocorridas durante o processo de fermentação. O SLD 3x3 utilizado contemplou modelos cúbico completos para cada etapa do processo de fermentação, com coeficientes de determinação significativos e relevantes, de 0,98 para o estágio inicial da fermentação, associado a 0 hora (sementes não fermentadas), 0,82 e 0,83 para 84 e 168 horas de fermentação, respectivamente. Esses resultados forneceram insights valiosos sobre as mudanças químicas que ocorreram durante a fermentação das sementes de cacau e destacaram a importância de compreender as características dos extratos em diferentes estágios de fermentação. O SLD 3x3, gerado pelo Chemistika, permitiu uma abordagem mais abrangente ao investigar o perfil químico e a evolução das propriedades dos extratos oriundos da fermentação das sementes de cacau, bem como visualizar as melhores condições para obtenção de extratos com quantidades de sinais mais expressivas, o que poderia facilitar os trabalhos de desreplicação molecular, quando necessário.

Palavras-chave: Simplex-Lattice; automatização de análises, software, tratamento de dados, modelos estatísticos

1 INTRODUÇÃO

A utilização da polpa fermentada do cacau (*Theobroma cacao* L.) não se limita apenas à produção de chocolate, mas também volta-se à obtenção de fitoquímicos com perspectivas de diversas aplicações industriais e/ou biotecnológicas. Estudos recentes têm explorado as propriedades bioativas da polpa fermentada de cacau, destacando seu potencial como fonte de compostos fenólicos e outros fitoquímicos com propriedades antioxidantes e anti-inflamatórias (ANDÚJAR et al., 2012; BELWAL et al., 2022).

Tais moléculas têm despertado interesse por investigações em campos como a Medicina, por meio de aplicações farmacêuticas, nutrição e Ciência dos Alimentos, devido aos seus potenciais benefícios para a saúde humana, incluindo atividades cardioprotetoras, neuroprotetoras e anticancerígenas (APROTOSOAIE et al., 2016; CIMINI et al., 2013; ZIĘBA; MAKAREWICZ-WUJEC; KOZŁOWSKA-WOJCIECHOWSKA, 2019).

Um elemento central deste processo é a fermentação da polpa que recobre as sementes de cacau, desencadeando transformações químicas significativas nesse substrato. A atividade microbiana é fundamental para a formação de uma gama diversificada de compostos químicos, resultantes do metabolismo de açúcares, ácidos orgânicos, proteínas e micronutrientes presentes que envolve as sementes de cacau (BART-PLANGE; BARYEH, 2003).

A polpa de cacau é composta de 82% a 87% de água, 10% a 15% de açúcares, 1% a 5% de pectina, 1% a 3% de ácido cítrico, 0,1% a 0,4% de outros ácidos não voláteis (como ácido málico), 0,5% a 0,7% de proteínas e 8% a 10% de minerais e oligoelementos (PUERARI; MAGALHÃES; SCHWAN, 2012). Dos açúcares presentes, cerca de 60% é sacarose e 39% é uma mistura de glicose e frutose. A concentração de sacarose, glicose e frutose é influenciada pela variedade e pelo estágio de maturação dos frutos; as vagens não maduras contêm uma proporção maior de sacarose, enquanto as vagens maduras contêm principalmente frutose e glicose. O pH da polpa é relativamente baixo (pH 3,0 – 4,0), principalmente devido ao seu teor de ácido cítrico (GUEHI et al., 2010). O alto teor de pectina e

outros polissacarídeos (celulose, hemicelulose, lignina) torna a polpa viscosa, pegajosa e coesa (MOZZI; RAYA; VIGNOLO, 2010).

A fermentação da polpa, aderida às sementes do cacau, envolve basicamente três estágios e é caracterizada por uma sucessão microbiana ao longo do processo e por alterações bioquímicas nas sementes fermentadas. Embora haja inconsistência nos dados sobre a cocobiota, considerável variedade de micro-organismos autóctones foi identificada durante as fermentações de cacau, e basicamente quatro grupos se destacam: leveduras, bactérias láticas (LABs), bactérias acéticas (AABs) e micro-organismos diversos (TAYLOR et al., 2022). Tais micro-organismos são oriundos do processo de pós-colheita, em especial do manuseio de facas para a abertura dos frutos, da superfície das mãos dos trabalhadores que os manejam, das folhas de bananeiras que são inoculadas nas leiras de fermentação e ainda das superfícies dos cochos onde ocorrem os processos fermentativos (DE VUYST; WECKX, 2016a).

Figueroa-Hernández et al. (2019), tomando como base o estudo de Kadow et al. (2015) sinaliza, claramente, os três estágios do processo de fermentação das sementes de cacau. O primeiro estágio da fermentação das sementes de cacau, nas primeiras 24 horas, inclui a despolimerização da pectina pela ação de enzimas pectinóliticas, sintetizadas pelas leveduras, sob condições anaeróbicas, e temperatura em torno 45^{o}C. Além disso, as leveduras metabolizam os açúcares disponíveis em etanol, concomitantemente produzindo dióxido de carbono, ácidos orgânicos e glicerol. *Hanseniaspora*, *Saccharomyces*, *Kluyveromyces* e *Pichia* são os gêneros predominantes nesse estágio da fermentação (PAPALEXANDRATOU et al., 2013).

O terceiro ou último estágio da fermentação das sementes de cacau, as LABs diminuem em população enquanto ocorre um aumento na população de AABs. Esse momento é marcado pela bioconversão exotérmica aeróbica do etanol em ácido acético pela atividade de AABs (CAMU et al., 2008), o que resulta no amarronzamento das sementes de cacau e morte do seu embrião. Estudos têm identificado várias espécies de AABs envolvidas nesse processo, como *Acetobacter pasteurianus* e *Acetobacter aceti*, que são capazes de realizar essa conversão

química (SARBU; CSUTAK, 2019). Para mais, no último estágio da fermentação, bactérias esporulantes do genêro *Bacillus* têm sido isoladas de sementes fermentadas de cacau, entretanto, o papel desse grupo microbiano ainda permanece pouco explorado e conhecido (OUATTARA et al., 2011).

Dessa forma, a compreensão da dinâmica microbiana envolvida na fermentação das sementes de cacau, bem como das rotas metabólicas seguidas por esses micro-organismos passa ser essencial para a prospecção de metabólitos diversos e para o desenvolvimento de estratégias que visem controlar, efetivamente, um processo fermentativo que ainda é considerável selvagem. No entanto, é relevante destacar que a composição da cocobiota pode variar em função das regiões produtoras de cacau, variedades e híbridos de *Theobroma cacao* e até mesmo das práticas de pós-colheita e condições operacionais e estruturais dos processos locais de fermentação (PASSOS; LOPEZ; SILVA, 1984; SCHWAN; WHEALS, 2004).

Com o intuito de compreender o complexo processo fermentativo do cacau, este trabalho explorou racionalmente, por um Design Simplex-Lattice (Simplex-Lattice Design – SLD) 3x3, os exsudatos provenientes de sementes de cacau em diferentes estágios de fermentação. A teoria Simplex-Lattice é aplicada a experimentos envolvendo misturas de '*q*' componentes e visa prever empiricamente a resposta de qualquer combinação desses componentes. Essencialmente, a resposta é influenciada apenas pela proporção dos constituintes na mistura, e não pela quantidade total presente. Essa abordagem permite uma compreensão mais precisa do efeito das proporções dos componentes na resposta observada, independentemente da quantidade total da mistura (GORMAN; HINMAN, 1962). Ao isolar a influência da proporção dos itens variáveis, é possível obter informações valiosas sobre as relações entre os diferentes componentes e a resposta resultante, contribuindo para a otimização e o aprimoramento dos processos de mistura (LAMBRAKIS, 1968).

No SLD 3x3 foram avaliadas misturas extrativas, compostas por três solventes: hexano, acetato de etila e etanol. Ao planejamento foram inseridos pontos adicionais e ponto central, totalizando 14 ensaios para a avaliação de cada

estágio da fermentação (realizados em duplicata): 0 hora (período inicial da fermentação, em que as sementes são consideradas não fermentadas), 84 horas (3,5 dias de fermentação das sementes) e 168 horas (7 dias de fermentação das sementes). Para explorar o perfil químico da fermentação do cacau, em todos os ensaios experimentais foi avaliada, como variável resposta, a quantidade de sinais espectrais obtidos nos extratos dos exsudatos das sementes de cacau nos diferentes estágios de fermentação.

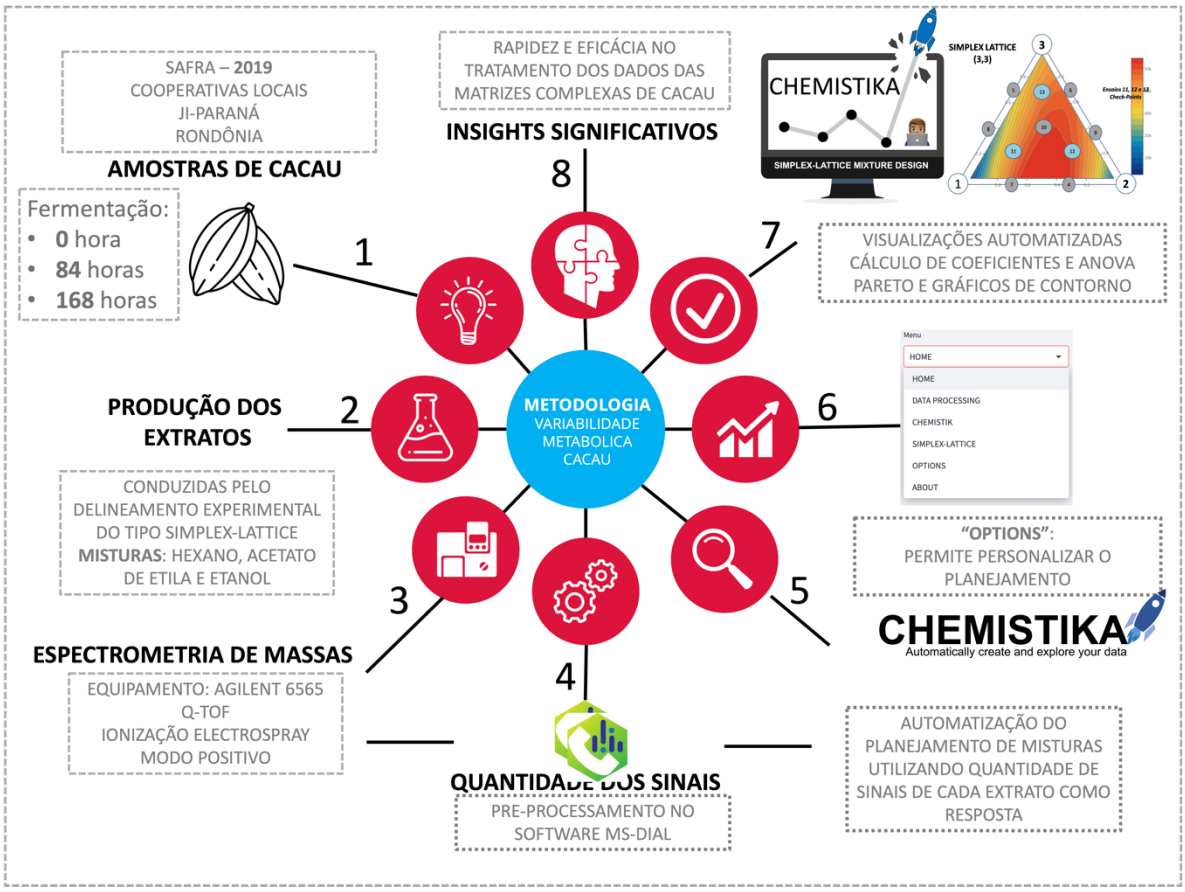
O tratamento estatístico de dados tem sido realizado por softwares comerciais especializados, como o MiniTab (ALIN, 2010) e o Statistica (HILBE, 2007; SARUMATHI et al., 2015) com o objetivo de reduzir o tempo e evitar cálculos longos e tediosos. Com o avanço de novas linguagens de programação e o crescimento da comunidade Python (SRINATH, 2017), surgem novos softwares (gratuitos e de código aberto) que solucionam problemas comuns no ambiente profissional (NOSRATI, 2011) e oferecem maior simplicidade e ferramentas específicas para nichos de aplicação. Dessa forma, o API Chemistika, apresentado no capítulo 3 desta tese, foi empregado para a construção do SLD 3x3 e respectivas análises estatísticas, o que permitiu uma análise direcionada, automatizada e integrada facilitando a exploração de matrizes naturais complexas e adicionando valor aos resultados obtidos.

2 METODOLOGIA

A metodologia desse capítulo foi desenvolvida conforme marcha analítica apresentada no item Metodologia do capítulo 3 (páginas 106 a 110), que descreve a obtenção das matrizes complexas de exsudatos de sementes fermentadas de cacau, sua preparação para análise de massas e a realização das análises utilizando espectrometria de massas. Utilizou-se o software MS-Dial para processamento dos dados e o aplicativo Chemistika para obtenção dos planejamentos associados à quantidade de sinais em cada extrato contido no SLD 3x3. A Figura 30 apresenta um resumo da metodologia aplicada no presente capítulo.

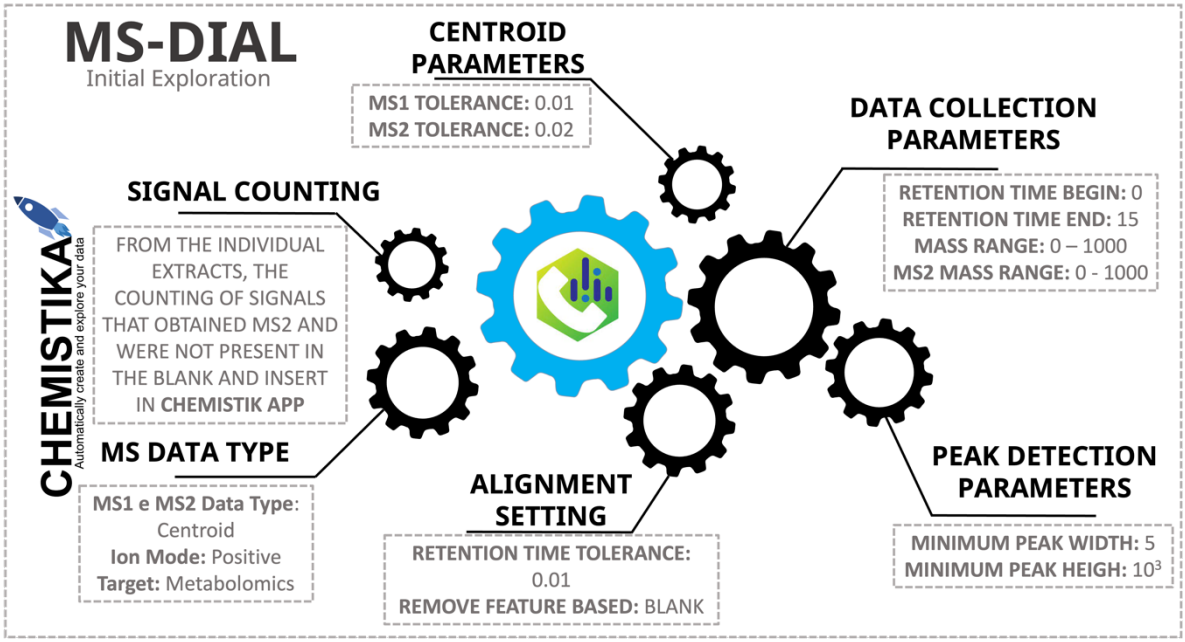
Na ocasião, o processamento dos dados espectrais foi realizado pelo software MS-Dial. O MS-Dial (TSUGAWA et al., 2015) é um software de análise de dados de espectrometria de massas desenvolvido para o processamento e interpretação de dados metabolômicos (e lipidômicos), o qual possibilita a adequação de parâmetros e configurações para auxiliar na análise e identificação de compostos presentes em amostras complexas. O software foi utilizado para obter a quantidade de sinal de cada extrato, levando em conta apenas sinais acima de 10^3 de intensidade e que apresentaram fragmentações MS/MS e não estavam presentes no branco. As demais adaptações aos parâmetros do software podem ser consultadas na Figura 31.

Figura 30 - Resumo da metodologia utilizada para exploração das matrizes complexas das sementes de cacau em diferentes estágios de fermentação a partir dos resultados do SLD 3x3.



Fonte: Elaborado pelo autor (2023).

Figura 31 - Configurações estabelecidas para o software MS-DIAL, utilizado para efetuar a contagem de sinais espectrais identificados nos dos extratos brutos de sementes de cacau em diferentes estágios de fermentação.



Fonte: Elaborado pelo autor (2023).

3 RESULTADOS E DISCUSSÃO

Cada fase da fermentação das sementes de cacau (0 hora, 84 horas e 168 horas) foi examinada individualmente usando a API Chemistika, com o intuito de estudar as diferenças entre os perfis químicos dos extratos. No planejamento Simplex-Lattice 3x3 proposto, a quantidade de sinais serviu como variável resposta neste processo exploratório visando entender como a duração da fermentação das sementes de cacau pode influenciar na variabilidade metabólica.

3.1 Variabilidade metabólica das matrizes complexas de sementes de cacau não fermentadas – ponto 0 hora

A análise da variabilidade metabólica das matrizes complexas de cacau no ponto 0 hora, quando as sementes ainda não passaram pelo processo fermentativo, revelou informações importantes sobre as características químicas presentes nessa etapa inicial.

Dentre os experimentos realizados, destaca-se o extrato proveniente do experimento 2 do SLD 3x3, no qual foi utilizado 100% de acetato de etila (variável B) como solvente extrator. Esse extrato apresentou um perfil químico com cerca de 300 sinais moleculares, indicando uma ampla diversidade de compostos presentes Figura 31-A. Os experimentos 1 (100% hexano – variável A) e 3 (100% etanol – variável C), demonstraram aproximadamente 209 e 161 sinais, respectivamente, revelando menor variabilidade metabólica das condições experimentais conduzidas com apenas um solvente extrator.

É importante ressaltar que o experimento com maior variabilidade metabólica não está diretamente relacionado à quantidade de extrato produzido em termos de massa (em miligramas) ou à presença de moléculas com potencial atividade biológica. A variabilidade metabólica indica a presença de uma ampla gama de moléculas que podem ser isoladas e exploradas em processos de desreplicação, auxiliando na identificação de compostos com propriedades bioativas.

Ao analisar os resultados gerados pelo API Chemistika, constatou-se a possibilidade de construir um modelo estatístico, altamente robusto, com um

coeficiente de determinação (R^2) de aproximadamente 0,98. Esse modelo foi composto por 10 coeficientes significativos (Figura 32-B e 32-C).

Apesar da regressão ser elevada, a análise de variância (ANOVA) sugeriu uma pequena falta de ajuste do modelo utilizado, pois o valor *p* calculado foi de 0,02 (sendo necessário um valor maior que 0,05 para refutar a falta de ajuste) e o F calculado 4,54 (sendo que o tabelado é de 3,28). Essa diferença pode ser um indício de que o modelo estatístico não conseguiu capturar completamente a variação dos dados experimentais e pode estar relacionada à diferença entre as previsões do modelo e a média dos experimentos.

Desse modo, ao elevar essas diferenças ao quadrado e somá-las, obteve-se a soma quadrática associada à falta de ajuste. Para mais, essa falta de ajuste pode ser atribuída a diferentes razões, como a presença de efeitos não lineares, interações complexas entre os componentes da mistura, influência de fatores não considerados no modelo ou até mesmo a complexidade da própria matriz natural.

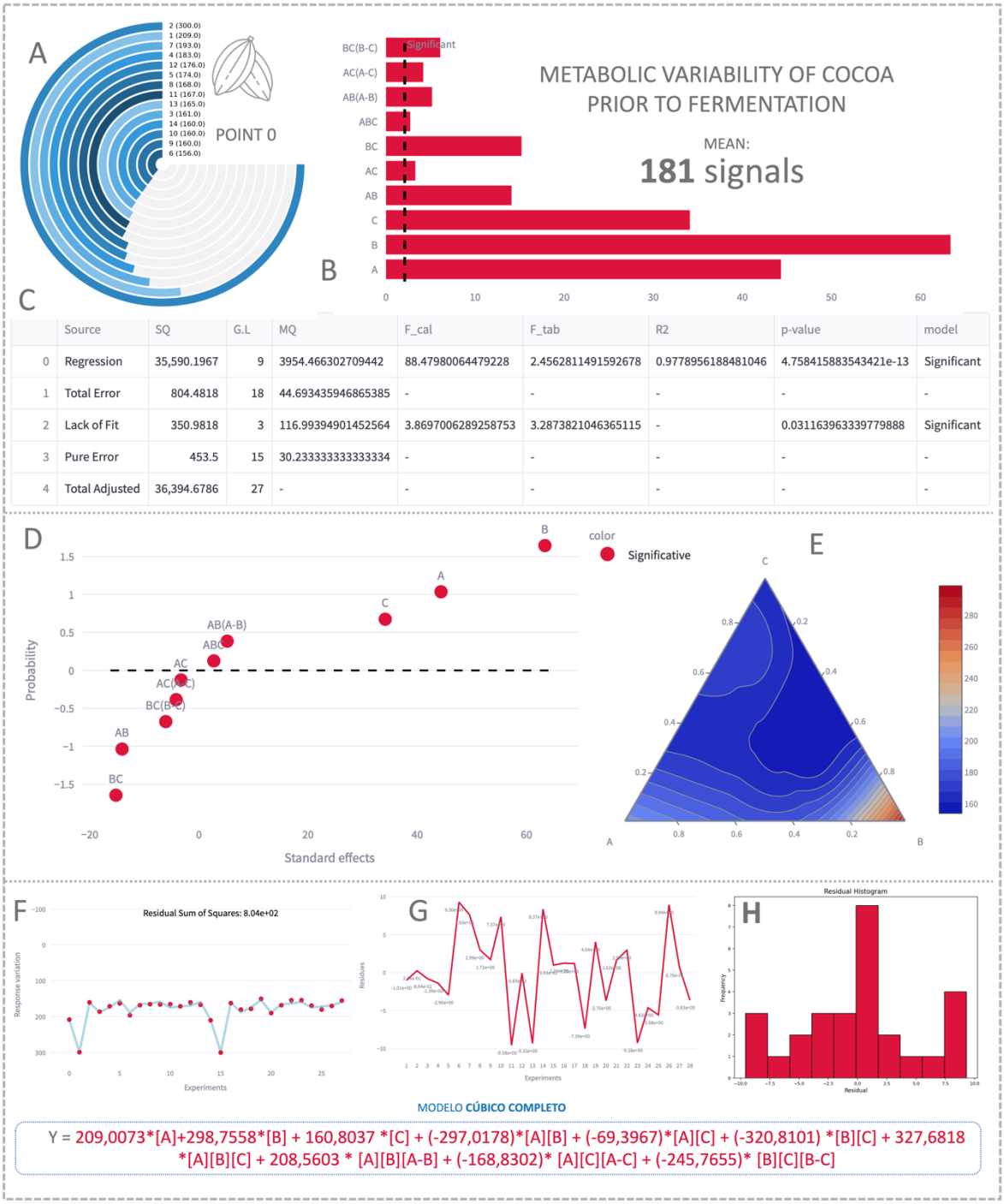
Dentre os 10 coeficientes que compõem o modelo de previsão da quantidade de sinais dos extratos das sementes de cacau não fermentadas, destacaram-se os coeficientes B (acetato de etila) e A (hexano), que corresponderam aos experimentos que produziram as maiores quantidades de sinais (experimentos 2 e 1). O coeficiente B está associado a um valor de z igual a 1,64, o que indica que sua contribuição como efeito está 1,64 desvios-padrão acima da média dos demais. Por sua vez, o coeficiente A se afasta positivamente da média dos experimentos em aproximadamente 1 desvio-padrão.

Por outro lado, existem coeficientes antagônicos que podem reduzir a quantidade de sinais nos extratos das sementes de cacau não fermentadas. Destacaram-se os coeficientes de interação quadrática BC (acetato de etila.etanol) e AB (hexano.acetato de etila), que seguem a mesma proporção dos coeficientes B e A quando comparados aos valores de z. Isso sugeriu que quanto maior a interação secundária entre os solventes BC ou AB, menor será a quantidade de sinais observados. Essas informações foram congruentes com o modelo, que relacionou a interação cúbica modulada pela diferença entre os coeficientes, como o caso da interação AB(A-B) (hexano.acetato de etila.[hexano-acetato de etila]), que

apresentou significância no modelo, indicando que quanto menor a interação entre A e B, maior será a quantidade de sinais nos extratos. Ademais, existiram efeitos negativos causados por outros coeficientes, como AC(A-C) (hexano.etanol.[hexano-etanol]) e BC(B-C) (acetato de etila.etanol[acetato de etila-etanol]), indicando que quanto maior a diferença entre A e C (ou B e C), mais negativa será a contribuição desse coeficiente, atuando de forma adversa para a variabilidade metabólica dos extratos (Figura 32-D).

No mapa de contorno apresentado na Figura 32-E, foi possível observar que a região de resposta mais intensa está associada ao vértice B, cujo coeficiente é maior e mais significativo para o modelo. Essa observação indicou uma tendência do modelo estatístico em não conseguir descrever adequadamente o comportamento dos dados de forma global. Apesar do modelo se aproximar dos pontos reais, conforme demonstrado na Figura 32-F, os resíduos gerados apresentam uma distribuição tanto na parte positiva quanto na parte negativa do gráfico (Figura 32-G), exibindo alguns padrões. Isso indicou a presença de correlações entre os resíduos deixados em alguns experimentos, como nos casos dos experimentos 2-5 e 6-9, que mostraram um comportamento de queda bastante similar. Além disso, o histograma dos resíduos (Figura 32-H) apresentou extremidades mais agrupadas, não seguindo agrupar os dados em uma distribuição gaussiana.

Figura 32 - Análises estatísticas no ponto 0 hora (sementes não fermentadas de cacau). A) Quantidade de sinal em cada ensaio. B) Gráfico de Pareto. C) ANOVA. D) Gráfico de dispersão coef. padronizados. E) Gráfico de Contorno. F) Previsão do modelo. G) Distribuição residual. H) Histograma residual.



	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	35,590.1967	9	3954.466302709442	88.47980064479228	2.4562811491592678	0.9778956188481046	4.758415883543421e-13	Significant
1	Total Error	804.4818	18	44.693435946865385	-	-	-	-	-
2	Lack of Fit	350.9818	3	116.99394901452564	3.8697006289258753	3.2873821046365115	-	0.031163963339779888	Significant
3	Pure Error	453.5	15	30.233333333333334	-	-	-	-	-
4	Total Adjusted	36,394.6786	27	-	-	-	-	-	-

Fonte: Elaborado pelo autor (2023).

3.2 Variabilidade metabólica das matrizes complexas de sementes de cacau após 84 horas de fermentação

Verificou-se que a quantidade de sinais presentes em cada um dos extratos sofreu mudanças expressivas, o que pode ser associado à intensa atividade da cocobiota sobre os substratos disponíveis (

Figura 33-A). Na marca de 84 horas de fermentação, os experimentos que anteriormente mostraram as maiores quantidades de sinais tiveram uma mudança expressiva. O experimento 2, realizado exclusivamente com acetato de etila (B), que anteriormente apresentava alta variabilidade, agora exibiu a menor quantidade de sinais (115). Da mesma forma, o experimento 1, que foi conduzido apenas com hexano (A), e que na marca de 0 hora tinha o maior número de sinais, agora está entre os experimentos com as menores variabilidades (150). O experimento 3, realizado exclusivamente com etanol (C), também mostrou uma quantidade menor de sinais (126).

O modelo proposto para prever a variabilidade metabólica após 84 horas de fermentação revelou algumas características interessantes quando comparado com o ponto 0 hora. Embora nenhuma interação cúbica modulada tenha demonstrado significância estatística para o modelo, esses coeficientes foram mantidos no modelo final Figura 33-B e 33-C. A razão para essa decisão reside no fato de que, mesmo que não sejam estatisticamente significativos, eles contribuíram para a obtenção do coeficiente de determinação (R^2), que atingiu o valor de 0,82.

O coeficiente de determinação (R^2) de 0,82 indicou que aproximadamente 82% da variabilidade metabólica observada após 84 horas de fermentação pode ser explicada pelo modelo proposto. Isso sugeriu que o modelo tivesse capacidade de prever a quantidade de sinais metabólicos com base nas variáveis consideradas, embora as interações cúbicas moduladas não tenham contribuído significativamente para essa previsão.

Os coeficientes utilizados na criação do modelo preditivo revelaram informações importantes sobre as interações entre os solventes e a variabilidade metabólica nos extratos de cacau após 84 horas de fermentação. Os 3 coeficientes

mais expressivos do modelo estão associados aos efeitos dos coeficientes primários, com destaque para coeficiente A (hexano), que se sobressaiu à média dos demais efeitos por uma diferença de 1,64 desvios-padrão nas respostas do modelo. Entre os três coeficientes modulados AB(A-B) (hexano.acetato de etila.[hexano-acetato de etila]), AC(A-C) (hexano.etanol.[hexano-etanol]) e BC(B-C) (acetato de etila.etanol[acetato de etila-etanol]), que envolveram a multiplicação pela diferença dos efeitos, nenhum deles mostrou-se estatisticamente significativo.

No entanto, o coeficiente terciário ABC (hexano.acetato de etila.etanol) se destacou, sendo o único responsável por conferir a característica cúbica ao modelo, cujo valor padronizado é de -2,50. Ao relacioná-lo com o valor de z, considerando as faixas de uma distribuição normal, observou-se que esse coeficiente está negativamente afastado da média dos demais por 1,64 desvios-padrão. Esse efeito antagônico indicou que a presença de interações mais intensas entre os três solventes está associada a uma menor variabilidade nos extratos brutos das sementes fermentadas de cacau. Esses resultados indicaram que a presença de interações complexas entre os solventes hexano, acetato de etila e etanol tiveram impactos negativos na variabilidade metabólica dos extratos de cacau durante o processo de fermentação. Quanto mais intensas forem essas interações, menor será a diversidade de compostos presentes nos extratos (Figura 33-D).

A análise da curva de contorno na Figura 33-E revelou uma distribuição abrangente da variabilidade metabólica dentro do espaço definido pelo SLD 3x3. Embora exista uma ampla distribuição dos sinais químicos ao longo da curva de contorno, foi possível identificar uma região com maior variabilidade metabólica. Neste ponto experimental, o experimento 7 se destacou como a região com a maior variabilidade química. Esse experimento foi uma composição de 66,6% de hexano e 33,4% de acetato de etila. A presença desses solventes em proporções específicas resultou em uma ampla diversidade de sinais metabólicos, indicando uma maior complexidade química.

Essa observação ressalta a importância da composição dos solventes utilizados na extração metabólica e como ela influenciou a variabilidade química dos extratos de sementes fermentadas de cacau. O experimento 7, com sua

combinação específica de solventes demonstrou ser uma opção promissora para explorar a riqueza metabólica do cacau, e direcioná-la para outros fins, como produção de bioativos para indústrias cosméticas, alimentícias e farmacêuticas pode ser uma alternativa.

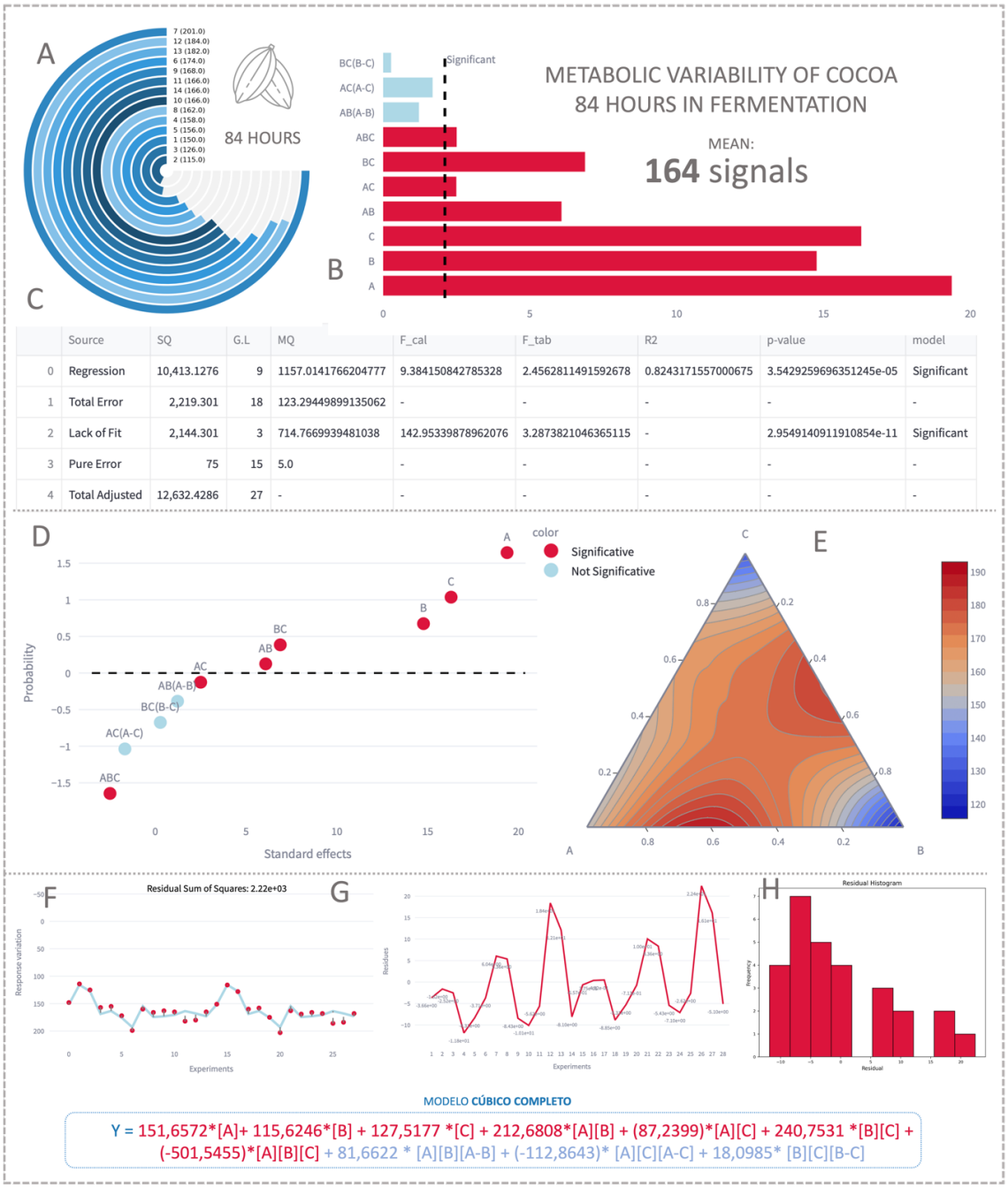
Além disso, é importante considerar que a variabilidade metabólica não se limita apenas a essa região específica, pois a curva de contorno indicou que havia uma distribuição abrangente de sinais metabólicos em todo o espaço do plano experimental. Essa abrangência da variabilidade sugeriu a presença de diferentes compostos químicos e interações complexas, fornecendo uma visão mais completa da diversidade metabólica da fermentação das sementes de cacau.

Embora o coeficiente ABC tenha se mostrado o único terciário significativo, outros coeficientes poderiam desempenhar papéis importantes nas interações entre os solventes e na variabilidade metabólica dos extratos das sementes fermentadas de cacau. Essa constatação foi especialmente relevante quando observaram-se os resíduos gerados pela previsão do modelo (Figura 32-F). Ao analisar os resíduos, foi possível observar que eles estão mais distribuídos na parte negativa do gráfico, conforme ilustrado na Figura 33-G. Essa distribuição assimétrica indica uma falta de homogeneidade na dispersão dos resíduos, sugerindo que o modelo não conseguiu capturar completamente a variação dos dados experimentais.

Ademais, a falta de ajuste foi evidenciada pela representação dos resíduos no histograma, conforme apresentado na Figura 33-H. Nesse histograma, notou-se que os dados não seguiram uma distribuição gaussiana, havendo uma descontinuidade no centro do gráfico. Essa discrepância na distribuição dos resíduos reforçou a inadequação do modelo em descrever adequadamente a variabilidade metabólica dos extratos de sementes de cacau fermentadas.

A falta de ajuste do modelo ressaltou a complexidade do processo de fermentação do cacau e a necessidade de explorar outras abordagens ou variáveis para melhorar a previsão da variabilidade metabólica desses extratos. Esses resultados ainda destacaram a complexidade do processo de fermentação do cacau e a necessidade de considerar múltiplos fatores na busca por extratos com alta variabilidade metabólica.

Figura 33 - Análises estatísticas no ponto 84 horas de fermentação. A) Quantidade de sinal em cada experimento. B) Gráfico de Pareto. C) ANOVA. D) Gráfico de dispersão coef. padronizados. E) Gráfico de Contorno. F) Previsão do modelo. G) Distribuição residual. H) Histograma residual.



	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	10,413.1276	9	1157.0141766204777	9.384150842785328	2.4562811491592678	0.8243171557000675	3.5429259696351245e-05	Significant
1	Total Error	2,219.301	18	123.29449899135062	-	-	-	-	-
2	Lack of Fit	2,144.301	3	714.7669939481038	142.95339878962076	3.2873821046365115	-	2.9549140911910854e-11	Significant
3	Pure Error	75	15	5.0	-	-	-	-	-
4	Total Adjusted	12,632.4286	27	-	-	-	-	-	-

Fonte: Elaborado pelo autor (2023).

3.3 Variabilidade metabólica das matrizes complexas de sementes de cacau após 168 horas de fermentação.

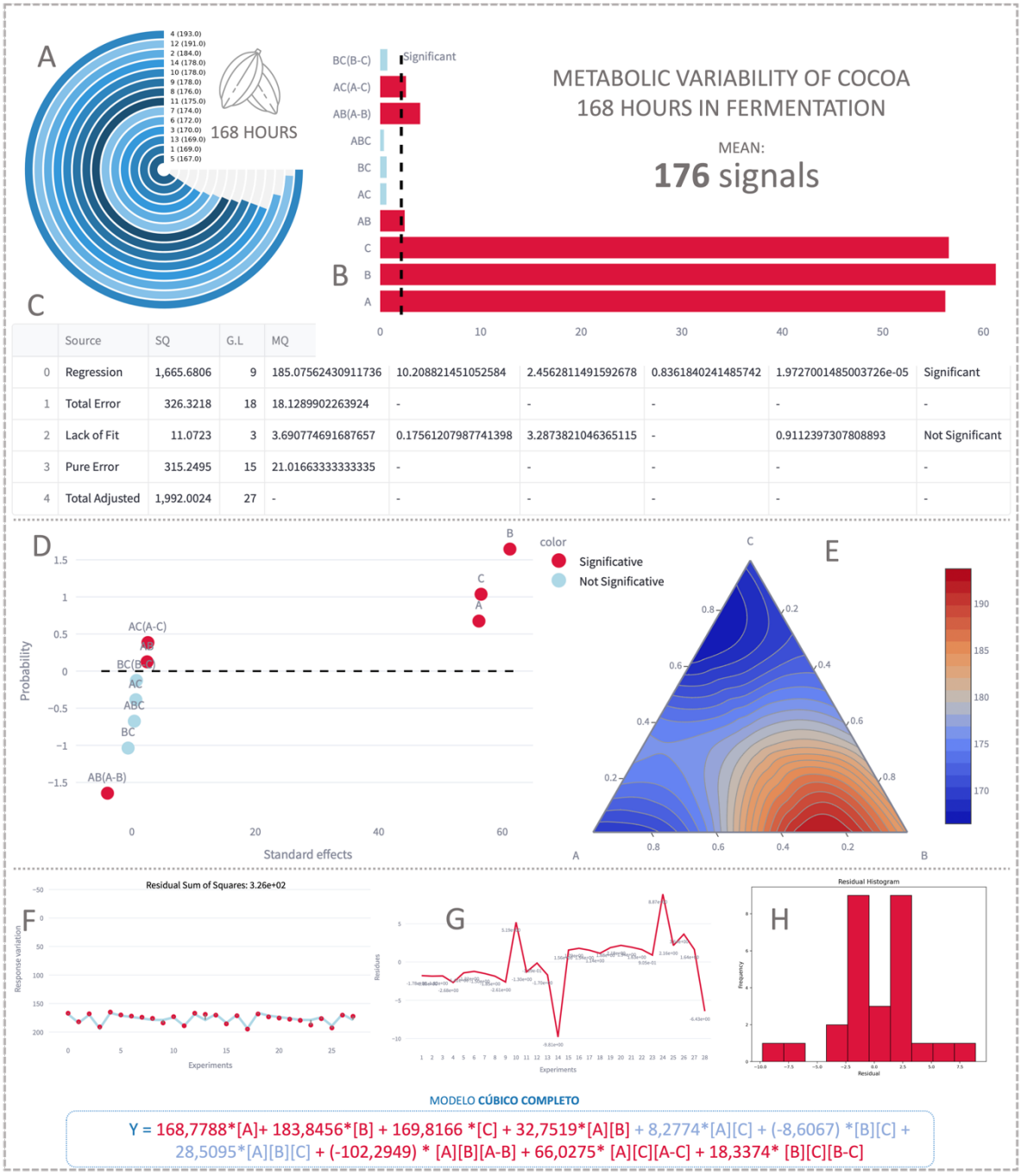
A análise estatística realizada para delinear o processo final da fermentação das sementes de cacau revelou uma tendência consistente entre os extratos de todas os estágios explorados. Conforme demonstrado na

Figura 34-A, as composições de misturas com a maior quantidade de sinais estão associadas ao experimento 4 (193), composto por 33,4% de hexano (A) e 66,6% de acetato de etila (B), seguido pelo experimento 12 (191), conduzido com 16,6% de hexano (A), 66,6% de acetato de etila (B) e 16,6% de etanol (C), e o experimento 2 (184), realizado com 100% de acetato de etila.

O melhor modelo estatístico desenvolvido para prever a variabilidade metabólica na fase final da fermentação apresentou um coeficiente de determinação de 0,83, determinado pelo modelo cúbico completo, com 6 coeficientes significativos Figura 34-B e 33-C). O coeficiente mais influente nas respostas foi o B (acetato de etila), que se destacou positivamente em relação aos demais, com uma diferença de 1,64 desvios-padrão quando relacionado ao valor de z distribuído em faixas gaussianas. Por outro lado, o coeficiente AB(A-B) (hexano.acetato de etila.[hexano-acetato de etila]) apresentou um efeito antagônico ao modelo, indicando que quanto maior a diferença entre A e B, menor é a variabilidade metabólica nos extratos (Figura 34-D).

Apesar de não explicar os 17% restantes da complexa variabilidade metabólica associada ao ponto final da fermentação, o modelo ajustou-se bem às respostas, apresentando uma distribuição residual adequada e um histograma com tendência a seguir uma distribuição gaussiana, caso seja considerado o agrupamento das faixas centrais. Ao analisar as informações de forma mais precisa, foi possível obter uma distribuição gaussiana dos resíduos melhor do que os modelos anteriores criados em outros pontos (Figura 34-F, 33-G e 33-H).

Figura 34 - Análises estatísticas no ponto de 168 horas de fermentação. A) Quantidade de sinal em cada ensaio. B) Gráfico de Pareto. C) ANOVA. D) Gráfico de dispersão coef. padronizados. E) Gráfico de Contorno. F) Previsão do modelo. G) Distribuição residual. H) Histograma residual.



	Source	SQ	G.L	MQ					
0	Regression	1,665.6806	9	185.07562430911736	10.208821451052584	2.4562811491592678	0.8361840241485742	1.9727001485003726e-05	Significant
1	Total Error	326.3218	18	18.1289902263924	-	-	-	-	-
2	Lack of Fit	11.0723	3	3.690774691687657	0.17561207987741398	3.2873821046365115	-	0.9112397307808893	Not Significant
3	Pure Error	315.2495	15	21.01663333333335	-	-	-	-	-
4	Total Adjusted	1,992.0024	27	-	-	-	-	-	-

Fonte: Elaborado pelo autor (2023).

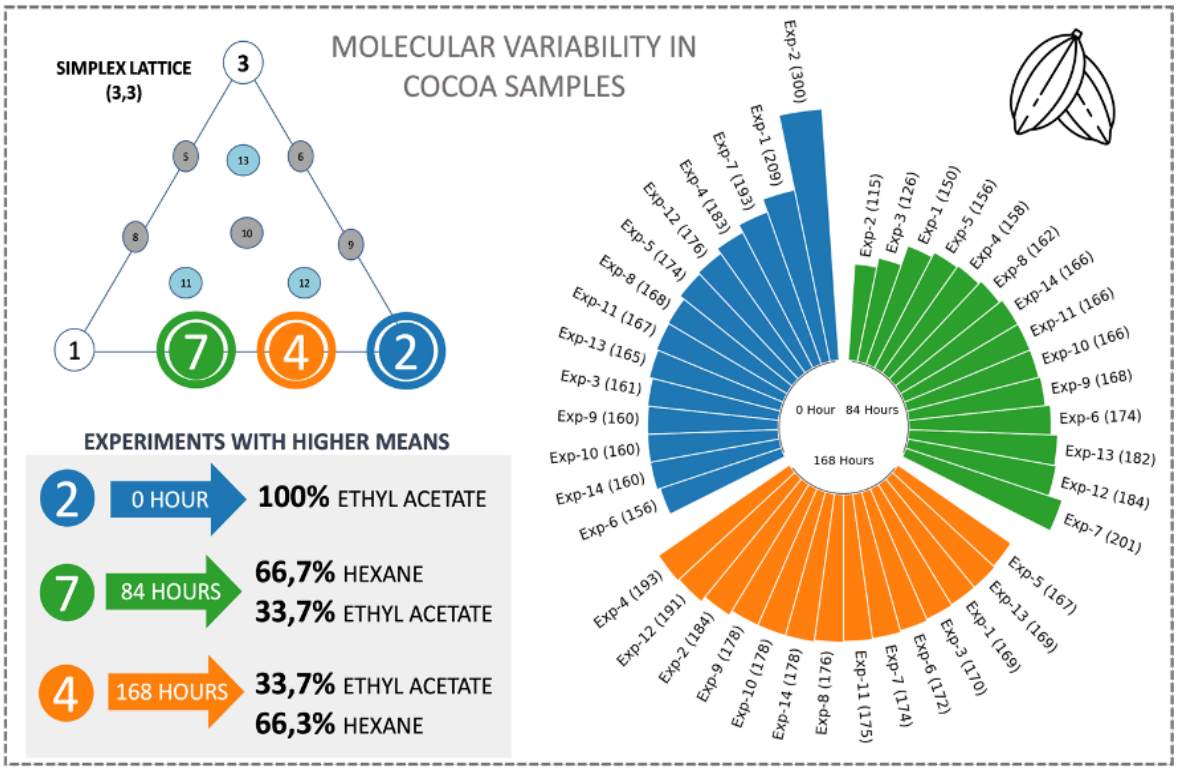
A combinação de todos os coeficientes resultou em um modelo estatístico ajustado, evidenciado por um gráfico de contorno com uma área intensa na parte centro-direita do espaço Simplex-Lattice (Figura 34-E).

Essa região indica a localização mais promissora para obter extratos com a maior quantidade de sinais, tornando-os ideais para estudos direcionados à desreplicação molecular. Essa abordagem permitiria uma maior possibilidade de exploração computacional dos dados, ampliando o espaço químico e aumentando o potencial de identificação de substâncias únicas. A presença de uma maior diversidade de sinais nesses extratos aumenta a probabilidade de encontrar moléculas com propriedades bioativas distintas, abrindo novas oportunidades para o desenvolvimento de produtos farmacêuticos, cosméticos ou agroquímicos.

3.4 Variabilidade metabólica no experimento 2 ao longo do processo fermentativo das sementes de cacau

A variabilidade metabólica nos três estágios do processo fermentativo das sementes de cacau pode ser representada e sintetizada na Figura 35.

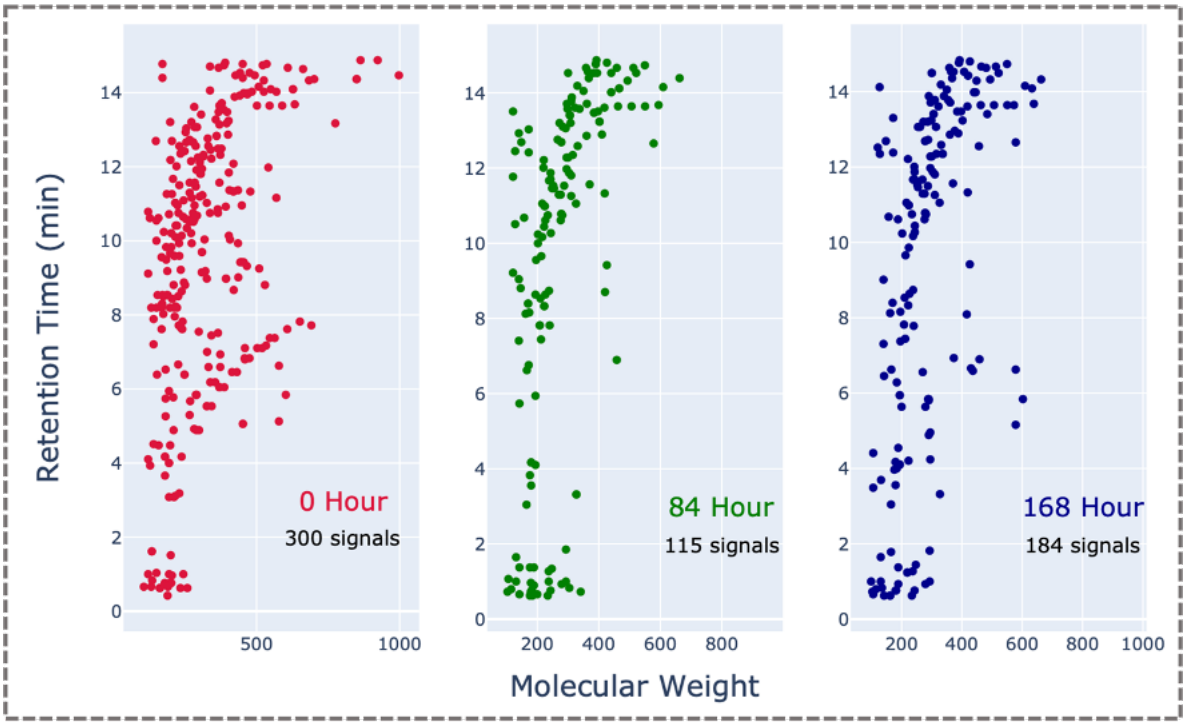
Figura 35 - Variabilidade metabólica fermentação de sementes cacau, utilizando a média da quantidade de sinais dos extratos obtidos a partir do SLD 3x3.



Fonte: Elaborado pelo autor

O experimento 2 se destacou como um ensaio notável e exemplar para discussão devido à sua variação dinâmica no número de sinais ao longo do processo de fermentação. No início da etapa fermentativa, ele registrou a maior quantidade de sinais, indicando uma alta variabilidade metabólica. No entanto, após 84 horas de fermentação, houve redução notável nos sinais, sugerindo uma mudança na composição metabólica. Porém, no final do processo, o número de sinais aumentou novamente, destacando a complexidade e a dinâmica das mudanças metabólicas que ocorrem durante a fermentação das sementes de cacau. Esta flutuação no número de sinais ao longo do processo de fermentação faz do experimento 2 um excelente exemplo para aprofundar a discussão sobre a variabilidade metabólica associada à fermentação do cacau (Figura 36).

Figura 36 - Variabilidade metabólica no experimento 2 ao longo do processo fermentativo das sementes de cacau.



Fonte: Elaborado pelo autor (2023).

Notou-se que o extrato desse experimento, obtido a partir de sementes de cacau não fermentadas e extraído com 100% de acetato de etila, apresentou uma maior diversidade de sinais, inclusive com a presença de moléculas de alto peso molecular (acima de 800 Dalton), algo não observado nos extratos obtidos em 84 e 168 horas de fermentação. Essas moléculas com altas massas moleculares poderiam ter servido como substrato para a sucessão microbiana durante a fermentação, sendo convertidas em outras estruturas moleculares.

Além disso, é interessante destacar que não foi observada uma tendência de diminuição dos sinais ao longo do processo de fermentação. No ponto de 84 horas, houve uma redução de aproximadamente 62% na quantidade de sinais, porém, quando comparado ao final do processo fermentativo, a quantidade de sinais volta a aumentar, mesmo apresentando 184 sinais moleculares, 39% abaixo do que foi evidenciado no extrato das sementes de cacau não fermentadas.

Tais resultados indicam que a fermentação das sementes de cacau é um processo dinâmico, com alterações significativas na composição metabólica ao longo do tempo. A presença de moléculas de massa molecular mais elevada no início da fermentação sugere uma metabolização microbiana intensa nessa fase inicial, com posterior transformação e diversificação dos compostos ao longo do tempo.

A presença de moléculas de alto peso molecular no início da fermentação pode ser devido à disponibilidade inicial de compostos poliméricos, ou ainda a presença de açúcares e compostos bioquímicos nas sementes de cacau frescas, que servem como substrato para a cocobiota fermentativa. Conforme a fermentação progrediu, compostos de alto peso molecular poderiam ter sido metabolizados pelos microrganismos em estruturas mais simples, levando a uma redução temporária na diversidade de sinais detectados. A subsequente recuperação na diversidade de sinais no final do processo fermentativo poderia ser atribuída à biossíntese de novos compostos pelos micro-organismos presentes, indicando uma transformação e reorganização do perfil metabólico durante o processo de fermentação.

Além disso, a redução inicial e o posterior aumento na diversidade de sinais pode também ser um reflexo das mudanças na composição da cocobiota ao longo do processo fermentativo. Diferentes espécies de micro-organismos, com diferentes capacidades metabólicas, podem dominar em diferentes estágios da fermentação, levando a variações no perfil metabólico dos extratos. A diversidade metabólica inicial pode ser atribuída à presença de uma comunidade microbiana diversificada nas fases iniciais, que é posteriormente substituída por espécies mais especializadas à medida que o processo fermentativo avança.

Adicionalmente, o ambiente de fermentação estaria se tornando mais estressante ao longo do tempo, levando a uma produção aumentada de metabólitos secundários por parte dos microrganismos presentes, em resposta ao estresse. Esses metabólitos secundários seriam responsáveis pelo aumento na diversidade de sinais observados no final do processo fermentativo.

4 CONCLUSÃO

Durante a fermentação das sementes de cacau, diferentes micro-organismos desempenham papéis específicos na remoção da polpa e na produção de metabólitos essenciais. As leveduras realizam a decomposição da pectina e a fermentação anaeróbica dos açúcares, enquanto as bactérias láticas convertem os açúcares e o ácido cítrico em ácido lático, ácido acético e manitol em condições microaerofílicas. Por sua vez, as bactérias acéticas promovem a bioconversão do etanol em ácido acético em condições aeróbicas. Notou-se que pela exploração das matrizes complexas oriundas da fermentação das sementes de cacau, esses processos desempenham um papel fundamental na formação na variabilidade metabólica dos extratos, sobretudo, quando preparados a partir de um delineamento experimental.

Esses resultados destacam a importância deste tipo de abordagem experimental, sobretudo, do tipo Simplex-Lattice e da utilização de ferramentas computacionais como o Chemistika para a exploração do perfil químico em estudos de misturas. As análises estatísticas contribuíram para a compreensão do processo de fermentação das sementes do fruto e podem auxiliar na otimização da produção de extratos com perfis metabólicos desejados.

Para mais, a abordagem adotada neste estudo proporcionou uma compreensão mais aprofundada das propriedades químicas dos extratos, fornecendo insights valiosos para o estudo da fermentação do cacau, abrindo caminhos para a investigação de suas propriedades moleculares aplicadas a outras áreas além do chocolate.

As ferramentas estatísticas utilizadas permitiram identificar as regiões mais promissoras no espaço Simplex-Lattice por meio dos modelos estatísticos abrangendo três fases distintas da fermentação das sementes de cacau. Os extratos obtidos das sementes *in natura* apresentaram, em média, a maior quantidade de sinais moleculares (187), destacando-se a extração com 100% de acetato de etila, que produziu a maior variabilidade metabólica (300 picos).

Em 84 horas de processo fermentativo, as matrizes complexas apresentaram variações no perfil químico, resultando em uma redução média de 12,3% na quantidade de sinais (164 sinais moleculares). Destacou-se a composição de misturas de 66,66% de hexano e 33,34% de acetato de etila, que resultou em 201 picos, representando a melhor composição extratora durante esta etapa do processo fermentativo, com 33% menos sinais do que a melhor composição do primeiro estágio da fermentação, que foi de 300 sinais.

Em 168 horas de fermentação, os sinais apresentaram leve aumento médio, totalizando 176 sinais, com destaque para a composição de mistura de 33,33% de hexano e 66,67% de acetato de etila, que permitiu a extração de 193 sinais moleculares. Porém, mesmo que, em média, o final do processo fermentativo tenha contemplado maior variabilidade quando comparado com a etapa de 84 horas, a melhor composição de misturas no final do processo fermentativo apresentou menos sinais, com uma redução de 12,4% para a melhor mistura extratora (sinalizando 193 picos).

Ademais, é possível que a fermentação das sementes de cacau seja um processo sequencial, onde diferentes grupos de compostos são metabolizados em diferentes estágios do processo, levando a variações temporais na diversidade de sinais observados. Essa hipótese estaria em consonância com a natureza dinâmica e complexa da fermentação do cacau, que envolve a interação de uma grande variedade de fatores, incluindo a composição inicial das sementes, a atividade da cocobiota, as condições ambientais e o tempo de fermentação.

Esses resultados forneceram uma orientação valiosa para direcionar investigações futuras e maximizar a eficiência na descoberta de compostos bioativos nessa fonte de matéria-prima. Além disso, mostraram que o complexo processo fermentativo causou alterações significativas e interessantes na composição química dos extratos, indicando composições mais assertivas para explorar matrizes mais ricas em sinais. Isso é especialmente relevante para usuários que buscam explorar matrizes complexas desse fruto utilizando ferramentas de desreplicação ou que têm interesse em identificar, isolar e elucidar compostos inéditos dessa rica matriz.

CAPÍTULO 6 – DINÂMICA MOLECULAR DAS ANOTAÇÕES PRESENTES NAS MISTURAS COMPLEXAS DO PROCESSO FERMENTATIVO DO CACAU

RESUMO

Este capítulo contempla a análise das trajetórias, decorrentes da abordagem de dinâmica molecular de 10 anotações moleculares que demonstraram potencial para atuarem em quatro bioreceptores associados a doenças respiratórias, tais como SARS-CoV-2 e asma. Essas anotações foram identificadas pelo software LUMIOS (Label Using Machine In Organic Samples), o qual automatizou as respostas obtidas por meio de algoritmos de inteligência artificial e testes de docagem molecular. Utilizou-se o algoritmo GROMACS para explorar tais anotações e analisar as interações finais entre os ligantes (anotações) e as proteínas, realizando cálculos energéticos com base nas forças intermoleculares. Durante os estudos de docagem molecular, as moléculas de trealose e catequina, procianidina, ácido ftálico, adenina, indol-3-acetamida, teobromina e anidrido ftálico, foram identificadas como estruturas promissoras para modular os alvos biomacromoleculares propostos. Visando estudar o comportamento das trajetórias de tais anotações nos sítios ativos das proteínas, realizou-se um estudo guiado por dinâmica molecular, a qual conduziu cálculos de energia ao longo de um período de 100 nanossegundos, permitindo ampliar a compreensão de como tais metabólitos interagem com as proteínas-alvo associadas às doenças respiratórias. Tais análises permitiram avaliar as forças intermoleculares envolvidas nessas interações, fornecendo resultados valiosos sobre os mecanismos de ligação e estabilidade dos complexos formados, conhecendo e explorando possíveis sítios de ligação, além de determinar a eficácia das anotações moleculares e fornecer subsídios iniciais para o desenvolvimento de novas terapias direcionadas ao tratamento de doenças associadas ao trato respiratório. Dentre as anotações moleculares incorporadas nestas análises, destaca-se a efetividade da molécula de catequina, modulando três dos quatro alvos receptores (associados à asmas e SARS-CoV-2), além da trealose

e ácido ftálico, os quais apresentaram afinidade por alvos de doenças associadas à asma.

Palavras-chave: *theobroma cacao*; dinâmica molecular; produtos naturais; doenças respiratórias

1 INTRODUÇÃO

O cacau, originário das florestas tropicais da América Central e do Sul (IGAWA; DE TOLEDO; ANJOS, 2022), é conhecido mundialmente não apenas como a base do chocolate, mas também como um reservatório de compostos bioativos com potenciais benefícios à saúde (COOPER et al., 2008). Por séculos, as civilizações pré-colombianas já reconheciam e aproveitavam as propriedades medicinais deste fruto (COQ-HUELVA; TORRES-NAVARRETE; BUENO-SUÁREZ, 2018; MOTAMAYOR et al., 2002), mas foi somente nas últimas décadas que a ciência moderna começou a descobrir e explorar a complexidade e potencialidade das moléculas presentes no cacau, valorizando e melhorando seu cultivo (SCHROTH; HARVEY, 2007).

Rico em flavonoides, teobromina e uma série de outros compostos, o cacau tem sido associado à promoção de saúde cardiovascular (ZIĘBA; MAKAREWICZ-WUJEC; KOZŁOWSKA-WOJCIECHOWSKA, 2019), melhorias na função cerebral, proteção contra o estresse oxidativo (NABAVI et al., 2015) e potencial anti-inflamatório (FREDHOLM; SMIT, 2011; KHAN et al., 2014b). Estudos têm indicado que o consumo regular e moderado de produtos oriundos do cacau pode ajudar na regulação da pressão arterial, melhoria da sensibilidade à insulina (GRASSI et al., 2005) e até mesmo na promoção do bem-estar e melhoria do humor (FUSAR-POLI et al., 2022).

Além de proporcionar benefícios à saúde e ao bem-estar, os produtos derivados do cacau, cujas propriedades são grandemente influenciadas pelo processo fermentativo, carregam em si um universo de complexidade biológica e química. A fermentação do cacau é um processo biológico essencial para o desenvolvimento do sabor e aroma característicos do chocolate. Este processo depende inteiramente da ação de micro-organismos presentes no ambiente e no próprio fruto do cacau. Durante a fermentação, esses micro-organismos trabalham em conjunto, degradando e transformando compostos, o que influencia diretamente a qualidade do produto (SCHWAN; WHEALS, 2004). Inicialmente, as sementes de cacau são envolvidas por uma polpa açucarada, rica em sacarose. Os micro-

organismos, ao entrar em contato com esta polpa, começam a consumir esses açúcares. As leveduras, como *Saccharomyces cerevisiae*, são os primeiros micro-organismos a agir, convertendo a sacarose em álcool, principalmente etanol, e dióxido de carbono (AFOAKWA et al., 2013; KONGOR et al., 2016b).

Conforme a fermentação avança, bactérias lácticas e acéticas começam a se destacar. Estas bactérias, incluindo *Lactobacillus* e *Acetobacter*, consomem o etanol produzido pelas leveduras, transformando-o em ácido láctico e ácido acético (MOREIRA et al., 2013). Estes ácidos desempenham um papel crucial na diminuição do pH do ambiente, o que leva à morte das sementes de cacau e inicia o processo de germinação.

Neste estudo, buscou-se explorar as propriedades moleculares do cacau que transcendem seu uso tradicional na produção de chocolate. Em particular, esta investigação focou em potenciais biomoléculas derivadas do cacau que poderiam despertar interesse no campo farmacológico (BALENTIC et al., 2018; TSLJCHIE, 1991). Motivados pela recente pandemia de coronavírus, investigou-se a capacidade de tais moléculas de interagirem com alvos associados a doenças respiratórias, como asma e SARS-CoV-2 (COVID-19) por meio de ferramentas computacionais de alta performance, envolvendo docagem e dinâmica moleculares, técnicas capazes de revelar insights valiosos na modulação entre ligantes e proteínas associadas a determinadas doenças (ŚLEDŹ; CAFLISCH, 2018).

No campo da química, as ferramentas *in silico* revolucionaram o modo como se conduz a pesquisa e o desenvolvimento (DZOBO, 2022), oferecendo uma alternativa mais eficiente e frequentemente mais econômica às abordagens experimentais tradicionais (MEDEMA; FISCHBACH, 2015; ROMANO; TATONETTI, 2019). Elas permitem que os pesquisadores explorem, simulem e prevejam propriedades moleculares e interações, economizando tempo e recursos ao reduzir a necessidade de ensaios físicos (JORGENSEN, 2004).

Dentre as inúmeras técnicas do estado da arte, englobando tecnologias de Big Data para quimioinformática (GAUDÊNCIO et al., 2023), inteligência artificial baseada em reconhecimento de padrões utilizando aprendizado de máquina como redes neurais artificiais (SHI et al., 2023) (utilizadas nos capítulos anteriores), nesta

etapa se faz presente a relevância e a pertinência da docagem e dinâmica molecular, usadas para entender e prever o comportamento de moléculas e suas interações.

No entanto, tais técnicas se diferenciam no modo como abordam e no que focam. A docagem molecular busca prever a posição e orientação ótima de uma molécula quando se liga a uma macromolécula alvo (LEACH; SHOICHET; PEISHOFF, 2006). A dinâmica molecular, por outro lado, simula o movimento natural de moléculas ao longo do tempo (KUMAR et al., 2020). Apesar de suas diferenças, ambas são ferramentas complementares na química computacional, e neste trabalho, utilizou-se a docagem molecular como um filtro para seleção de estruturas que pudessem se efetivar no sítio ativo da proteína.

A abordagem de dinâmica molecular foi escolhida por conta de uma série de vantagens, como de fornecer uma visão temporal das interações, considerando a flexibilidade inerente das moléculas; simular condições mais próximas do ambiente biológico real e fornece informações detalhadas em nível atômico (HOLLINGSWORTH; DROR, 2018). Reforça-se, neste capítulo, que a abordagem computacional tem desempenhado um papel cada vez mais significativo no avanço da química de produtos naturais, especialmente quando se trata de investigar interações moleculares e prever o comportamento de compostos em sistemas biológicos complexos (NOVAK et al., 2021). Especificamente, as técnicas de docagem e dinâmica molecular oferecem uma janela única para sondar e entender a relação entre moléculas e alvos potenciais, que aqui foram exclusivamente pautados em receptores associados a doenças respiratórias.

No contexto do cacau, a fermentação é um processo crítico que pode levar à formação de uma variedade de compostos bioativos (MACHONIS et al., 2012). A investigação de moléculas derivadas de extratos obtidos a partir da fermentação do cacau, através de ferramentas computacionais, pode revelar potenciais aplicações terapêuticas e benefícios à saúde ainda não explorados (YAÑEZ et al., 2021b). Esta pesquisa, portanto, se concentra em empregar metodologias *in silico* para explorar, de maneira racional, as propriedades e potenciais aplicações de tais compostos oriundos do cacau fermentado.

2 METODOLOGIA

2.1 Exploração das matrizes complexas de cacau com o software LUMIOS:

Os dados de espectrometria de massas, derivados das matrizes de cacau, foram submetidos a um processamento rigoroso e direcionados ao sistema de manipulação de dados multitarefas da plataforma Label Using Machine In Organic Samples – LUMIOS. Esta plataforma, cujos detalhes operacionais foram extensivamente explorados no Capítulo 1 e 2, foi fundamental para a triagem inicial.

O principal objetivo desta triagem foi identificar as anotações moleculares que possuíam afinidade pelos quatro alvos biomoleculares específicos. Estes alvos foram meticulosamente inseridos no núcleo de docagem da plataforma LUMIOS, uma operação que facilitou a modulação e acesso ao sítio ativo das proteínas, região onde ocorrem as interações químicas.

Na fase de implementação da dinâmica molecular, adotou-se uma abordagem seletiva, concentrando-se em 10 anotações moleculares específicas extraídas das matrizes de cacau (catequina, trealose, procianidina, teobromina, adenina, indol-3-acetamida, ácido ftálico, anidrido ftálico, fenilalanina e tirosina). Estas anotações se destacaram durante a triagem inicial devido à sua afinidade significativa para com os alvos moleculares, permitindo a identificação das melhores poses de cada anotação, isto é, as orientações e disposições que possibilitavam o acesso e a modulação efetiva do sítio ativo proteico.

De posse das anotações selecionadas, avançou-se para a construção de um estudo de trajetória molecular. Utilizou-se o algoritmo GROMACS, uma ferramenta poderosa que permite análises detalhadas das interações e movimentos moleculares dentro do sítio ativo (VAN DER SPOEL et al., 2005). Esta abordagem de docagem e dinâmica molecular possibilitou sondar profundamente os processos complexos que governam as interações entre as anotações moleculares e os alvos biomoleculares.

Esta seleção consciente de 10 anotações moleculares garantiu que o estudo de dinâmica molecular fosse robusto e informativo, propiciando uma compreensão

aprofundada das interações que ocorrem nos sítios ativos das proteínas e dos processos subjacentes que regem essas relações moleculares.

2.2 Dinâmica molecular

A dinâmica molecular é uma ferramenta essencial para refinar os resultados obtidos por meio de técnicas de docagem, proporcionando uma visão mais aprofundada das interações biomoleculares. Para este propósito, é importante mencionar que utilizou-se a plataforma LUMIOS (VIEIRA et al., 2023), exclusivamente para desreplicação e a etapa de docagem molecular (a plataforma não efetuada a dinâmica molecular em si). As etapas subsequentes, incluindo a dinâmica molecular, são realizadas com a ajuda do algoritmo GROMACS. Neste estudo, utilizou-se a arquitetura CUDA (Compute Unified Device Architecture) tecnologia disponibilizada por placas de vídeo fabricadas pela empresa NVIDIA, para acelerar as simulações de dinâmica molecular, operando em um sistema equipado com um processador AMD Ryzen 7, 80 GB de memória RAM e uma placa de vídeo NVIDIA RTX 3060. Uma representação esquemática da metodologia pode ser visualizada na Figura 37.

Figura 37 - Resumo de proposta metodológica para condução de estudos voltados à Dinâmica Molecular.

Fonte: Elaborado pelo autor (2023).

O complexo formado entre o ligante e a proteína foi analisado utilizando o software de dinâmica molecular de código aberto, GROMACS, instalado no sistema operacional LINUX (Ubuntu 22.04.03). Inicialmente, os arquivos do ligante e da proteína foram preparados, processo que incluiu a otimização das estruturas, adição de íons e solventes apropriados, e a definição da posição do ligante no receptor, baseada nos resultados da docagem obtidos no LUMIOS (que disponibiliza o complexo a partir da melhor pose do ligante).

O campo de força CHARMM, atualizado em julho de 2022, foi selecionado para descrever as interações intermoleculares no sistema, garantindo assim a precisão e confiabilidade das trajetórias de simulação. Um campo de força é composto por equações matemáticas e parâmetros que descrevem as interações entre átomos, incluindo ligações covalentes (como ligações simples, duplas, etc.), ângulos entre ligações, interações de van der Waals e interações eletrostáticas. Essas equações são usadas para calcular as forças exercidas entre os átomos e, assim, prever como eles se moverão ao longo do tempo. Ao realizar simulações de

dinâmica molecular, o campo de força é usado para calcular as posições dos átomos em cada passo de tempo, permitindo a observação de como a conformação de uma molécula muda temporalmente e como ela interage com outras moléculas em seu ambiente.

Em seguida, foram aplicadas condições de contorno periódicas, usando o método Particle Mesh Ewald (PME) para interações eletrostáticas de longo alcance, e realizada a minimização de energia para remover tensões no sistema. Este passo foi seguido pela aplicação de restrições de posição e equilíbrio da temperatura e pressão do sistema, tirando proveito da arquitetura tecnológica CUDA para acelerar esses processos computacionalmente intensivos.

A simulação de dinâmica molecular foi então executada sem restrições por 100 nanosegundos, salvando as coordenadas estruturais a cada 10 picosegundos para análise subsequente. Esta etapa também foi beneficiada pela eficiência proporcionada pela arquitetura CUDA. Após a simulação, várias métricas de análise foram calculadas, incluindo Root Mean Square Deviation (RMSD), Root Mean Square Flutuation (RMSF), raio de giro (RG), número de ligações de hidrogênio e energia livre de Gibbs. A energia livre entre o receptor e o ligante foi analisada utilizando o método Mechanics/Poisson-Boltzmann Surface Area (MMPBSA) (KUMARI; KUMAR; LYNN, 2014), fornecendo uma visão detalhada das interações moleculares no sistema. A análise exploratória dos dados foi conduzida pela plataforma CHEIC – Chemical Image Classificator (VIEIRA et al., 2023), que abriga algoritmos para sistematização e automatização do tratamento dos dados oriundos do GROMACS, gerando visualizações e análises detalhadas dos resultados de dinâmica molecular.

A combinação da plataforma LUMIOS, CHEIC e GROMACS, apoiada pelo campo de força CHARMM atualizado e pela aceleração da arquitetura CUDA, possibilitou a identificação e análise detalhada das interações entre ligantes e proteínas, agregando potencial para o desenvolvimento de fármacos e de terapias relacionadas ao tratamento de doenças respiratórias.

3 RESULTADOS E DISCUSSÃO

Para uma análise eficiente e comparativa dos resultados derivados das trajetórias moleculares originadas da dinâmica molecular, decidiu-se estruturar as discussões em quatro grupos distintos. Esta organização foi baseada nos receptores propostos, agrupando-se as moléculas de acordo com o receptor a qual apresentaram alta afinidade e que foram projetadas nos estudos fornecidos pela dinâmica molecular.

- **Grupo 1NC6:** Neste conjunto, o foco é na interação da molécula de procianidina com o receptor 1NC6, proteína associada à asma. As informações obtidas para este complexo possibilitaram analisar detalhadamente as interações e as trajetórias moleculares visando entender como a procianidina liga-se e modula a atividade do receptor 1NC6.
- **Grupo 6VVU:** Este agrupamento contempla uma discussão mais complexa, considerando que envolve as interações de várias moléculas – adenina, catequina, indol-3-acetamida, fenilalanina, tirosina, teobromina, trealose e anidrido ftálico e ácido ftálico – com o receptor 6VVU. Neste caso, a proposta foi de explorar como cada uma dessas moléculas interage individualmente com o receptor e efetuar uma discussão sobre qual estrutura é mais adequada a ser uma candidata melhor que o ligante original associado a tal proteína.
- **Grupos 4DD8 e 7P2G:** Nestes conjuntos, a análise estará centrada na catequina e na trealose, as demais anotações não foram efetivas na modulação destes receptores. Ambas as moléculas foram estudadas em relação aos receptores 4DD8 e 7P2G, permitindo uma comparação direta das trajetórias moleculares e interações no sítio ativo desse receptor.

Ressalta-se que as comparações serão realizadas em pares. Nesse processo, cada anotação será avaliada em relação ao ligante originalmente co-cristalizado com a proteína, aqui denominado ligante-padrão. Ao estabelecer um limiar comparativo, será possível aplicar análises estatísticas a fim de obter insights

valiosos e identificar anotações derivadas do cacau que possam emergir como candidatos a fármacos.

As análises efetuadas e discutidas nos capítulos anteriores, somadas às contribuições da dinâmica molecular sugerem um aspecto particularmente interessante sobre o potencial dessas anotações de atuarem em bioreceptores associados a doenças respiratórias. Esta linha de investigação pode levar à identificação de novos compostos bioativos derivados do cacau que tenham potencial terapêutico, enriquecendo assim o repertório de tratamentos disponíveis para estas doenças.

3.1 Dinâmica Molecular – grupo receptor 1NC6

A única anotação molecular indicada inicialmente pela técnica de docagem como uma possível candidata a atuar na modulação da proteína 1NC6, associada à asma (COSTANZO et al., 2003b), foi a molécula de procianidina. A procianidina é um composto bioativo que tem sido objeto de vários estudos devido às suas propriedades antioxidantes e anti-inflamatórias. Pesquisas recentes sugerem que a procianidina pode ter efeitos benéficos em várias condições de saúde, incluindo doenças respiratórias. Por exemplo, um estudo conduzido por Zhou (ZHOU et al., 2015) descobriu que a procianidina pode aliviar a inflamação das vias aéreas e a hiperresponsividade em ratos, acometidos de asma alérgica, sugerindo seu potencial como uma nova estratégia terapêutica para a asma.

Outro estudo, conduzido por Jiang e colaboradores (JIANG et al., 2020), demonstrou que a procianidina pode suprimir a resposta inflamatória em células epiteliais brônquicas humanas, fornecendo um novo insight sobre os mecanismos subjacentes da inflamação das vias aéreas. Estes apontamentos corroboram com a sinalização de que a molécula de procianidina possa ser uma candidata a modular o alvo 1NC6, uma vez que ela apresenta afinidade pelo receptor, confirmada pelo processo de docagem molecular, maior que o ligante padrão.

A análise da dinâmica molecular aqui proposta visa analisar, inicialmente, a energia total associada ao complexo, que se traduz na afinidade entre o ligante e

proteína (HOU et al., 2011). A energia total é composta pelas contribuições de 6 fatores, sendo eles:

- **VDWAALS**: Refere-se à energia de Van der Waals. Esta é a energia associada às forças de atração ou repulsão entre moléculas que não são resultado de uma ligação covalente ou uma interação iônica. Essas forças podem surgir de dipolos induzidos ou flutuantes em moléculas (WANG; WANG; KOLLMAN, 1999).
- **EEL:** É uma sigla para Energia Eletrostática. Esta é a energia associada às forças entre partículas carregadas (AKSIMENTIEV; SCHULTEN, 2005). Em um sistema biológico, como uma proteína, esta seria a energia resultante das interações entre os aminoácidos carregados.
- **ESURF:** Esta sigla se refere à Energia de Superfície. Em contextos de modelagem molecular, é uma forma de estimar a energia necessária para manter a área de superfície de uma molécula ou proteína, que pode ser importante em interações como a adesão de proteínas (DU et al., 2016).
- **GGAS:** Energia de um sistema no estado gasoso. Em simulações de dinâmica molecular, refere-se à energia total de um sistema no estado gasoso, ou seja, na ausência de solvente. Este termo geralmente inclui as energias de van der Waals e eletrostáticas, que são as interações entre todos os átomos do receptor e do ligante que não são ligados covalentemente (GENHEDEN; RYDE, 2015).
- **GSOLV:** Refere-se à energia envolvida no processo de solvatação, que é a interação e atração de moléculas de solvente (como a água) com moléculas solúveis ou partículas (como íons ou uma proteína) (BARRIL et al., 2001) .
- **EGB:** é a energia envolvida no processo de solvatação, ou seja, a interação de um soluto com o solvente em uma solução. Este é um componente específico da energia de solvatação polar. É calculado usando o método Generalized Born (GB), que é uma aproximação da teoria do solvato contínuo (ONUFRIEV; CASE, 2019).

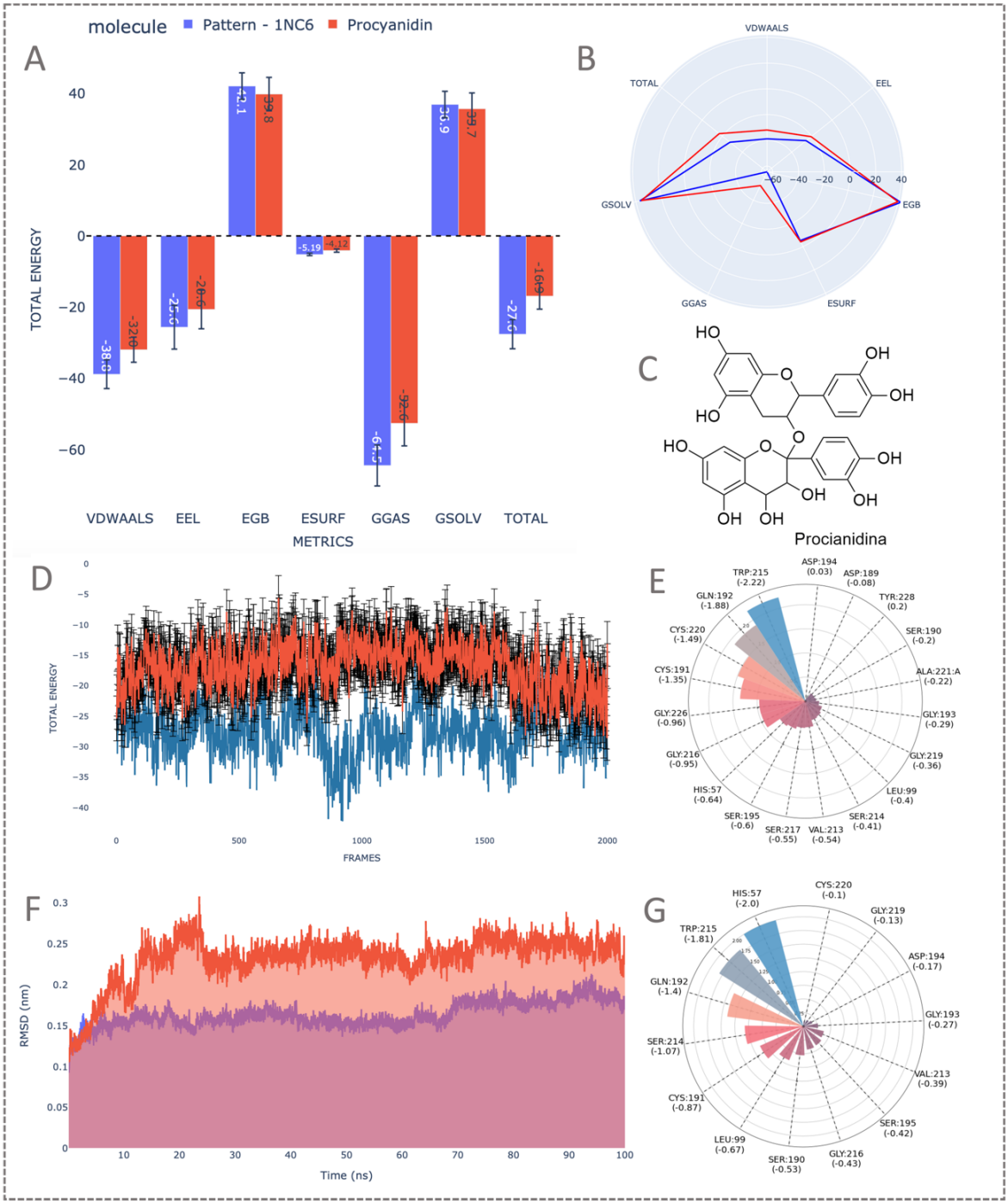
Em análise aos gráficos apresentados na Figura 38, nota-se que o ligante padrão, originalmente co-cristalizado, apresenta uma interação geral mais forte com a proteína, como indicado pelo valor médio mais negativo da energia total (-27,59 kcal/mol para o ligante-padrão versus -16,90 kcal/mol para a procianidina). Isso sugere que o ligante-padrão pode formar um complexo mais estável com a proteína do que a procianidina.

A força desta interação é influenciada por vários fatores. Por exemplo, a interação de van der Waals entre o ligante-padrão e a proteína é mais forte, como indicado pelo valor médio de VDWAALS mais negativo para o ligante-padrão (-38,83 kcal/mol) em comparação com a procianidina (-31,96 kcal/mol). Além disso, o ligante-padrão apresenta uma interação eletrostática mais forte com a proteína (EEL), como indicado pelo valor médio de EEL mais negativo (-25,63 kcal/mol para o ligante-padrão versus -20,64 kcal/mol para a procianidina).

Os valores de EGB e ESURF indicam que a solvatação da procianidina é ligeiramente mais favorável. No entanto, a diferença nesses valores é bastante pequena, e esses fatores parecem ser menos influentes no resultado do que as interações de van der Waals e as interações eletrostáticas.

Contudo, vale mencionar que estes são apenas valores médios, e a variação em torno de tais valores também é importante. Por exemplo, a interação de van der Waals do ligante-padrão com a proteína parece ter uma variabilidade maior do que a da procianidina (desvio-padrão de 4,04 kcal/mol para o ligante-padrão versus 3,72 kcal/mol para a procianidina). Isso pode ter implicações para a robustez e a reprodutibilidade dessas interações.

Figura 38 - Resultados de Dinâmica Molecular (procianidina e receptor 1NC6). A) Distribuição energética. B) Gráfico de Radar para energias. C) Estrutura Procian.. D) Energia total durante 100ns. E) Decomp. residual (padrão). F) RMSD e G) Decomp. residual (ligante).



Fonte: elaborado pelo autor (2023).

Ao examinar isoladamente a energia total, observa-se uma diferença expressiva entre o ligante-padrão e a procianidina. Mesmo quando considerado o desvio padrão máximo, as estimativas para esses dois ligantes não coincidem, sugerindo que o ligante-padrão possa ser mais adequado para modular o alvo 1NC6 do que a procianidina. Entretanto, para uma compreensão mais ampla e precisa do sistema, outras métricas também foram levadas em consideração. Estas incluem a densidade, temperatura, pressão, energia potencial, Área de Superfície Acessível ao Solvente (SASA), raio de giro molecular, Raiz da Flutuação Quadrática Média (RMSF) e o Raiz do Desvio Quadrático Médio (RMSD). Além disso, serão considerados aspectos específicos da interação proteína-ligante, que envolvem uma análise detalhada da decomposição residual dos aminoácidos responsáveis pela constituição do sítio ativo da proteína e dos ligantes.

Tabela 8 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 1NC6 e Procianidina.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
1NC6 vs Procian.	Densidade	0,95	Falha ao rejeitar H0	Sem diferença
1NC6 vs Procian.	Temperatura	0,36	Falha ao rejeitar H0	Sem diferença
1NC6 vs Procian.	Pressão	0,32	Falha ao rejeitar H0	Sem diferença
1NC6 vs Procian.	Potencial	0,09	Falha ao rejeitar H0	Sem diferença
1NC6 vs Procian.	SASA	0,65	Falha ao rejeitar H0	Sem diferença
1NC6 vs Procian.	Raio Giro	0,00	Rejeitar H0	Há diferença
1NC6 vs Procian.	RMSF	0,00	Rejeitar H0	Há diferença
1NC6 vs Procian.	RMSD	0,00	Rejeitar H0	Há diferença

Hipótese nula (H0): Assume-se de que não há diferença significativa entre os dois ligantes

Com base na Tabela 8, que mostra os *p*-valores do teste de hipóteses para diferentes métricas em dinâmica molecular, pode-se concluir que: Na comparação entre o ligante-padrão e a procianidina, não há diferenças estatisticamente significativas em termos de densidade (a qual representa a densidade média de moléculas no sistema), temperatura (relativa ao estado termodinâmico do sistema), pressão (pressão exercida no sistema), potencial (energia potencial total do sistema) e a Área de Superfície Acessível ao Solvente (SASA - medida da superfície de uma proteína que é acessível ao solvente). Os *p*-valores para essas métricas

são todos superiores a 0,05 (nível de significância α), o que indica a falha em rejeitar a hipótese nula (H^0). Isso sugere que, para essas métricas, as propriedades do ligante-padrão e da procianidina não diferem significativamente.

Em contrapartida, ao observar o raio de giro (medida da extensão espacial do ligante), a Raiz da Flutuação Quadrática Média (RMSF – indica o grau de flexibilidade ou movimento de cada resíduo na proteína) e a Raiz do Desvio Quadrático Médio (RMSD – mede a diferença média na posição de cada átomo, usualmente para comparar duas estruturas), os *p*-valores são zero. Isso implica na rejeição da hipótese nula, indicando que há diferenças significativas nestas métricas entre o ligante-padrão e a procianidina.

A procianidina apresenta características similares ao ligante-padrão em termos de densidade, temperatura, pressão, energia potencial e SASA. No entanto, exibe diferenças significativas quando se considera o raio de giro, RMSF e RMSD, sugerindo comportamentos distintos destes ligantes nessas particularidades encontradas nos diferentes sistemas complexos.

A respeito das métricas que apresentam diferenças significativas, ressalta-se que:

O raio de giro (RG) é uma métrica que indica o tamanho da molécula, mais precisamente o quanto uma molécula se estende a partir do seu centro de massa. Quanto maior o RG, mais "estendida" está a molécula no sítio proteico.

O ligante padrão apresenta um RG de 1,64 (0,005 nm) enquanto a procianidina possui 1,65 (0,001 nm). Os valores são muito próximos, indicando que ambas as moléculas têm tamanhos similares. No entanto, a procianidina tem um valor de RG ligeiramente superior e um desvio padrão menor, o que significa que sua conformação é um pouco mais estendida que o padrão.

Sobre a Raiz da Flutuação Quadrática Média (RMSF), responsável por medir a mobilidade média de cada átomo em relação à sua posição ao longo da simulação, faz um indicativo de que quanto maior o valor RMSF maior flexibilidade molecular. Então, comparativamente, a procianidina (0,11 0,06 nm) exibe uma flutuação média ligeiramente superior à do ligante padrão (0,08 0,05 nm),

sugerindo que a procianidina pode ter regiões ou átomos mais móveis ou flexíveis no decorrer da simulação.

Outra métrica que sugere diferenças estatísticas significativas é o RMSD, que mede o quanto a estrutura de uma molécula desvia, em média, de uma referência ao longo do tempo. Valores mais altos de RMSD indicam que a conformação da molécula se alterou mais em relação à estrutura de referência. Na Figura 38-F é possível visualizar graficamente a diferença entre os dois compostos. Ademais, retratando numericamente, evidencia-se que a procianidina apresenta um RMSD de 0,23 (□ 0,03 nm), que é maior do que o ligante padrão 0,16 (□ 0,02 nm)). Isso implica que a procianidina pode ter sofrido mais alterações conformacionais ou deslocamentos durante a simulação em comparação com o ligante padrão.

As informações apresentadas na Figura 38-E-G com a Tabela 9, diz respeito ao processo da decomposição residual envolvendo os ligantes e o sítio reacional. Esta abordagem foi utilizada neste trabalho utilizando a decomposição residual como uma ferramenta valiosa para entender as interações moleculares em um nível mais detalhado visando a compreensão dos mecanismos biológicos entre as anotações e os receptores.

Tabela 9 - Decomposição residual referente ao complexo formado entre a proteína 1NC6 e ligante Procianidina.

RESÍDUO	VALOR PADRÃO 1NC6 kcal/mol	VALOR LIGANTE PROCIANIDINA kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
TRP:215	-2,22	-1,81	Comum	H bond	vdW
GLN:192	-1,88	-1,40	Comum	Pi-donor .H.B*	Pi-donor H. B*
CYS:220	-1,49	-0,10	Comum	H bond	vdW
CYS:191	-1,35	-0,87	Comum	vdW	vdW
GLY:226	-0,96	----	Padrão	vdW	vdW
GLY:216	-0,95	-0,43	Comum	vdW	H bond
HIS:57	-0,64	-2,00	Comum	vdW	Pi-pi-t-shaped
SER:195	-0,60	-0,42	Comum	vdW	H bond
SER:217	-0,55	----	Padrão	vdW	vdW
VAL:213	-0,54	-0,39	Comum	vdW	vdW
SER:214	-0,41	-1,07	Comum	vdW	vdW
LEU:99	-0,40	-0,67	Comum	vdW	Pi-alkyl

GLY:219	-0,36	-0,13	Comum	H bond	H bond
GLY:193	-0,29	-0,27	Comum	vdW	vdW
ALA:221:A	-0,22	----	Padrão	vdW	vdW
SER:190	-0,20	-0,53	Comum	vdW	H bond
ASP:189	-0,08	----	Padrão	H bond	vdW
ASP:194	0,03	-0,17	Comum	vdW	vdW
TYR:228	0,20	----	Padrão	vdW	vdW
SOMA	**-12,91**	**-10,25**			

vdW: forças de van der Waals

H.B: ligação de hidrogênio convencional

A Tabela 9 foi construída para fornecer uma visão detalhada do papel de cada resíduo de aminoácido do sítio ativo da proteína, visando entender as diferenças no comportamento de ligação entre os dois ligantes e na elaboração de projetos futuros de otimização de possíveis ligantes que possam modular essa proteína. Nela, apresenta-se a decomposição energética das interações entre os resíduos de aminoácidos do sítio ativo de uma proteína com dois ligantes distintos: o ligante-padrão cujo nome IUPAC é (2S,4R)-1-Acetil-N-[(1S)-4-[(aminoiminometil)amino]-1-(2-benzotiazolilcarbonil)butil]-4-hidroxi-2-pirrolidinecarboxamide, de fórmula molecular $C_{20}H_{26}N_6O_4S$, e a procianidina. Essa decomposição é feita para entender como cada resíduo do sítio reacional contribui para a ligação do ligante à proteína, tanto em termos de energia (em kcal/mol) quanto do tipo de interação (por exemplo, ligação de hidrogênio, interações de van der Waals, etc.).

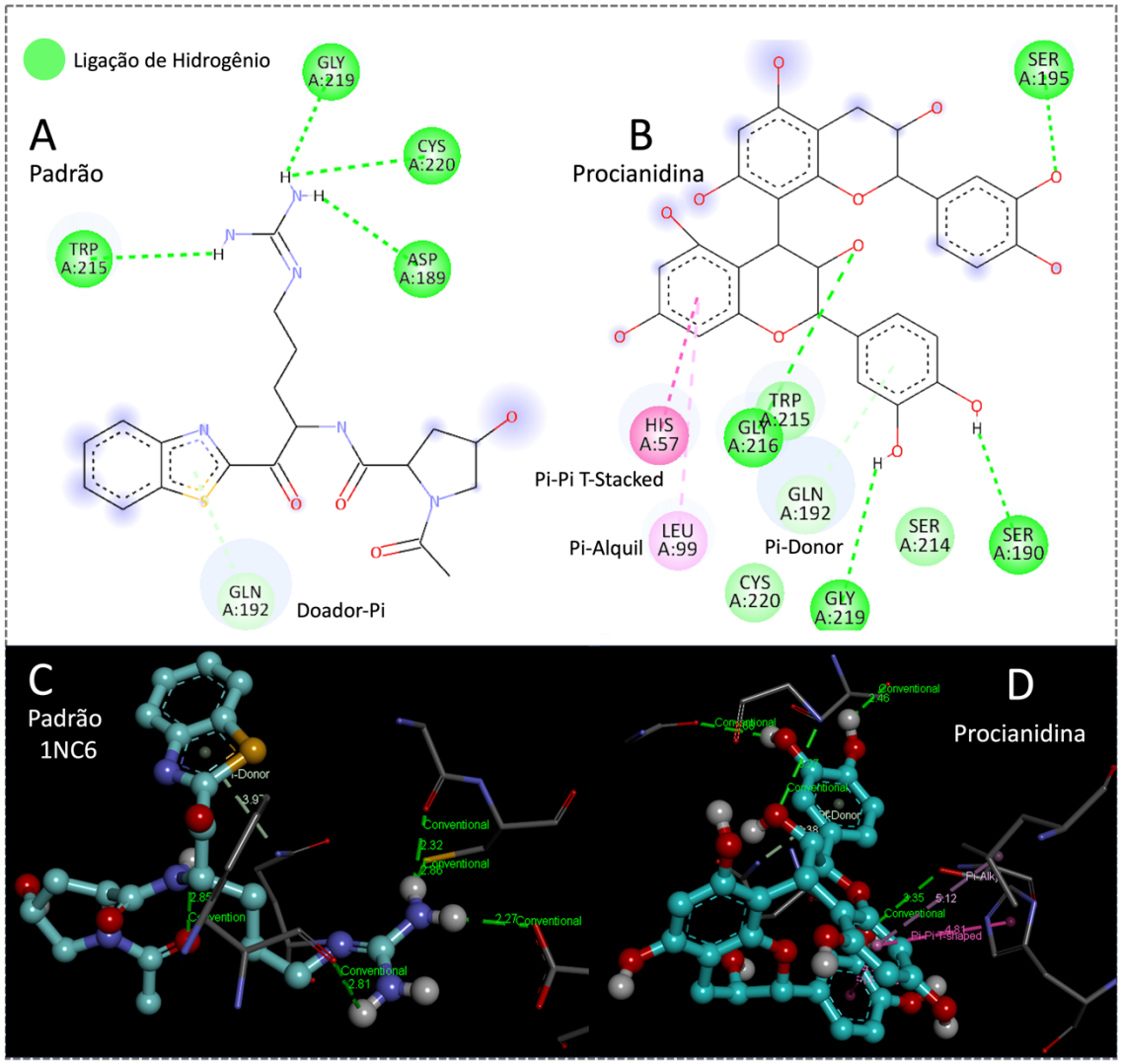
Através da decomposição das energias associadas aos resíduos de aminoácidos, obteve-se tais apontamentos:

- As energias associadas a cada resíduo de aminoácido no sítio ativo foram em sua maioria negativas, o que é esperado, pois valores negativos de energia implicam interações favoráveis.
- Resíduos que interagiram com ambos os ligantes (marcados como "comum") e outros que interagiram apenas com o ligante-padrão (marcados como "padrão"). Isso sugeriu que a procianidina poderia não ser capaz de interagir com todos os resíduos que o ligante-padrão interage, o que poderia levar a uma ligação menos estável e menos eficaz.

- O tipo de interação entre os resíduos e os ligantes também variou. Por exemplo, triptofano (TRP:215) formou uma ligação de hidrogênio com o ligante-padrão e uma interação de van der Waals com a procianidina.
- A soma total das interações para cada ligante mostrou que o ligante-padrão teve uma energia total mais negativa (-12,91 kcal/mol) em comparação com a procianidina (-10,25 kcal/mol). Isso sugeriu que o ligante-padrão poderia ter uma ligação mais forte ou favorável ao sítio ativo em comparação com a procianidina.
- A diferença de energia e os tipos de interação entre os dois ligantes e a proteína pôde indicar possíveis diferenças no modo de ligação e na eficácia da ligação de cada ligante ao sítio ativo da proteína.

A Figura 39 traz a representação das interações dos ligantes no sítio ativo da proteína 1NC6. Com destaque das representações bidimensionais e tridimensionais, que permitem sugerir as possíveis interações entre o ligante e os resíduos de aminoácidos do sítio proteico.

Figura 39 - Representação do ligante padrão (A e C) e a procianidina (B e D) no sítio ativo da proteína 1NC6



Fonte: Elaborado pelo autor (2023).

Embora a análise esteja focada nos resíduos proteicos, é evidente que forças intramoleculares (intrinsicamente presentes na molécula) desempenham um papel crucial na conformação assumida no sítio reacional. Na molécula de procianidina, a densa movimentação de carga, potencializada pelo fenômeno de ressonância, favorece a ocorrência de interações de empilhamento pi-pi- entre seus anéis aromáticos. Tais forças intramoleculares, por outro lado, não foram observadas no ligante padrão nesta pose específica.

Portanto, após uma avaliação abrangente do comportamento da molécula de procianidina no alvo 1NC6 em relação ao ligante co-cristalizado, observou-se que ambos ligantes têm tamanhos semelhantes como indicado pelos seus raios de giro. No entanto, a procianidina pareceu ser ligeiramente mais flexível e ter sofrido maiores alterações conformacionais durante a simulação em comparação com o ligante padrão, e mesmo que a procianidina tenha demonstrado afinidade pelo referido alvo, seus valores de interação não superaram aqueles do ligante original.

3.2 Dinâmica Molecular – grupo receptor 6VVU

O alvo biomolecular 6VVU, relacionado à asma (MAUN et al., 2020b), destacou-se por possuir um elevado número de anotações com alta afinidade pelo sítio proteico. Entre essas anotações estão ácido ftálico, anidrido ftálico, adenina, catequina, indol-3-acetamida, teobromina, trealose, fenilalanina e tirosina. Neste contexto, dar-se-á ênfase à discussão sobre as anotações que demonstraram potencial em modular a proteína 6VVU (à luz da dinâmica molecular), especificamente, catequina, trealose e ácido ftálico.

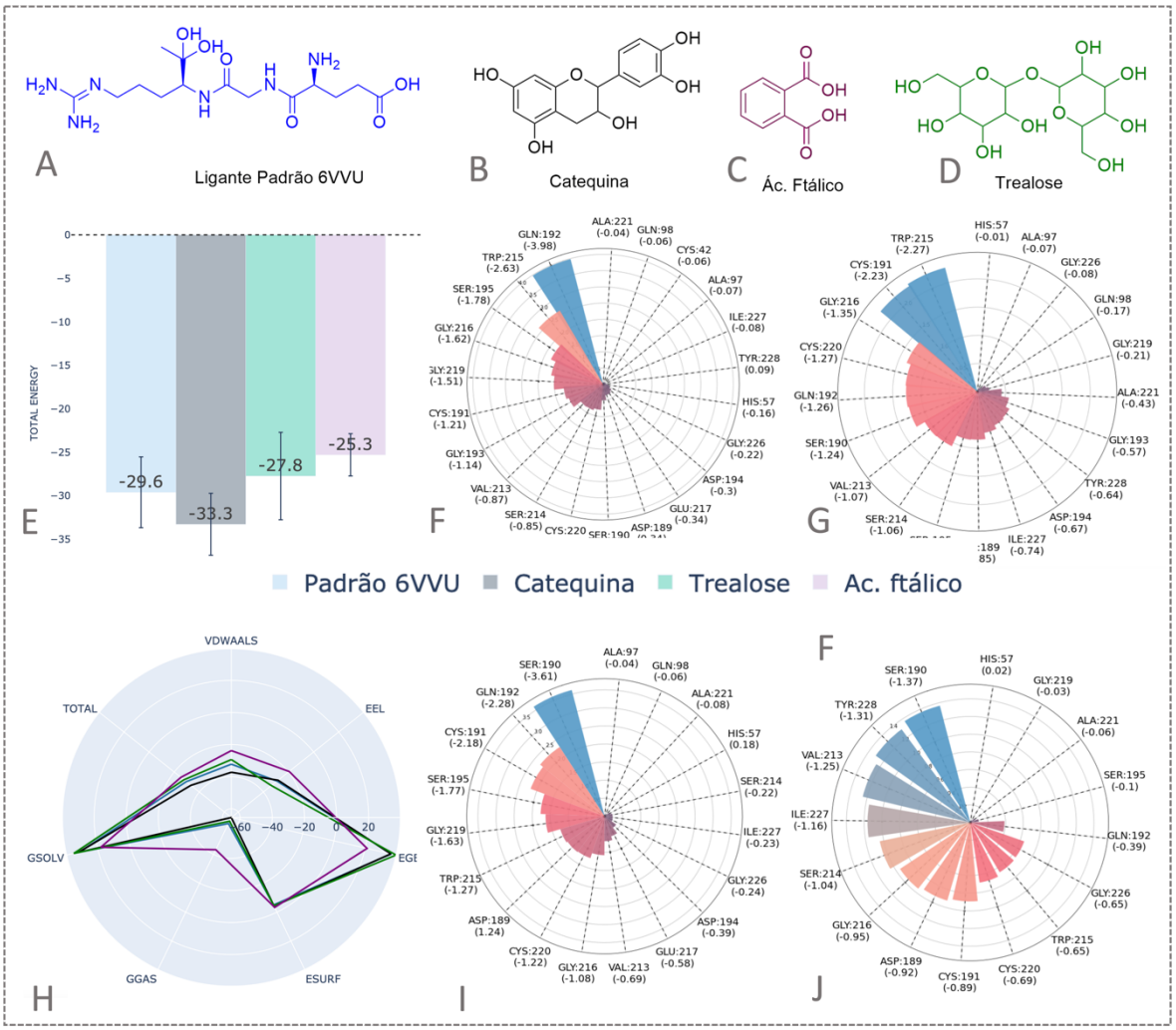
Ainda não há relatos de que o ácido ftálico possua atividade biológica em doenças respiratórias (e toda discussão referente a esta classe de moléculas foi efetuada no capítulo 2), mas que estão sendo aplicados em estudos envolvendo propriedades alelopáticas, antimicrobianas, inseticidas, além de indícios de que podem potencializar a capacidade de plantas, algas e micro-organismos a se adaptarem eficientemente a estresses bióticos e abióticos (HUANG et al., 2021).

Por outro lado, devido às suas propriedades anti-inflamatórias e antioxidantes, as catequinas podem desempenhar um papel na proteção contra doenças respiratórias inflamatórias, como a asma e a doenças pulmonares obstrutivas (FURUSHIMA et al., 2020; GHAYUR; KHAN; GILANI, 2007; UMEDA et al., 2021b). Em relação à trealose, ela emerge como um potencial aliado no tratamento da silicose, uma doença pulmonar ocupacional que apresenta inflamação persistente e fibrose irreversível (HE et al., 2020). Esta doença é desencadeada principalmente pela fagocitose de partículas de sílica cristalina por

macrófagos alveolares, provocando inflamação, apoptose e fibrose pulmonar (LI et al., 2016).

As demais estruturas, por sua vez, registraram valores reduzidos de energia total em interação com o receptor (quando comparadas ao ligante originalmente co-cristalizado com o receptor 6VVU). Detalhes, gráficos e tabelas referentes a estes resultados estão disponíveis no anexo do capítulo 6.

Figura 40 - Resultados da DM das anotações com melhores valores de afinidade pela proteína 6VVU (A – D). Energia total dos ligantes (E). Gráfico de radar para a energia (H) decomp. residual padrão (F), catequina (G), trealose (I) e ác. ftálico (J).



Fonte: elaborado pelo autor (2023).

A análise dos dados oriundos da dinâmica molecular das três anotações mais promissoras a interagirem com o alvo 6VVU, quando comparadas com o ligante co-cristalizado, permitiu discernir nuances específicas das interações moleculares e do comportamento termodinâmico de cada entidade química. Essa investigação visa realçar a capacidade diferencial de cada molécula para se associar ao sítio ativo, e fornecer uma abordagem robusta para aprimorar a compreensão das características que tornam um ligante mais adequado do que outro.

O destaque na Figura 40-H, que é representado pelo gráfico de radar, reforçou a notável variação do ácido ftálico nas métricas de GGAS e EGB. Essas diferenças também foram acompanhadas de leves desvios em WDWAALS e EEL. A estrutura do ácido ftálico, caracterizada por um anel aromático único e duas funções ácido carboxílico nas posições orto e meta, pode ser um influenciador direto nesse comportamento, por ser uma estrutura pequena (menor massa molecular entre os compostos aqui comparados), e devido ao fenômeno da ressonância no anel aromático, pôde permitir que interações associadas a essa região o tornasse favorável a se ligar neste sítio reacional. Esta configuração sugeriu que o ácido ftálico tivesse menos capacidade de estabelecer interações no sítio reacional, possivelmente devido à sua disposição espacial e características eletrônicas.

Expandindo a análise para a energia total dos quatro sistemas (três anotações e o ligante padrão), observou-se uma sobreposição nos resultados quando considerados os desvios-padrão. Tal sobreposição destacou a importância de avaliar os resultados sob múltiplas perspectivas. Particularmente na Figura 40-E, é apresentada a afinidade da molécula de catequina pelo alvo proteico 6VVU, superando inclusive o ligante padrão, seguida pela trealose e, posteriormente, pelo ácido ftálico. A catequina, representada em cinza no gráfico, destacou-se não apenas por sua afinidade, mas também pelas métricas adicionais de RMSF, que indicam sua flexibilidade quando ligada ao alvo. A maior variação de RMSF da catequina em relação ao ligante padrão sugeriu uma dinâmica diferenciada, possivelmente levando a uma interação mais adaptável com o sítio de ligação.

Em relação às avaliações estatísticas, o ácido ftálico e a trealose demonstraram diferenças significativas nas métricas de Raio de Giro e RMSD em

relação ao ligante padrão, conforme mostrado nas Tabela Suplementar 21 e Tabela Suplementar 24, respectivamente. Estas métricas foram cruciais, pois o Raio de Giro é interpretado como uma medida do tamanho efetivo da molécula, e o RMSD faz o indicativo do quanto a conformação de uma molécula varia ao longo do tempo. As diferenças observadas para ambas as métricas sugeriram que essas moléculas puderam adotar conformações distintas e tivessem tamanhos efetivos diferentes em comparação com o ligante padrão.

Essas descobertas evidenciaram uma complexidade subjacente nas interações destas moléculas com o alvo proteico 6VVU. As variações nas métricas, especialmente no Raio de Giro e RMSD, assim como a diferença destacada de RMSF para a catequina, lançam luz sobre as distintas dinâmicas e conformações que essas moléculas podem adotar no referido alvo.

Com o intuito de se explorar o potencial terapêutico destas moléculas e compreender de maneira sistemática suas interações no ambiente proteico, efetuou-se a decomposição residual dos ligantes, associando com informações oriundas da docagem molecular, para sugerir possíveis interações e energias associadas as eles.

As Tabela Suplementar 9, e Tabela Suplementar 15 detalham as interações energéticas entre o conjunto de resíduos de um alvo proteico (6VVU) e os ligantes analisados, sempre comparando-os com o padrão. A decomposição residual permite uma compreensão granular das interações em nível atômico ou de resíduos e é fundamental para entender as nuances de ligação de diferentes moléculas ao mesmo alvo proteico.

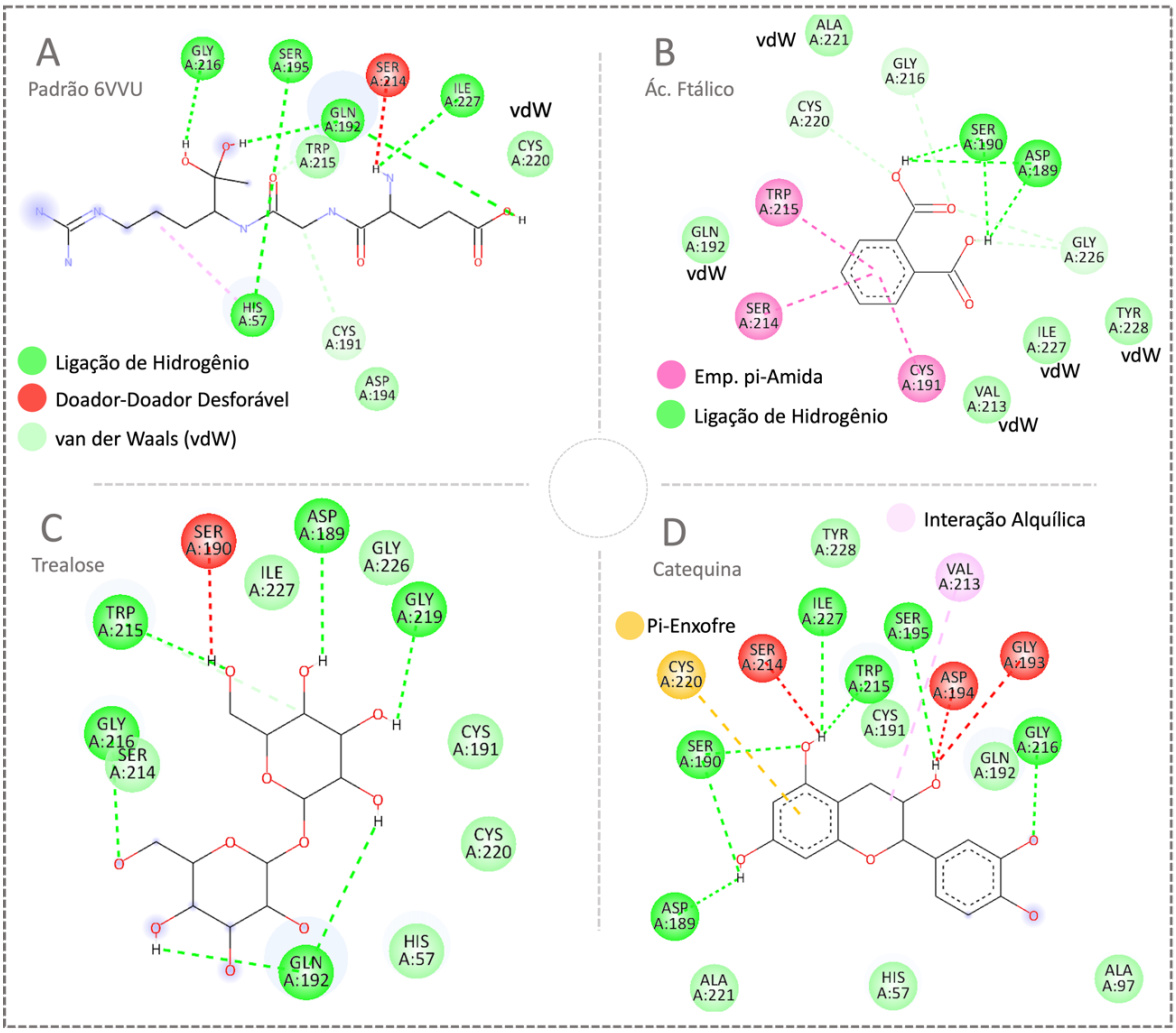
Primeiramente, observou-se os valores energéticos de interação para cada resíduo. Um valor negativo indica uma interação favorável (estabilizante), enquanto um valor positivo indica uma interação desfavorável (destabilizante). O ligante padrão para a proteína 6VVU apresentou energia total de -17,78 kcal/mol, servindo como uma referência para avaliar a eficácia dos demais ligantes. Ademais, a Figura 41 permite examinar as interações dos quatro ligantes (padrão, Ácido Ftálico, Catequina e Trealose) com os resíduos da proteína 6VVU, mostrando a diversidade de comportamentos e eficiência interacional do sítio ativo.

Em termos de energia total baseada na composição residual, o Ácido Ftálico apresentou -11,44 kcal/mol, ressaltando-se a forte interação com o resíduo glutamina (GLN:192). Nota-se que, para alguns resíduos, tanto o ligante padrão quanto o Ácido Ftálico compartilharam interações do tipo ligação de hidrogênio convencional, como é o caso da serina (SER190) e aspartato (ASP:185), sugerindo que o Ácido Ftálico poderia experimentar interações semelhantes ao ligante padrão, mas com intensidades diferentes.

A catequina, por sua vez, apresentou energia total de -17,06 kcal/mol. Ela também realizou interações mais fortes com resíduos semelhantes, como glutamina (GLN:192), mas em magnitudes diferentes em comparação ao ácido ftálico. Em contraste com o ácido ftálico, a Catequina apresentou predominantemente interações de tipo vdW (Van der Waals), que são extremamente importantes para indicar uma interação mais homogênea com a proteína. Porém, a diversidade de interações como pi-enxofre ocorrendo devido à estabilização aromática e o resíduo de aminoácido de cisteína (CYS-220), bem como as inúmeras quantidades de ligações de hidrogênio, configuram a alta afinidade deste composto no sítio reacional da proteína 6VVU.

Já a trealose, com uma energia total de -16,15 kcal/mol, destacou-se particularmente pela sua interação com o resíduo serina (SER:190), que possui grupamento hidroxila em sua constituição, que quando se aproxima de outro grupamento de mesma natureza, causa essa sobreposição entre grupos doadores de prótons, denominadas doadores-doares desfavoráveis, que mesmo não efetivando uma ligação de hidrogênio convencional, efetuou um papel interacional importante no sítio de reação da proteína 6VVU. Além disso, vale destacar que a molécula de trealose apresenta vários grupamentos hidroxílicos, que são funções com alto potencial de interação com alvo, com destaque para o resíduo de aminoácido de glutamina (GLN: 192), em que a carbonila do grupamento amida efetuou duas ligações de hidrogênio com partes distintas da molécula de trealose. Esta estrutura, embora tenha menos resíduos interagindo do que a catequina, efetuou interações tão robustas quanto ela.

Figura 41 - Interações no sítio reacional da proteína 6VVU envolvendo (A) – ligante padrão, (B) – ácido ftálico, (C) – Trealose e (D) – Catequina



Fonte: elaborado pelo autor (2023)

Em análise geral, ao considerar as energias totais, a catequina surge como o ligante mais eficaz, seguida de perto pela trealose e, por último, pelo ácido ftálico. No entanto, é relevante notar a diversidade de interações apresentadas pelo ácido ftálico, que sugeriu interações mais complexas com a proteína em comparação aos outros ligantes. Além disso, a forte afinidade da trealose pelo resíduo da serina (SER:190) pode ser um foco de investigação em estudos subsequentes ou na otimização de ligantes.

Em síntese, embora a catequina tenha se apresentado como o ligante mais eficaz, cada ligante teve suas próprias características únicas de interação que

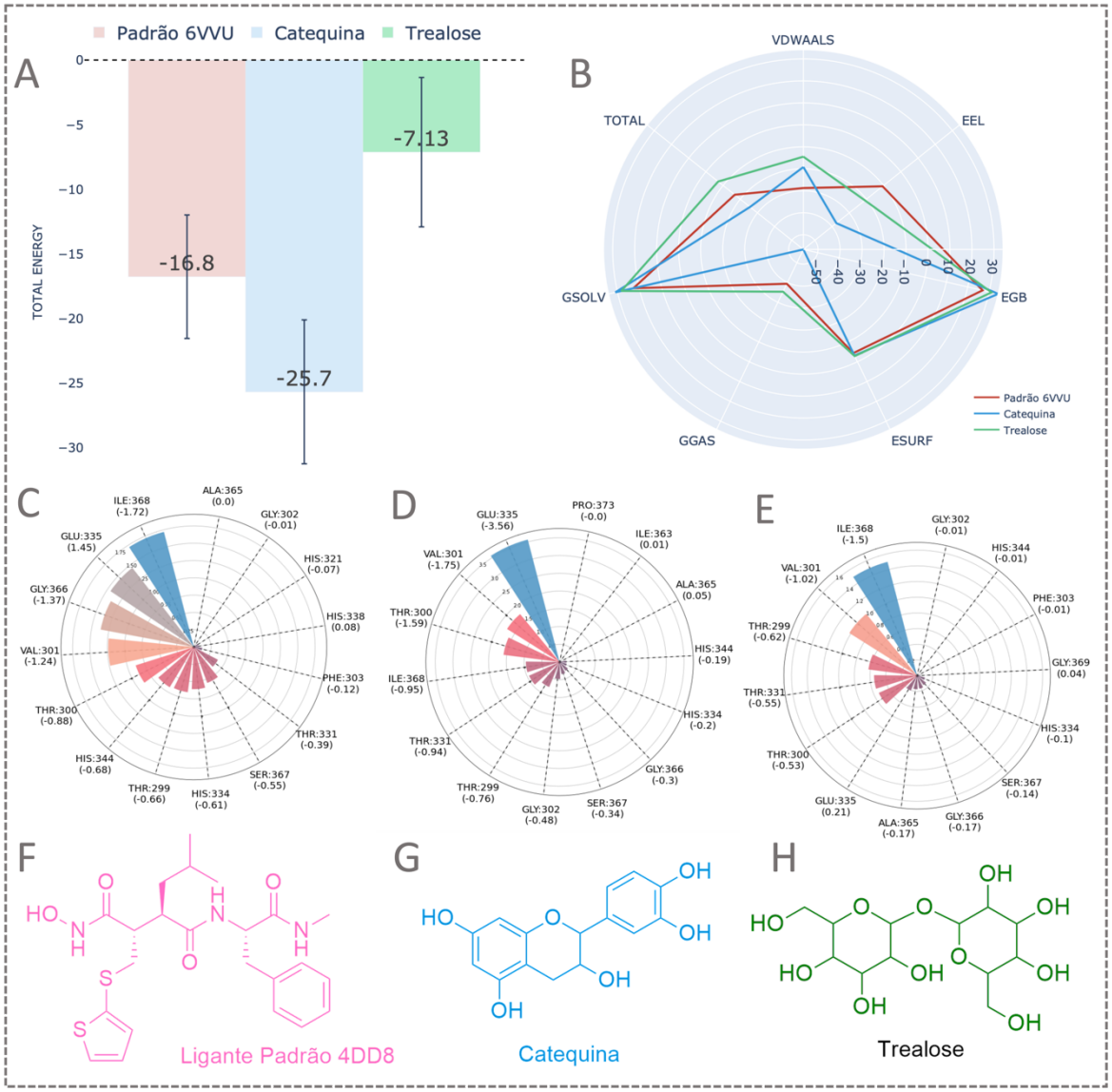
poderão ser exploradas em diferentes aplicações ou na otimização de um ligante específico. Estas observações podem guiar futuros estudos para aperfeiçoar ainda mais as moléculas, criando derivados que maximizam as interações desejadas. É também essencial observar a variedade de tipos de interações que ocorrem entre os ligantes e o alvo, como ligações de hidrogênio, interações de van der Waals, empilhamentos pi, entre outras. Estes detalhes são cruciais quando se considera a otimização de ligantes para melhor se ligarem a um alvo proteico.

3.3 Dinâmica Molecular – grupo receptor 4DD8

Apenas a molécula de catequina e trealose apresentam resultados promissores na interação com o receptor 4DD8 (também associado à doença respiratória da asma) (HALL et al., 2012b). Associado a este alvo molecular está a molécula de (2S,3R)-N-hidroxi-N'-[(2S)-1-(metilamino)-1-oxo-3-phenilpropan-2-il]-3-(2-metilpropil)-2-(tiofen-2-ilsulfanilmetill)butanediamide, de fórmula molecular $C_{23}H_{31}N_3O_4S_2$ (massa molecular de 477,6 $g.mol^{-1}$), e foi utilizado como limiar comparativo das demais anotações estudadas.

Devido à capacidade plural de conseguir atuar em diferentes alvos biomacromoleculares, as moléculas de catequina e trealose apresentam uma propriedade denominada promiscuidade química (HU; BAJORATH, 2013). Compostos que atuam sobre diferentes alvos despertam interesse não somente no campo da polifarmacologia, mas também ajudam a aprofundar a compreensão sobre a maneira como moléculas menores estabelecem interações particulares em variados contextos de locais de ligação (FELDMANN et al., 2019).

Figura 42 - Resultados da DM das anotações com melhores afinidades pela proteína 4DD8 (F – G). Energia total dos três ligantes (A). Gráfico de radar valores de energia (B) e gráficos de decomp. residual para o padrão (C), catequina (D), trealose (E).



Fonte: Elaborado pelo autor (2023).

Nesta classe de moléculas envolvidas na modulação da proteína 4DD8, destacou-se novamente a molécula de catequina frente ao ligante de referência, que apresentou valores absolutos de energia de ligação de -25,7 kcal/mol, comparados com -16,8 kcal/mol do padrão e -7,13 kcal/mol para a molécula de trealose. Porém, ao considerar-se os desvios-padrão é possível que em algum momento da trajetória molecular, a interação com o alvo proteico apresentasse

valores de afinidade semelhantes, como apresentado no gráfico de barras da Figura 42-A.

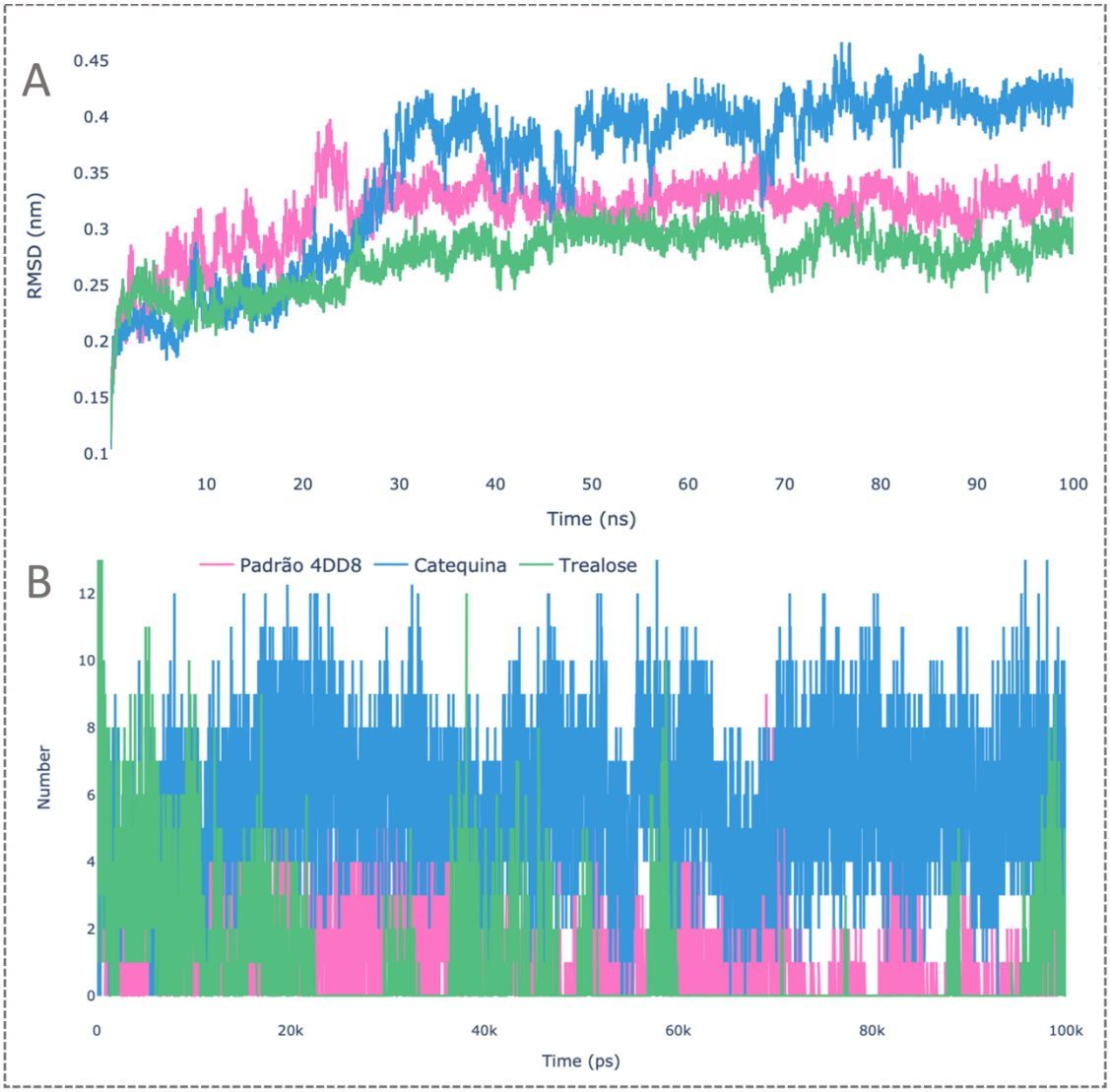
As Tabela Suplementar 18 e 19 compilam informações sobre as métricas obtidas da trajetória de dinâmica molecular da catequina e da trealose, respectivamente, permitindo análises compreensivas de que para a maioria das métricas, incluindo Densidade, Temperatura, Pressão, Potencial, SASA e RMSF, não houve diferenças significativas entre a proteína na presença do ligante padrão 4DD8 e os ligantes investigados.

Contudo, duas métricas revelaram-se notavelmente distintas: Raio de Giro e RMSD. Para ambas as comparações, diferenças significativas nesses parâmetros foram observadas, sugerindo modificações estruturais ou dinâmicas na presença destes ligantes em comparação com o ligante padrão. Uma hipótese para os valores distintos do Raio de Giro, que é uma métrica que quantifica a distribuição espacial de uma molécula, pode estar relacionado às diferentes extensões e flexibilidade que os ligantes possam apresentar. O ligante padrão apresenta massa molecular menor (266,11 g/mol) quando comparados aos demais ligantes (catequina com 290,08 g/mol e trealose (342,12 g/mol). Embora, estruturalmente, as moléculas apresentem ciclos (incluindo aromáticos), que permitem interações similares ao sítio reacional a diferença estatística no raio de giro sugere que a proteína, ao interagir com trealose ou catequina, pode se reconfigurar-se de uma maneira diferente do que faz com o ligante padrão 4DD8.

O ligante padrão associado à proteína 4DD8 demonstrou média de RMSD de 0,314 (± 0,03 nm), o que indica uma conformação relativamente estável e consistente para este ligante. Em contraste, a catequina apresentou um RMSD médio mais elevado, de 0,35 (± 0,07). Este valor sugere que a catequina pode adotar conformações ligeiramente mais variadas quando ligada ao sítio de ligação, possivelmente devido a diferentes modos de interação ou maior flexibilidade molecular. Por último, a trealose se destacou como o ligante com a conformação mais estável, tendo o menor RMSD médio de 0,27 (± 0,02 nm). Este valor indica que a trealose pode ter uma interação mais específica e estável com o sítio de

ligação da proteína 4DD8. As diferenças de RMSD entre os três ligantes podem ser visualizadas na Figura 43-A.

Figura 43 - Comparação das métricas de RMSD (A) e número de ligações de hidrogênio (B) dos complexos formados entre o ligante padrão (rosa), catequina (azul) e trealose (verde) com a proteína 4DD8.



Fonte: elaborado pelo autor

A estabilidade observada na conformação da trealose pôde indicar um modo de ligação mais específico com o sítio de ligação da proteína, enquanto a maior variabilidade da catequina sugeriu múltiplos possíveis modos de interação ou pontos de interação. No entanto, até o momento, estas hipóteses são apontadas

com cautela, pois outras métricas e ensaios complementares são necessários para fornecer um panorama completo das interações em estudo, como a decomposição residual das interações dos ligantes com o sítio ativo da proteína 4DD8 e a possibilidade de formação de ligação de hidrogênio, que em estudos de dinâmica molecular podem proporcionar insights cruciais sobre a natureza e estabilidade das interações moleculares, particularmente nas interações proteína-ligante (GUÀRDIA et al., 2005).

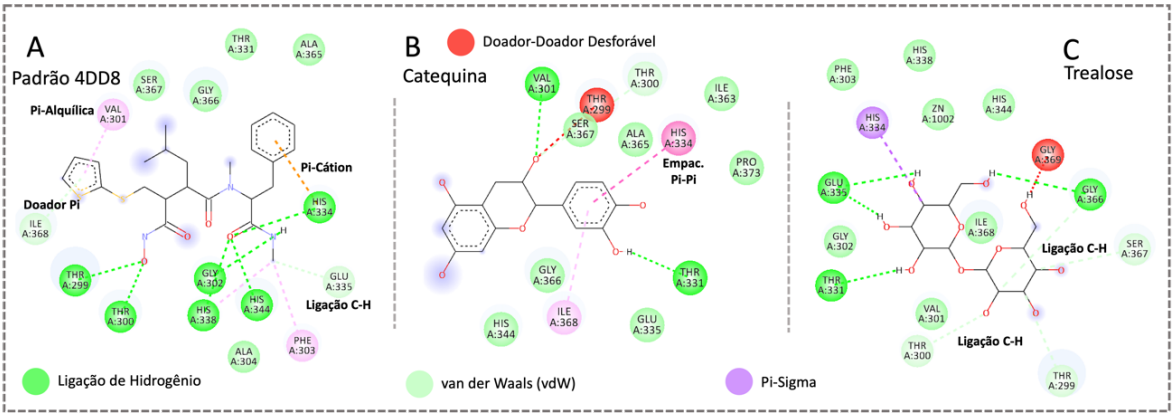
Essas ligações são vitais na determinação da especificidade das interações, pois contribuem significativamente para a estabilidade da complexação. Além disso, o exame das ligações de hidrogênio pode elucidar os mecanismos subjacentes pelos quais um ligante se acopla ou se dissocia de seu alvo proteico. Do ponto de vista do design de fármacos, as ligações de hidrogênio são frequentemente consideradas de grande importância para a atividade biológica (YANG; FANG; JI, 2016). A formação dessas interações pode ser um guia na orientação de modificações moleculares para aprimorar tanto a afinidade quanto a especificidade de um ligante em relação ao seu alvo. Além das interações proteína-ligante, as ligações de hidrogênio também desempenham um papel crítico no dobramento, estabilidade e função das proteínas (GHIANDONI; CALDEWEYHER, 2023) .

No contexto deste estudo, observou-se que, ao longo da trajetória molecular de 100 ns, o ligante padrão apresenta média de 1,03 ligações de hidrogênio com um desvio padrão de mais ou menos 1,25. Em contraste, a catequina, apresentou média de 5,8 ligações de hidrogênio com um desvio de mais ou menos 2,03. Neste caso, a trealose, que havia apresentado resultados melhores em RMSD, exibiu apenas 0,83 ligações de hidrogênio, com um desvio de mais ou menos 1,6. Essas diferenças na formação de ligações de hidrogênio podem indicar variações na maneira como cada ligante interage com a proteína-alvo e, consequentemente, podem ter implicações significativas para a sua atividade biológica e estabilidade da complexação (HUBBARD; KAMRAN HAIDER, 2010). As ligações de hidrogênio do grupo comparativo podem ser visualizadas na Figura 43-B.

No estudo de decomposição residual efetuado entre os ligantes e o sítio reacional da proteína 4DD8, observou-se que em relação à interação energética

residual, a catequina apresentou um valor mais negativo de -11,01 kcal/mol, indicando uma interação mais forte com o receptor em comparação ao valor do ligante padrão, de -6,78 kcal/mol. Por outro lado, a trealose revelou uma interação menos intensa, com uma energia de -4,59 kcal/mol. Isso sugeriu que a catequina pudesse formar um complexo mais estável com o alvo biomacromolecular, enquanto a trealose poderia potencialmente formar um complexo menos estável ou mais dinâmico. As sugestões das interações podem ser visualizadas na Figura 44, e os valores associados a cada um dos resíduos estão compilados na Tabela Suplementar 7 (sobre catequina) e Tabela Suplementar 8 (trealose).

Figura 44 - Interações no sítio reacional da proteína 4DD8 envolvendo (A) – ligante padrão, (B) – Catequina, (C) – Trealose.



Fonte: elaborado pelo autor (2023).

Examinando interações específicas, foi observado que a catequina mostrou a maior diferença positiva de 1,07 kcal/mol no resíduo de glicina (GLY:366), implicando uma interação mais fraca nesse local em comparação com o padrão. No entanto, a interação mais forte da catequina, de -5,01 kcal/mol, foi observada no glutamato (GLU:335).

Enquanto o padrão 4DD8 mostrou uma ligação de hidrogênio com o resíduo de treonina (THR:300), a catequina apresentou uma interação do tipo Carbono-Hidrogênio, efetuada entre o hidrogênio do carbono alfa do resíduo de aminoácido treonina (THR: 300) e um dos oxigênios da catequina. Do mesmo modo, a trealose efetuou interações do tipo Carbono-Hidrogênio com a THR:300. Porém, essa

diversidade de grupamentos funcionais dos ligantes e a capacidade de conformações variadas, sugeriu diferentes modos de ligação ou orientações no sítio de ligação, o que impactou em implicações funcionais e modulação da proteína de maneira diversa.

Além disso, tanto a catequina quanto a trealose apresentaram interações exclusivas. A catequina interagiu especificamente com os resíduos de isoleucina (ILE:363) e procianidina (PRO:373), enquanto a trealose mostrou uma interação exclusiva com o resíduo de glicina (GLY:369). Essas interações exclusivas sugeriram que cada ligante poderia potencialmente influenciar o padrão 4DD8 de maneiras únicas, levando a diferentes conformações ou funções do complexo ligante-alvo.

3.4 Dinâmica Molecular – grupo receptor 7P2G

Dada a natureza promíscua tanto da catequina quanto da trealose, ambas as moléculas mostraram potencial para se ligar em mais um alvo proteico, nessa vez, a associação foi feita ao sítio ativo do receptor 7P2G, proteína associada à síndrome aguda grave (SARS-CoV-2) (ROSSETTI et al., 2022a). Em concordância com informações disponíveis na literatura, como as apresentadas por HENSS e sua equipe (2021c), a catequina tem sido foco de estudos e implementada em projetos que exploram sua eficácia contra infecções por SARS-CoV-2. Embora ainda não exista um tratamento específico para COVID-19, estratégias terapêuticas complementares e alternativas estão sendo avaliadas. Dentre elas, destaca-se o uso da trealose, conforme destacado por PAJOKH e POURFRIDONI (2021), reforçando os achados *in silico* apresentados neste estudo.

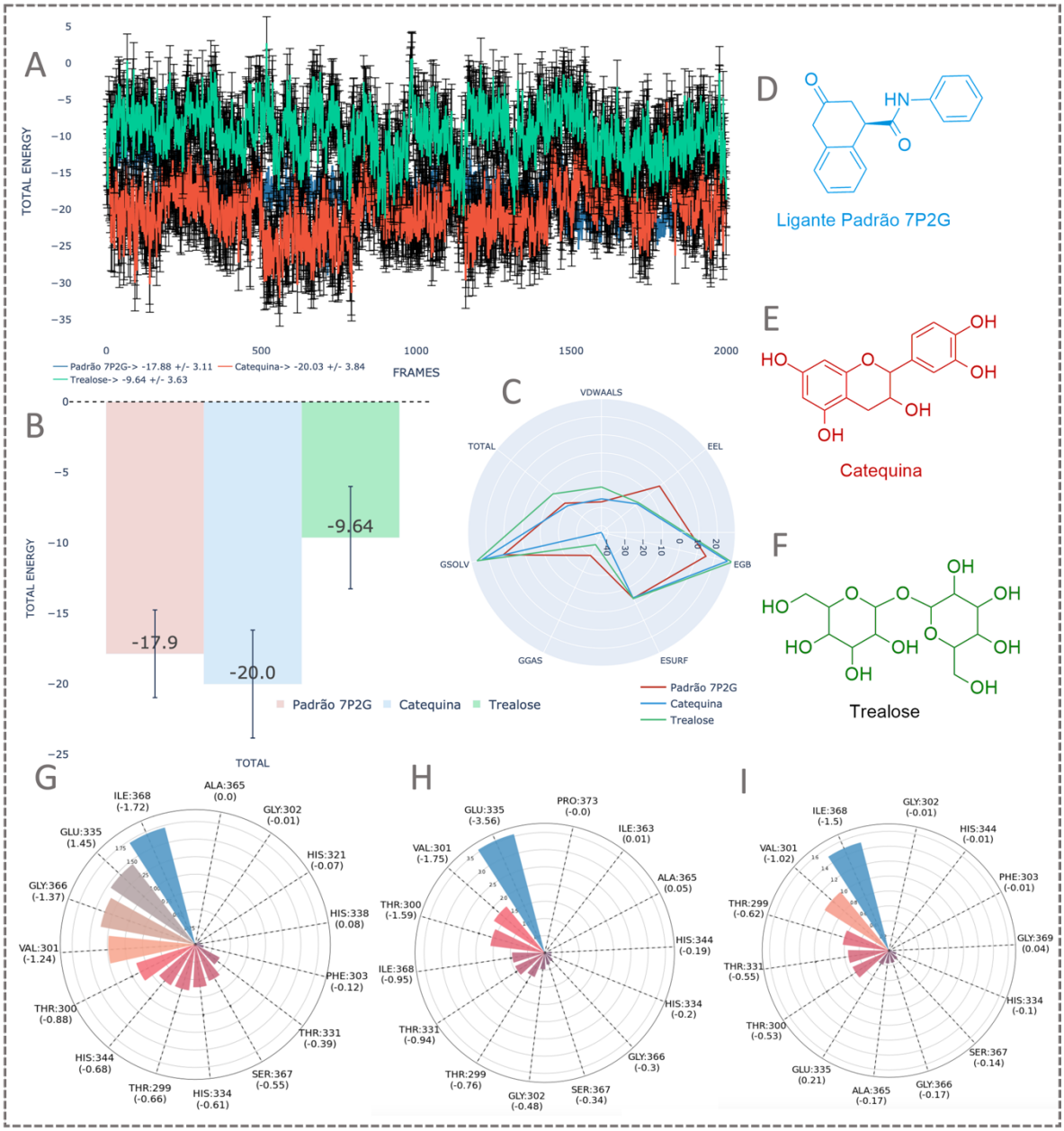
Do mesmo modo que os demais grupos de receptores discutidos, os resultados de dinâmica molecular são comparados com o ligante padrão, originalmente co-cristalizado com a proteína. O ligante associado ao alvo biomacromolecular 7P2G apresenta fórmula molecular $C_{16}H_{13}N_2O_2$ e nome IUPAC (4R)-N-(4-iodofenill)-2-oxo-3,4-dihidro-1H-quinolina-4-carboxamida. Analisando o gráfico de energia total de ligação, mostrado na Figura 45-A, B e C, que é notável

que a molécula de catequina (Figura 45-D) prevaleceu com melhor afinidade pelo alvo proteico ao longo de toda a trajetória de 100 ns de dinâmica molecular, seguida pelo ligante padrão (Figura 45-E) e, por último, a trealose (Figura 45-F). Neste caso, a molécula de trealose, mesmo considerando os desvios-padrão envolvidos, não conseguiu efetivamente modular a proteína como os outros ligantes.

Analisando-se as métricas complementares, contidas nas Tabela Suplementar 28Tabela Suplementar 29, notou-se que em relação a densidade, a similaridade entre o ligante padrão e os compostos trealose e catequina sugeriu que ambos pudessem ter uma ocupação espacial parecida no sítio ativo do receptor conduzindo a interações análogas, seja ligando-se a sub-regiões idênticas ou envolvendo-se com os mesmos resíduos de aminoácidos.

Do mesmo modo, sobre a temperatura e pressão, obteve-se valores semelhantes entre o padrão, a catequina e a trealose, indicando que, quando ligados ao receptor, esses ligantes induziram ou enfrentaram flutuações térmicas e pressões internas parecidas. Isso pode apontar que todos os ligantes estudados contribuíram para manter a estabilidade do receptor de forma análoga, sem causar perturbações estruturais expressivas.

Figura 45 - Resultados da DM das anotações com melhores afinidades pela proteína 7P2G (D, E e F). Energia total de ligação dos ligantes (A). Distribuição de energia e desvios-padrão (B), Energia total do sistema (C) e decomp. residual para os ligantes



Fonte: Elaborado pelo autor (2023).

A métrica SASA (Área de Superfície Acessível ao Solvente) foi outro ponto de convergência entre o padrão e os dois ligantes naturais. O fato de não haver diferença significativa sugeriu que, ao se ligarem ao receptor, todos os compostos apresentaram perfis de solvatação semelhantes, sugerindo que possam ter adotado

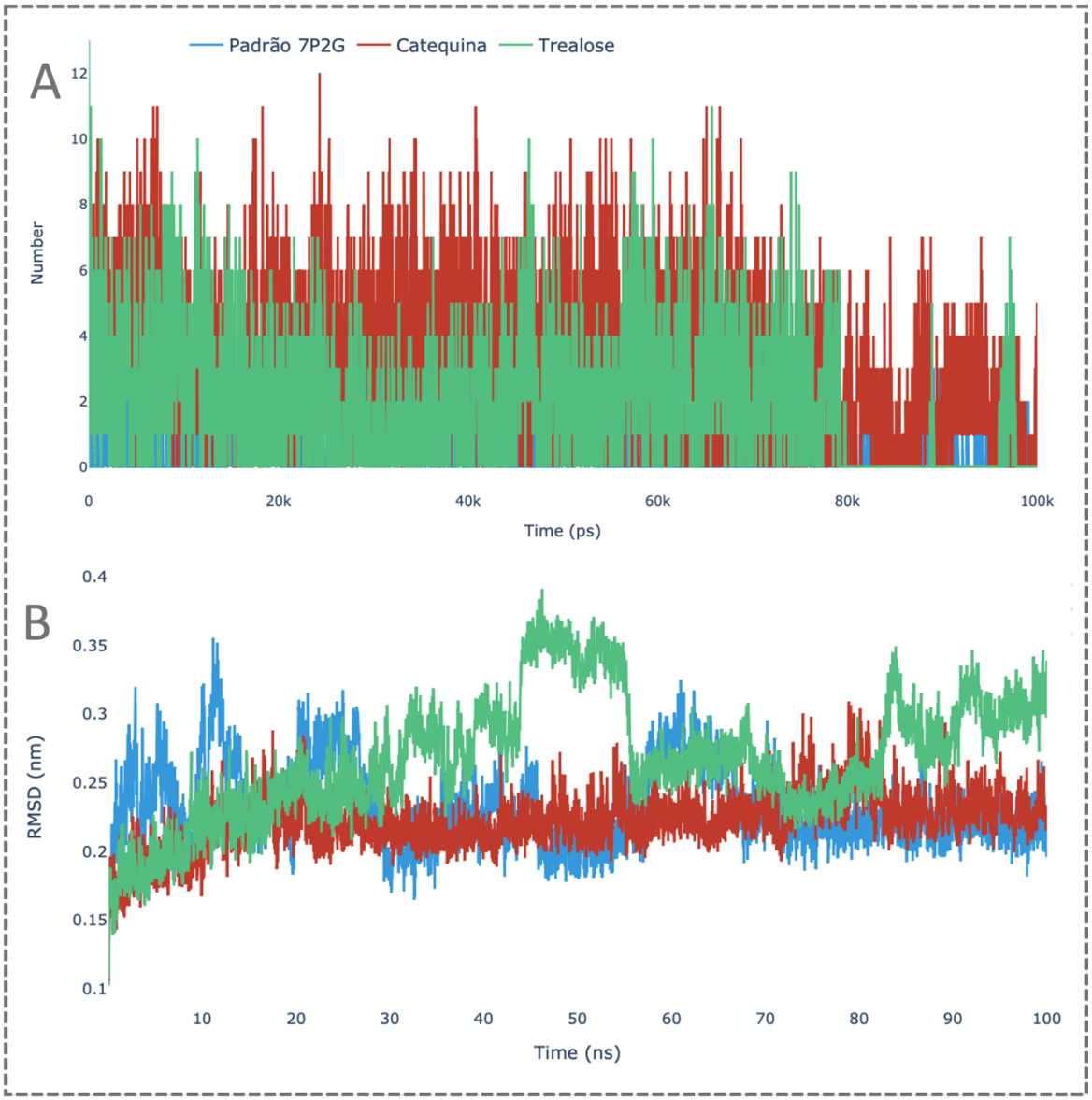
conformações no sítio de ligação que permitem que áreas moleculares entre em contato com o solvente, refletindo em propriedades de solvatação comparáveis.

Em relação ao raio de giro, a diferença significativa observada entre o padrão e ambos os compostos, catequina e trealose, sugeriu uma variação na compactação das estruturas quando ligadas ao receptor. Esta métrica avalia o grau de espalhamento da molécula em torno de seu centro de massa. Portanto, uma variação nessa métrica pode indicar que a catequina e a trealose induziram conformações estruturais diferentes no receptor, comparadas com ligante co-cristalizado à proteína 7P2G.

Quanto ao RMSD, contemplada na Figura 46-A, a diferença observada entre o ligante padrão e os dois compostos sugeriu que as conformações adotadas pelo receptor quando ligado a esses compostos são distintas. Esta métrica aponta a variação na posição dos átomos de uma proteína ao longo do tempo. Assim, um RMSD diferente pode indicar uma diferença na dinâmica estrutural ou na estabilidade da conformação induzida por cada ligante.

Por outro lado, em relação à quantidade de ligações de hidrogênio estabelecidas entre os ligantes e receptores (Figura 46-B) reforçou-se que o ligante padrão praticamente não efetuou sua interação por meio de ligações de hidrogênio, diferentemente da catequina e da trealose, que estruturalmente (devido a quantidade de grupamentos hidroxilas) efetivaram este tipo de interação com o bioreceptor.

Figura 46 - Comparação das métricas de RMSD (A) e número de ligações de hidrogênio (B) dos complexos formados entre o ligante padrão (azul), catequina (vermelho) e trealose (verde) com a proteína 7P2G.



Fonte: Elaborado pelo autor (2023).

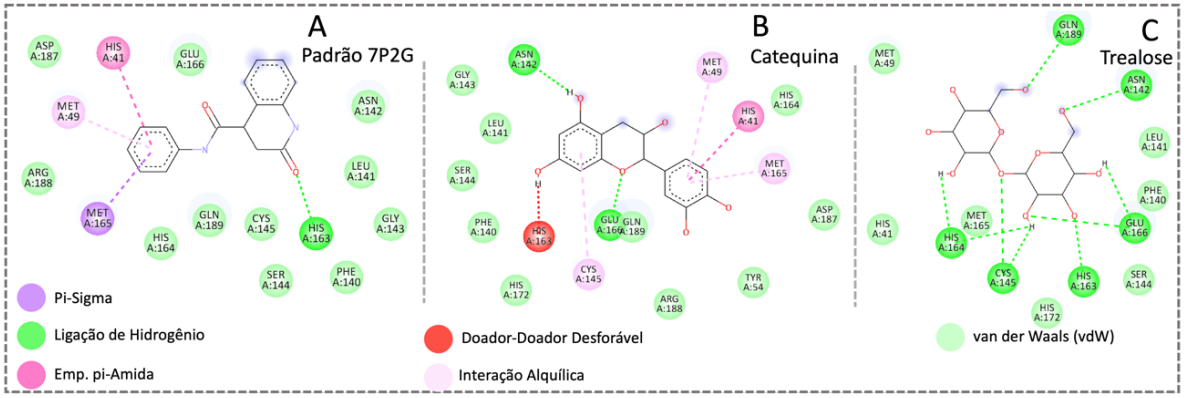
O RMSF (Raiz da Flutuação Quadrática Média), que também apresentou diferença significativa para a trealose em comparação com 7P2G, sugeriu que a ligação da trealose influenciou a flexibilidade de regiões específicas do receptor de maneira diferente do padrão.

Por fim, a diferença significativa na métrica de potencial energético entre catequina e em comparação com o padrão foi notável. Esta métrica avalia a energia

potencial do sistema, e uma variação sugeriu que a catequina induzisse estados energéticos distintos no receptor comparados ao padrão. A energia potencial comparativa entre padrão e trealose não foi estatisticamente distinta.

Ao avaliar as interações moleculares entre os resíduos do padrão 7P2G e os ligantes catequina e trealose, notou-se padrões distintos e significativos de ligação que ofereceram insights sobre o comportamento bioquímico destes ligantes. As sugestões das ligações existentes entre os ligantes e os resíduos de aminoácidos do sítio reacional estão mostrados na Figura 47.

Figura 47 - Interações no sítio reacional da proteína 7P2G envolvendo (A) – ligante padrão, (B) – Catequina, (C) – Trealose.



Fonte: Elaborado pelo autor (2023).

Para o ligante catequina, observou-se variações de interação quando comparado ao padrão do receptor 7P2G. Notavelmente, no resíduo de metionina (MET:165), a catequina formou uma ligação pi-alquílica, em comparação com a interação pi-sigma observada no ligante padrão. Esta mudança poderia potencialmente traduzir-se em uma ligação mais estável, como é sugerido pela diferença de energia de -0,45 kcal/mol.

Da mesma forma, os resíduos de aspartato (ASP:187) e de serina (SER:144) destacaram-se, mostrando diferenças substanciais na energia de ligação quando a catequina está envolvida. Tais diferenças, em combinação com a observação de que a catequina pode formar ligação de hidrogênio no resíduo de glutamato (GLU:166), indicaram que este ligante poderia ter uma afinidade distinta, ou até

mesmo maior, para o receptor em comparação com padrão. Além disso, os resíduos de histidina (HIS:172) e glicina (GLY:143), que não mostraram interação com padrão, efetuaram interações com a catequina. Esta capacidade da catequina de interagir com locais adicionais no receptor pode dar indícios da sua alta afinidade pelo ambiente ativo desta proteína.

Em contraste, a trealose apresentou outro conjunto distinto de interações. Os resíduos de metionina (MET:165) e (MET:49) mostraram diferenças positivas na energia de ligação em comparação com padrão, indicando que este formou uma ligação mais estável nesses locais. Interessante é a observação no resíduo de glutamina (GLN:189), onde a trealose efetuou uma ligação de hidrogênio, indicando uma ligação potencialmente mais forte e estável em comparação com a interação VdW do padrão. Além disso, a trealose mostrou interações com histidina (HIS:172) e glicina (GLY:143), resíduos que não se ligaram com o padrão. Corroborando com as demais análises observadas, a decomposição residual também faz o indicativo de que a trealose apresentou menos efetividade pelo sítio ativo da proteína 7P2G, quando comparada ao padrão e à catequina.

4 CONCLUSÃO

Tais análises apresentadas e discutidas neste capítulo teve por objetivo realizar uma varredura entre possíveis anotações moleculares de matrizes complexas de cacau, analisando-as como ligantes em potencial para modular alvos bioquímicos. Tal abordagem não apenas proporcionou uma compreensão mais profunda das interações moleculares que tais moléculas possam efetuar com sítios ativos das proteínas associadas a doenças respiratórias, mas também pavimentou o caminho para futuras investigações que poderão ter implicações significativas em áreas como o desenvolvimento de fármacos e a biologia estrutural.

Ao identificar e compreender essas peculiaridades, pode-se preparar estratégias mais inteligentes e rápidas de design de ligantes e buscar candidatos potencialmente mais eficazes para a interação com bioreceptores desta natureza, o que pode ter amplas implicações no campo do desenvolvimento de fármaco.

Em conclusão, apesar de todos as moléculas aqui discutidas serem potenciais para atuar em doenças respiratórias, destacam-se, especialmente, a catequina por formar um complexo altamente estável e específicos com 3 dos 4 receptores propostos (6VVU, 4DD8 e 7P2G). A trealose, com interação geral mais fraca, sugeriu associações mais transitórias e flexíveis, mas estatisticamente tão consistentes quanto aos ligantes padrões das proteínas 6VVU e 4DD8 (e também da catequina). E o ácido ftálico, que apresentou afinidade pela proteína 6VVU.

Estas descobertas propiciadas por análises computacionais robustas, como as de dinâmica molecular, coletivamente sugerem que, embora catequina, trealose e ácido ftálico possam associarem-se ao mesmo receptor, a natureza de suas interações moleculares varia significativamente, modulando o receptor de maneiras distintas, mas resultando em respostas bioquímicas similares.

CONSIDERAÇÕES FINAIS

Para consolidação dessa tese de doutorado realizou-se esforços visando o desenvolvimento de ferramentas computacionais que facilitassem a análise das matrizes complexas derivadas do processo fermentativo do cacau (*Theobroma cacao*), focando na desreplicação de entidades químicas e na sua aplicação em alvos biomacromoleculares associados a doenças respiratórias, como asma e SARS-CoV-2 (Covid-19).

O capítulo 1 detalhou a criação do LUMIOS (acrônimo para **L**abel **U**sing **M**achine **I**n **O**rganic **S**amples), uma plataforma web que automatiza o processamento de dados espectrais e indica anotações de moléculas potencialmente ativas em alvos moleculares, avaliadas por técnicas de docagem. O capítulo 2 demonstrou a aplicação bem-sucedida do LUMIOS na identificação de 13 moléculas distintas em várias etapas da fermentação, destacando as estruturas mais promissoras (catequina, trealose, procianidina, teobromina, adenina, ácido ftálico, anidrido ftálico, indol-3-acetamida, fenilalanina e tirosina).

O capítulo 3 introduziu o aplicativo Chemistika, que se integra ao LUMIOS para realizar análises estatísticas resultantes do planejamento de misturas do tipo Simplex-Lattice. Os capítulos 4 e 5 ilustraram, a utilização do Chemistika para explorar as matrizes de cacau, considerando as intensidades relativas das anotações, e para analisar a variabilidade molecular ao longo da fermentação, respectivamente. O capítulo final, capítulo 6, abordou a dinâmica molecular das anotações, utilizando 100 nanosegundos de análise de trajetória, e a exploração dos dados gerados pelo GROMACS através da plataforma CHEIC, uma ferramenta que processa e sistematiza de forma inteligente os arquivos de saída do GROMACS, novamente obtendo resultados promissores para a moléculas de catequina e trealose.

Dessa forma, esta tese não apenas contribuiu para a compreensão da complexa matriz fermentativa do cacau e suas potenciais aplicações no tratamento de doenças respiratórias, mas também entregou como produto, ferramentas computacionais valiosas que podem ser aplicadas em outros contextos de

pesquisa. A integração de várias ferramentas, incluindo LUMIOS, Chemistika e CHEIC, permitiu uma abordagem abrangente e eficiente para a análise de dados, desde a identificação inicial de moléculas até a análise detalhada da dinâmica molecular. Esta pesquisa, portanto, representa um avanço na aplicação de técnicas computacionais na exploração de matrizes biológicas complexas e na identificação de novos compostos potencialmente terapêuticos oriundos de produtos naturais.

REFERÊNCIAS

ABDEL-KAREEM, M. M.; RASMEY, A. M.; ZOHRI, A. A. The action mechanism and biocontrol potentiality of novel isolates of Saccharomyces cerevisiae against the aflatoxigenic Aspergillus flavus. **Letters in applied microbiology**, v. 68, n. 2, p. 104–111, 2019.

AFOAKWA, E. O. et al. Chemical composition and physical quality characteristics of Ghanaian cocoa beans as affected by pulp pre-conditioning and fermentation. **Journal of Food Science and Technology**, v. 50, n. 6, p. 1097–1105, dez. 2013.

AGYIRIFO, D. S. et al. Metagenomics analysis of cocoa bean fermentation microbiome identifying species diversity and putative functional capabilities. **Heliyon**, v. 5, n. 7, 2019.

AHMED, J. et al. SuperSweet—a resource on natural and artificial sweetening agents. **Nucleic acids research**, v. 39, n. suppl_1, p. D377–D382, 2010.

AKSIMENTIEV, A.; SCHULTEN, K. Imaging α-hemolysin with molecular dynamics: Ionic conductance, osmotic permeability, and the electrostatic potential map. **Biophysical Journal**, v. 88, n. 6, p. 3745–3761, 2005.

ALIN, A. Minitab. **Wiley interdisciplinary reviews: computational statistics**, v. 2, n. 6, p. 723–727, 2010.

ANDÚJAR, I. et al. **Cocoa polyphenols and their potential benefits for human health**. **Oxidative Medicine and Cellular Longevity**, 2012.

APROTOSOAIE, A. et al. The Cardiovascular Effects of Cocoa Polyphenols—An Overview. **Diseases**, v. 4, n. 4, p. 39, dez. 2016.

ARBORETTI, R. et al. **Design of Experiments and machine learning for product innovation: A systematic literature review**. **Quality and Reliability Engineering International.** John Wiley and Sons Ltd, 1 mar. 2022.

ARTRITH, N. et al. Best practices in machine learning for chemistry. **Nature chemistry**, v. 13, n. 6, p. 505–508, 2021.

ATANASOV, A. G. et al. Natural products in drug discovery: advances and opportunities. **Nature reviews Drug discovery**, v. 20, n. 3, p. 200–216, 2021.

AZCARATE, S. M.; PINTO, L.; GOICOECHEA, H. C. **Applications of mixture experiments for response surface methodology implementation in analytical methods development**. **Journal of Chemometrics.** John Wiley and Sons Ltd, dez. 2020.

BACH, E.; SCHYMANSKI, E. L.; ROUSU, J. Joint structural annotation of small molecules using liquid chromatography retention order and tandem mass spectrometry data. **Nature Machine Intelligence**, v. 4, n. 12, p. 1224–1237, 1 dez. 2022.

BAI, X. et al. **Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments**. **Pattern Recognition.** Elsevier Ltd, dez. 2021.

BAJORATH, J. Integration of virtual and high-throughput screening. **Nature Reviews Drug Discovery**, v. 1, n. 11, p. 882–894, 2002.

BALABAN, A. T. Chemical graphs. **Theoretica Chimica Acta**, v. 53, n. 4, p. 355–375, 1979.

BALENTIC, J. P. et al. Cocoa shell: A by-product with great potential for wide application. **Molecules**, v. 23, n.6, 2018.

BARRIL, X. et al. How accurate can molecular dynamics/linear response and Poisson-Boltzmann/solvent accessible surface calculations be for predicting relative binding affinities? Acetylcholinesterase huprine inhibitors as a test case. **Theoretical Chemistry Accounts**, v. 106, n. 1–2, p. 2–9, jun. 2001.

BART-PLANGE, A.; BARYEH, E. A. The physical properties of Category B cocoa beans. **Journal of Food Engineering**, v. 60, n. 3, p. 219–227, dez. 2003.

BATTISTELLA, C.; NONINO, F. Open innovation web-based platforms: The impact of different forms of motivation on collaboration. **Innovation: Management, Policy and Practice**, v. 14, n. 4, p. 557–575, 2012.

BAYADA, D. M.; HAMERSMA, H.; VAN GEERESTEIN, V. J. Molecular diversity and representativity in chemical databases. **Journal of chemical information and computer sciences**, v. 39, n. 1, p. 1–10, 1999.

BELWAL, T. et al. Bioactive Compounds from Cocoa Husk: Extraction, Analysis and Applications in Food Production Chain. **Foods**, v.11, n.6, p.798, mar. 2022.

BENDER, A. et al. “Bayes affinity fingerprints” improve retrieval rates in virtual screening and define orthogonal bioactivity space: when are multitarget drugs a feasible concept? **Journal of chemical information and modeling**, v. 46, n. 6, p. 2445–2456, 2006.

BERRY, N. M. et al. Impact of cocoa flavanol consumption on blood pressure responsiveness to exercise. **British Journal of Nutrition**, v. 103, n. 10, p. 1480–1484, 2010.

BHATT, D. et al. **CNN variants for computer vision: History, architecture, application, challenges and future scope**. **Electronics**, v.10, n. 20, p. 2470, out. 2021.

BOLL, M. et al. Microbial degradation of phthalates: biochemistry and environmental implications. **Environmental Microbiology Reports**, v. 12, n. 1, p. 3-15, fev. 2020.

BOOZARI, M.; HOSSEINZADEH, H. Natural products for COVID-19 prevention and treatment regarding to previous coronavirus infections and novel studies. **Phytotherapy Research**, v. 35, n. 2, p. 864–876, 2021.

BRUNETTO, M. DEL R. et al. The effect of fermentation and roasting on free amino acids profile in Criollo cocoa (Theobroma cacao L.) grown in Venezuela. **Brazilian Journal of Food Technology**, v. 23, 2020.

BÜHLMANN, P.; YU, B. Discussion of "Additive logistic regression: A statistical view," by J. Friedman, T. Hastie and R. Tibshirani. **Ann. Statist**, v. 28, p. 377–386, 2000.

BUITRAGO-LOPEZ, A. et al. Chocolate consumption and cardiometabolic disorders: Systematic review and meta-analysis. **BMJ (Online)**, v. 343, n. 7825, 1 out. 2011.

CÁDIZ-GURREA, M. L. et al. Isolation, comprehensive characterization and antioxidant activities of Theobroma cacao extract. **Journal of Functional Foods**, v. 10, p. 485–498, 2014.

CAMANDOLA, S.; PLICK, N.; MATTSON, M. P. Impact of Coffee and Cacao Purine Metabolites on Neuroplasticity and Neurodegenerative Disease. **Neurochemical Research**, v. 44, n. 1, p. 214–227, 15 jan. 2019.

CAMU, N. et al. Influence of turning and environmental contamination on the dynamics of populations of lactic acid and acetic acid bacteria involved in spontaneous cocoa bean heap fermentation in Ghana. **Applied and Environmental Microbiology**, v. 74, n. 1, p. 86–98, jan. 2008.

CASTRO-ALAYO, E. M. et al. Formation of aromatic compounds precursors during fermentation of Criollo and Forastero cocoa. **Heliyon**, v. 5, n. 1, 2019.

ÇELIK, Ö.; ALTUNAYDIN, S. S. A Research on Machine Learning Methods and Its Applications. **Journal of Educational Technology & Online Learning**, v. 1, n. 3, p. 25–40, 2018.

CHAN, H. C. S. et al. Advancing Drug Discovery via Artificial Intelligence. **Trends in Pharmacological** Sciences, v. 40, n. 8, p. 592-604, ago. 2019.

CHAPMAN, A. G.; ATKINSON, D. E. Adenine nucleotide concentrations and turnover rates. Their correlation with biological activity in bacteria and yeast. **Advances in microbial physiology**, v. 15, p. 253–306, 1977.

CHASALOW, S. D.; BRAND, R. J. Algorithm AS 299: Generation of Simplex Lattice Points. **Journal of the Royal Statistical Society. Series C (Applied Statistics)**, v. 44, n. 4, p. 534-545, 1995.

CHEN, D. et al. **GREEN TEA AND TEA POLYPHENOLS IN CANCER PREVENTIONFrontiers in Bioscience**. [s.l: s.n.].

CHOI, J. et al. Exploring the chemical space of protein–protein interaction inhibitors through machine learning. **Scientific Reports**, v. 11, n. 1, dez. 2021.

CHOURASIA, M. et al. EGCG, a green tea catechin, as a potential therapeutic agent for symptomatic and asymptomatic SARS-CoV-2 infection. **Molecules**, v. 26, n. 5, p. 1200, 2021.

CIEPLINSKI, T. et al. We should at least be able to design molecules that dock well. **arXiv preprint arXiv:2006.16955**, 2020.

CIMINI, A. et al. Cocoa powder triggers neuroprotective and preventive effects in a human Alzheimer's disease model by modulating BDNF signaling pathway. **Journal of Cellular Biochemistry**, v. 114, n. 10, p. 2209–2220, out. 2013.

COOPER, K. A. et al. Cocoa and health: A decade of research. **British Journal of Nutrition**, v. 99, n. 1, p. 1-11, jan. 2008.

COQ-HUELVA, D.; TORRES-NAVARRETE, B.; BUENO-SUÁREZ, C. Indigenous worldviews and Western conventions: Sumak Kawsay and cocoa production in Ecuadorian Amazonia. **Agriculture and Human Values**, v. 35, n. 1, p. 163–179, 1 mar. 2018.

CORLEY, D. G.; DURLEY, R. C. Strategies for database dereplication of natural products. **Journal of natural products**, v. 57, n. 11, p. 1484–1490, 1994.

CORTEZ, D. et al. Changes in bioactive compounds during fermentation of cocoa (Theobroma cacao) harvested in Amazonas-Peru. **Current Research in Food Science**, v. 6, p.1000494, jan. 2023.

COSTANZO, M. J. et al. Potent, small-molecule inhibitors of human mast cell tryptase. Antiasthmatic action of a dipeptide-based transition-state analogue containing a benzothiazole ketone. **Journal of medicinal chemistry**, v. 46, n. 18, p. 3865–3876, 2003a.

COVA, T. F. G. G.; PAIS, A. A. C. C. Deep Learning for Deep Chemistry: Optimizing the Prediction of Chemical Patterns. **Frontiers in Chemistry**, v. 7, p.809, nov. 2019.

CRAGG, G. M.; NEWMAN, D. J. Natural products: a continuing source of novel drug leads. **Biochimica et Biophysica Acta (BBA)-General Subjects**, v. 1830, n. 6, p. 3670–3695, 2013.

CROTEAU, R. et al. Natural products (secondary metabolites). **Biochemistry and molecular biology of plants**, v. 24, p. 1250–1319, 2000.

CURTIS, M. J. et al. Planning experiments: Updated guidance on experimental design and analysis and their reporting III. **British Journal of Pharmacology**, v. 179, n. 15, p. 3907-3913, ago. 2022.

DA SILVEIRA, N. J. F. et al. Web services for molecular docking simulations. Em: **Docking Screens for Drug Discovery**, p. 221-229, 2019.

DAS, S.; DEY, A.; ROY, N. Applications of Artificial Intelligence in Machine Learning: Review and Prospect. **International Journal of Computer Applications**, v. 115, n. 9, 2015.

DASIMAN, R. et al. A Review of Procyanidin: Updates on current bioactivities and potential health benefits. **Biointerface Res. Appl. Chem**, v. 12, n. 5, p. 5918–5940, 2022.

DAYAN, P.; NIV, Y. Reinforcement learning: The Good, The Bad and The Ugly. **Current Opinion in Neurobiology**, v. 18, n. 2, p. 185-196, abr. 2008.

DE BRITO, E. S. et al. Structural and chemical changes in cocoa (Theobroma cacao L) during fermentation, drying and roasting. **Journal of the Science of Food and Agriculture**, v. 81, n. 2, p. 281–288, 2001.

DE QUEIROZ, L. N. et al. New substances of Equisetum hyemale L. extracts and their in vivo antitumoral effect against oral squamous cell carcinoma. **Journal of Ethnopharmacology**, v.303, p. 116043, 2022.

DE VUYST, L.; LEROY, F. Functional role of yeasts, lactic acid bacteria and acetic acid bacteria in cocoa fermentation processes. **FEMS Microbiology Reviews**, v. 44, n. 4, p. 432–453, 2020.

DE VUYST, L.; WECKX, S. The cocoa bean fermentation process: from ecosystem analysis to starter culture development. **Journal of Applied Microbiology**, v. 121, n. 1, p. 5–17, 2016a.

DEL CAMPO, M.; CARLSON, A.; MANNINGER, S. Towards Hallucinating Machines - Designing with Computational Vision. **International Journal of Architectural Computing**, v. 19, n. 1, p. 88–103, mar. 2021.

DELGADO-OSPINA, J. et al. The role of fungi in the cocoa production chain and the challenge of climate change. **Journal of Fungi**, v. 7, n. 3, p. 202, 2021.

DEMAIN, A. L. Importance of microbial natural products and the need to revitalize their discovery. **Journal of Industrial Microbiology and Biotechnology**, v. 41, n. 2, p. 185–201, 2014.

DENG, J. et al. Imagenet: A large-scale hierarchical image database. In: **2009 IEEE conference on computer vision and pattern recognition**. Ieee, 2009, p. 248-255.

DEUS, V. L. et al. Understanding amino acids and bioactive amines changes during on-farm cocoa fermentation. **Journal of Food Composition and Analysis**, v. 97, p. 103776, abr. 2021.

DIFRANCISCO-DONOGHUE, J. et al. Effects of Tyrosine on Parkinson's Disease: A Randomized, Double-Blind, Placebo-Controlled Trial. **Movement Disorders Clinical Practice**, v. 1, n. 4, p. 348–353, dez. 2014.

DMITRYJUK, M.; ŁOPIEŃSKA-BIERNAT, E.; FARJAN, M. The level of sugars and synthesis of trehalose in Ascaris suum tissues. **Journal of Helminthology**, v. 83, n. 3, p. 237–243, 2009.

DONG, X. et al. Web service infrastructure for chemoinformatics. **Journal of chemical information and modeling**, v. 47, n. 4, p. 1303–1307, 2007.

DORFMAN, L. J.; JARVIK, M. E. Comparative stimulant and diuretic actions of caffeine and theobromine in man. **Clinical Pharmacology & Therapeutics**, v. 11, n. 6, p. 869–872, 1970.

DREW, K. L. M. et al. Size estimation of chemical space: how big is it? **Journal of Pharmacy and Pharmacology**, v. 64, n. 4, p. 490–495, 2012.

DRICHE, E. H. et al. A new Streptomyces strain isolated from Saharan soil produces di-(2-ethylhexyl) phthalate, a metabolite active against methicillin-resistant Staphylococcus aureus. **Annals of microbiology**, v. 65, n. 3, p. 1341–1350, 2015.

DU, X. et al. Insights into protein–ligand interactions: Mechanisms, models, and methods. **International Journal of Molecular Sciences**, v. 17, n. 2, p. 144, 2016.

DUCA, D. et al. Indole-3-acetic acid in plant–microbe interactions. **Antonie Van Leeuwenhoek**, v. 106, n. 1, p. 85–125, 2014.

DUCA, D. R.; GLICK, B. R. Indole-3-acetic acid biosynthesis and its regulation in plant-associated bacteria. **Applied microbiology and biotechnology**, v. 104, p. 8607–8619, 2020.

DÜHRKOP, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. **Nature methods**, v. 16, n. 4, p. 299–302, 2019.

DUKE, J. A. Biological activity summary for cocoa (Theobroma cacao L.). **Journal of Medicinal Food**, v. 3, n. 2, p. 115–119, 2000.

DZOBO, K. The Role of Natural Products as Sources of Therapeutic Agents for Innovative Drug Discovery. **Comprehensive Pharmacology**, p. 408, 2022.

EALES, J. et al. Human health impacts of exposure to phthalate plasticizers: An overview of reviews. **Environment** International, v. 158, p. 106903, jan. 2022.

EKINS, S.; CLARK, A. M.; WILLIAMS, A. J. Open drug discovery teams: A chemistry mobile app for collaboration. **Molecular Informatics**, v. 31, n. 8, p. 585–597, ago. 2012.

ELHARROUSS, O. et al. Backbones-Review: Feature Extraction Networks for Deep Learning and Deep Reinforcement Learning Approaches. **arXiv preprint arXiv:2206.08016**, jun. 2022.

ELLAM, S.; WILLIAMSON, G. Cocoa and human health. **Annual Review of Nutrition**, v. 33, p. 105–128, jul. 2013.

ERTL, P.; ROHDE, B.; SELZER, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. **Journal of medicinal chemistry**, v. 43, n. 20, p. 3714–3717, 2000.

FEBRIANTO, N. A.; ZHU, F. Composition of methylxanthines, polyphenols, key odorant volatiles and minerals in 22 cocoa beans obtained from different geographic origins. **LWT**, v. 153, p. 112395, jan. 2022.

FEDURAEV, P. et al. Phenylalanine and tyrosine as exogenous precursors of wheat (triticum aestivum L.) secondary metabolism through PAL-associated pathways. **Plants**, v. 9, n. 4, p. 476, abr. 2020.

FELDMANN, C. et al. Identifying promiscuous compounds with activity against different target classes. **Molecules**, v. 24, n. 22, p. 4185, 2019.

FERNSTROM, J. D.; FERNSTROM, M. H. Tyrosine, Phenylalanine, and Catecholamine Synthesis and Function in the Brain. **The Journal of Nutrition**, v. 137, n. 6, p. 1539S-1547S, 2007.

FIGUEROA-HERNÁNDEZ, C. et al. The challenges and perspectives of the selection of starter cultures for fermented cocoa beans. **International Journal of Food Microbiology**, v. 301, p. 41–50, 2019.

FRANÇOIS, J.; PARROU, J. L. Reserve carbohydrates metabolism in the yeast Saccharomyces cerevisiae. **Fems microbiology reviews**, v. 25, n. 1, p. 125–145, 2001.

FRANZEN, M.; BORGERHOFF MULDER, M. Ecological, economic and social perspectives on cocoa production worldwide. **Biodiversity and Conservation**, v. 16, p. 3835-3849, dez. 2007.

FREDHOLM, B. B.; SMIT, H. J. Theobromine and the pharmacology of cocoa. **Methylxanthines**, p. 201–234, 2011.

FREIESLEBEN, J.; KEIM, J.; GRUTSCH, M. Machine learning and Design of Experiments: Alternative approaches or complementary methodologies for quality improvement? **Quality and Reliability Engineering International**, v. 36, n. 6, p. 1837–1848, out. 2020.

FURUSHIMA, D. et al. Prevention of acute upper respiratory infections by consumption of catechins in healthcare workers: A randomized, placebo-controlled trial. **Nutrients**, v. 12, n. 1, p. 4, jan. 2019.

FUSAR-POLI, L. et al. The effect of cocoa-rich products on depression, anxiety, and mood: A systematic review and meta-analysis. **Critical Reviews in Food Science and Nutrition**, v. 62, n. 28, p. 7905-7916, 2022.

GALLEGO, A. M. et al. Transcriptomic analyses of cacao flavonoids produced in photobioreactors. **BMC genomics**, v. 22, n. 1, p. 1–18, 2021.

GARCÍA-ORTEGÓN, M. et al. DOCKSTRING: easy molecular docking yields better benchmarks for ligand design. **Journal of Chemical Information and Modeling**, v. 62, n. 15, p. 3486-3502, ago. 2022.

GARG, A. K. et al. Trehalose accumulation in rice plants confers high tolerance levels to different abiotic stresses. **Proceedings of the National Academy of Sciences**, v. 99, n. 25, p. 15898–15903, 2002.

GARIBOTTO, G. et al. The metabolic conversion of phenylalanine into tyrosine in the human kidney: Does it have nutritional implications in renal patients? **Journal of Renal Nutrition**, v. 12, n. 1, p. 8–16, jan. 2002.

GAUDÊNCIO, S. P. et al. Advanced Methods for Natural Products Discovery: Bioactivity Screening, Dereplication, Metabolomics Profiling, Genomic Sequencing,

Databases and Informatic Tools, and Structure Elucidation. **Marine Drug**, v.21, n. 5, p. 308, 2023.

GAWEHN, E. et al. Advancing drug discovery via GPU-based deep learning. **Expert Opinion on Drug Discovery**, v. 13, n. 7, p. 579-582, jul. 2018.

GENHEDEN, S.; RYDE, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. **Expert Opinion on Drug Discovery** v. 10, n. 5, p. 449-461, 2015.

GHAYUR, M. N.; KHAN, H.; GILANI, A. H. Antispasmodic, bronchodilator and vasodilator activities of (+)-catechin, a naturally occurring flavonoid. **Archives of pharmacal research**, v. 30, p. 970-975, 2007.

GHIANDONI, G. M.; CALDEWEYHER, E. Fast calculation of hydrogen-bond strengths and free energy of hydration of small molecules. **Scientific Reports**, v. 13, n. 1, dez. 2023.

GHOSH, D. **A cinnamon-derived procyanidin type-A compound: A potential candidate molecule against coronaviruses including COVID-19**. **Journal of Ayurveda Case Reports**, v. 3, n. 4, p. 122-126, 2020.

GOETZ, M. et al. Extremely randomized trees based brain tumor segmentation. **Proceeding of BRATS challenge-MICCAI**, v. 14, p. 6–11, 2014.

GOHLKE, H.; KLEBE, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. **Angewandte Chemie International Edition**, v. 41, n. 15, p. 2644–2676, 2002.

GONZÁLEZ, M. P. et al. BCUT descriptors to predicting affinity toward A3 adenosine receptors. **Bioorganic & medicinal chemistry letters**, v. 15, n. 15, p. 3491–3495, 2005.

GORMAN, J. W.; HINMAN, J. E. Simplex lattice designs for multicomponent systems. **Technometrics**, v. 4, n. 4, p. 463–487, 1962.

GRASSI, D. et al. Cocoa reduces blood pressure and insulin resistance and improves endothelium-dependent vasodilation in hypertensives. **Hypertension**, v. 46, n. 2, p. 398–405, ago. 2005.

GROMSKI, P. S. et al. How to explore chemical space using algorithms and automation. **Nature Reviews Chemistry**, v. 3, n. 2, p. 119–128, 2019.

GU, J. et al. Recent advances in convolutional neural networks. **Pattern recognition**, v. 77, p. 354–377, 2018.

GUÀRDIA, E. et al. A molecular dynamics simulation study of hydrogen bonding in aqueous ionic solutions. **Journal of Molecular Liquids**, v. 117, n. 1-3, p. 63-67, 2005.

GUEHI, T. S. et al. Performance of different drying methods and their effects on the chemical quality attributes of raw cocoa material. **International Journal of Food Science and Technology**, v. 45, n. 8, p. 1564–1571, ago. 2010.

GUPTA, R. et al. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. **Molecular Diversity**, v. 25, n. 3, p. 1315–1360, 2021.

GUTIÉRREZ-RÍOS, H. G. et al. Yeasts as Producers of Flavor Precursors during Cocoa Bean Fermentation and Their Relevance as Starter Cultures: A Review. **Fermentation**, v. 8, n. 7, p. 331, 2022.

HABIB, M. R.; KARIM, M. R.; OTHERS. Antitumour evaluation of di-(2-ethylhexyl) phthalate (DEHP) isolated from Calotropis gigantea L. flower. **Acta Pharm**, v. 62, n. 4, p. 607–615, 2012.

HALL, T. et al. Structure of human ADAM-8 catalytic domain complexed with batimastat. **Acta Crystallographica Section F: Structural Biology and Crystallization Communications**, v. 68, n. 6, p. 616–621, 2012a.

HASTINGS, J. et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. **Nucleic acids research**, v. 41, n. D1, p. D456–D463, 2012.

HAUG, K. et al. MetaboLights: a resource evolving in response to the needs of its scientific community. **Nucleic acids research**, v. 48, n. D1, p. D440–D444, 2020.

HE, X. et al. Trehalose Alleviates Crystalline Silica-Induced Pulmonary Fibrosis via Activation of the TFEB-Mediated Autophagy-Lysosomal System in Alveolar Macrophages. **Cells**, v. 9, n. 1, jan. 2020.

HEINRICH, M. Ethnobotany and natural products: the search for new molecules, new treatments of old diseases or a better understanding of indigenous cultures? **Current Topics in Medicinal Chemistry**, v. 3, n. 2, p. 141–154, 2003.

HELGUERA, A. M. et al. Applications of 2D descriptors in drug design: a DRAGON tale. **Current topics in medicinal chemistry**, v. 8, n. 18, p. 1628–1655, 2008.

HENSS, L. et al. The green tea catechin epigallocatechin gallate inhibits SARS-CoV-2 infection. **The Journal of general virology**, v. 102, n. 4, 2021.

HESHMATISAFA, S.; SEPPÄNEN, M. Exploring API-driven business models: Lessons learned from Amadeus's digital transformation. **Digital Business**, v. 3, n. 1, p. 100055, jun. 2023.

HILBE, J. M. STATISTICA 7: an overview. **The American Statistician**, v. 61, n. 1, p. 91–94, 2007.

HILLMAN, E. T.; READNOUR, L. R.; SOLOMON, K. V. **Exploiting the natural product potential of fungi with integrated-omics and synthetic biology approaches**. **Current Opinion in Systems Biology**, v. 5, p. 50-56, 2017.

HOANG, V. L. T.; LI, Y.; KIM, S.-K. Cathepsin B inhibitory activities of phthalates isolated from a marine Pseudomonas strain. **Bioorganic & medicinal chemistry letters**, v. 18, n. 6, p. 2083–2088, 2008.

HOLLINGSWORTH, S. A.; DROR, R. O. **Molecular Dynamics Simulation for All**. **Neuron**, v. 99, n. 6, p. 1129-1143, 2018.

HOLTEN-ANDERSEN, L. et al. Combination of the cationic surfactant dimethyl dioctadecyl ammonium bromide and synthetic mycobacterial cord factor as an efficient adjuvant for tuberculosis subunit vaccines. **Infection and immunity**, v. 72, n. 3, p. 1608–1617, 2004.

HORWOOD, J.; NOUTAHI, E. Molecular Design in Synthetically Accessible Chemical Space via Deep Reinforcement Learning. **ACS Omega**, v. 5, n. 51, p. 32984–32994, dez. 2020.

HOU, T. et al. Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. **Journal of Chemical Information and Modeling**, v. 51, n. 1, p. 69–82, jan. 2011.

HU, Y.; BAJORATH, J. Compound promiscuity: What can we learn from current data? **Drug Discovery Today**, v. 18, n. 13-14, p. 644-650, 2013.

HUANG, B.; ZHANG, Y. Teaching an old dog new tricks: Drug discovery by repositioning natural products and their derivatives. **Drug Discovery Today**, v. 27, n. 7, p. 1936-1944, 2022.

HUANG, L. et al. Phthalic acid esters: Natural sources and biological activities. **Toxins**, v. 13, n. 7, p. 495, 2021.

HUANG, M.; LU, J.-J.; DING, J. Natural products in cancer therapy: past, present and future. **Natural Products and Bioprospecting**, v. 11, n. 1, p. 5–13, 2021.

HUBBARD, R. E.; KAMRAN HAIDER, M. Hydrogen Bonds in Proteins: Role and Strength. Em: **eLS**. [s.l.] Wiley, 2010.

HUBER, F. et al. matchms-processing and similarity evaluation of mass spectrometry data. **bioRxiv**, p. 2020.08. 06.239244, 2020.

HUEY, R.; MORRIS, G. M.; FORLI, S. Using AutoDock 4 and AutoDock vina with AutoDockTools: a tutorial. **The Scripps Research Institute Molecular Graphics Laboratory**, v. 10550, p. 92037, 2012.

HUNTER, S. V. Analysing and representing narrative data: The long and winding road. **Current narratives**, v. 1, n. 2, p. 44–54, 2010.

IGAWA, T. K.; DE TOLEDO, P. M.; ANJOS, L. J. S. Climate change could reduce and spatially reconfigure cocoa cultivation in the Brazilian Amazon by 2050. **PLoS ONE**, v. 17, n. 1 January, jan. 2022.

ITURRIAGA, G.; SUÁREZ, R.; NOVA-FRANCO, B. Trehalose metabolism: From osmoprotection to signaling. **International Journal of Molecular Sciences**, v. 10, n. 9, p. 3793-3810, 2009.

JAIN, R. S. et al. Review on methylxanthine, theobromine and theophylline. **Asian Journal of Pharmaceutical Analysis**, v. 10, n. 3, p. 173–174, 2020.

JIANG, Y. et al. Procyanidin B2 Suppresses Lipopolysaccharides-Induced Inflammation and Apoptosis in Human Type II Alveolar Epithelial Cells and Lung Fibroblasts. **Journal of Interferon and Cytokine Research**, v. 40, n. 1, p. 54–63, 1 jan. 2020.

JOHN, W. A. et al. Experimentally modelling cocoa bean fermentation reveals key factors and their influences. **Food Chemistry**, v. 302, n. July 2019, p. 125335, 2020.

JORGENSEN, W. L. The Many Roles of Computation in Drug Discovery. **Science**, v. 303, n. 5665, p. 1813-1818, 2004.

JOUANEH, T. M. M. et al. Incorporating LC-MS/MS Analysis and the Dereplication of Natural Product Samples into an Upper-Division Undergraduate Laboratory Course. **Journal of Chemical Education**, v. 99, n. 7, p. 2636–2642, jul. 2022.

KADOW, D. et al. Fermentation-like incubation of cocoa seeds (Theobroma cacao L.) - Reconstruction and guidance of the fermentation process. **LWT**, v. 62, n. 1, p. 357–361, 1 jun. 2015.

KANG, Y. et al. Natural language processing (NLP) in management research: A literature review. **Journal of Management Analytics**, v. 7, n. 2, p. 139-172, 2020.

KANWAL et al. Indole-3-acetamides: As Potential Antihyperglycemic and Antioxidant Agents; Synthesis, in Vitro α-Amylase Inhibitory Activity, Structure-Activity Relationship, and in Silico Studies. **ACS Omega**, v. 6, n. 3, p. 2264–2275, jan. 2021.

KATZ, D. L.; DOUGHTY, K.; ALI, A. Cocoa and chocolate in human health and disease. **Antioxidants and Redox Signaling**, v. 15, n. 10, p. 2779-2811, 2011.

KE, G. et al. Lightgbm: A highly efficient gradient boosting decision tree. **Advances in neural information processing systems**, v. 30, 2017.

KEITH, J. A. et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. **Chemical reviews**, v. 121, n. 16, p. 9816–9872, 2021.

KHAN, N. et al. Cocoa polyphenols and inflammatory markers of cardiovascular disease. **Nutrients**, v. 6, n. 2, p. 844–880, 2014.

KHANFAR, M. A.; TAHA, M. O. Elaborate ligand-based modeling coupled with multiple linear regression and k nearest neighbor QSAR analyses unveiled new nanomolar mTOR inhibitors. **Journal of chemical information and modeling**, v. 53, n. 10, p. 2587–2612, 2013.

KIEFER, B. A. et al. The identification of adenine in cacao products. **Journal of Liquid Chromatography**, v. 6, n. 5, p. 927–930, abr. 1983.

KIND, T.; FIEHN, O. Strategies for dereplication of natural compounds using high-resolution tandem mass spectrometry. **Phytochemistry Letters**, v. 21, p. 313-319, set. 2017.

KITCHEN, D. B. et al. Docking and scoring in virtual screening for drug discovery: methods and applications. **Nature reviews Drug discovery**, v. 3, n. 11, p. 935–949, 2004.

KJELDAHL, K.; BRO, R. Some common misunderstandings in chemometrics. **Journal of Chemometrics**, v. 24, n. 7–8, p. 558–564, 2010.

KOEHN, F. E.; CARTER, G. T. The evolving role of natural products in drug discovery. **Nature reviews Drug discovery**, v. 4, n. 3, p. 206–220, 2005.

KONGOR, J. E. et al. Factors influencing quality variation in cocoa (Theobroma cacao) bean flavour profile - A review. **Food Research International**, v. 82, p. 42-52, 2016.

KOULOURIDI, E. et al. A primer on natural product-based virtual screening. **Physical Sciences Reviews**, v. 4, n. 6, 2019.

KOYAMA, Y. et al. Metabolism of purine bases, nucleosides and alkaloids in theobromine-forming Theobroma cacao leaves. **Plant Physiology and Biochemistry**, v. 41, n. 11–12, p. 977–984, 2003.

KRZYWINSKI, M.; ALTMAN, N. Classification and regression trees. **Nature Methods**, v. 14, n. 8, p. 757–758, 2017.

KULKARNI, G. B. et al. Indole-3-acetic acid biosynthesis in Fusarium delphinoides strain GPK, a causal agent of Wilt in Chickpea. **Applied biochemistry and biotechnology**, v. 169, n. 4, p. 1292–1305, 2013.

KULKARNI VISHAKHA, S.; BUTTE KISHOR, D.; RATHOD SUDHA, S. Natural polymers–A comprehensive review. **International journal of research in pharmaceutical and biomedical sciences**, v. 3, n. 4, p. 1597–1613, 2012.

KUMAR, S. et al. Discovery of New Hydroxyethylamine Analogs against 3CLproProtein Target of SARS-CoV-2: Molecular Docking, Molecular Dynamics Simulation, and Structure-Activity Relationship Studies. **Journal of Chemical Information and Modeling**, v. 60, n. 12, p. 5754–5770, dez. 2020.

KUMARI, R.; KUMAR, R.; LYNN, A. G-mmpbsa -A GROMACS tool for high-throughput MM-PBSA calculations. **Journal of Chemical Information and Modeling**, v. 54, n. 7, p. 1951–1962, jul. 2014.

KUNTZ, I. D. et al. A geometric approach to macromolecule-ligand interactions. **Journal of molecular biology**, v. 161, n. 2, p. 269–288, 1982.

LABUTE, P. A widely applicable set of descriptors. **Journal of Molecular Graphics and Modelling**, v. 18, n. 4–5, p. 464–477, 2000.

LAMBRAKIS, D. P. Experiments with mixtures: A generalization of the simplex-lattice design. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 30, n. 1, p. 123–136, 1968.

LAMBROT, R. et al. Phthalates impair germ cell development in the human fetal testis in vitro without change in testosterone production. **Environmental health perspectives**, v. 117, n. 1, p. 32–37, 2009.

LANDRUM, G.; OTHERS. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. **Greg Landrum**, v. 8, p. 31, 2013.

LANGLYKKE, A. CRC Handbook of antibiotic compound (IV), Edited by Bardy J et al. **CRC, Boca Raton, FL**, 1980.

LEACH, A. R.; SHOICHET, B. K.; PEISHOFF, C. E. Prediction of protein-ligand interactions. Docking and scoring: Successes and gaps. **Journal of Medicinal Chemistry**, v. 49, n. 20, p. 5851-5855, 2006..

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. **Nature**, v. 521, n. 7553, p. 436-444, 2015.

LEE, H. W.; CHOI, I. W.; HA, S. K. Immunostimulatory Activities of Theobromine on Macrophages via the Activation of MAPK and NF-κB Signaling Pathways. **Current Issues in Molecular Biology**, v. 44, n. 9, p. 4216–4228, set. 2022.

LI, C. et al. Blocking the 4-1BB pathway ameliorates crystalline silica-induced lung inflammation and fibrosis in mice. **Theranostics**, v. 6, n. 12, p. 2052–2067, 2016.

LI MANNI, G. et al. The OpenMolcas Web: A Community-Driven Approach to Advancing Computational Chemistry. **Journal of Chemical Theory and Computation**, 2023.

LIMA, L. J. R. et al. Theobroma cacao L.,"the food of the gods": quality determinants of commercial cocoa beans, with particular reference to the impact of fermentation. **Critical reviews in food science and nutrition**, v. 51, n. 8, p. 731–761, 2011.

LIN, L.; XU, X. Indole-3-acetic acid production by endophytic Streptomyces sp. En-1 isolated from medicinal plants. **Current Microbiology**, v. 67, n. 2, p. 209–217, 2013.

LIU, Z. et al. DeepScreening: A deep learning-based screening web server for accelerating drug discovery. **Database**, v. 2019, n. 1, 2019.

LOTFY, W. A. et al. Production of di-(2-ethylhexyl) phthalate by Bacillus subtilis AD35: Isolation, purification, characterization and biological activities. **Microbial Pathogenesis**, v. 124, p. 89–100, nov. 2018.

LOWE, D. Chemical space is big. Really big. **MedChemComm**, v. 6, n. 1, p. 12, 2015.

LUNN, J. E. et al. Trehalose metabolism in plants. **The Plant Journal**, v. 79, n. 4, p. 544–567, 2014.

MACHONIS, P. R. et al. Method for the determination of catechin and epicatechin enantiomers in cocoa-Based ingredients and products by High-Performance Liquid chromatography: Single- Laboratory validation. **Journal of AOAC International**, v. 95, n. 2, p. 500–507, mar. 2012.

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. **Journal of Biomolecular Structure and Dynamics**, v. 40, n. 2, p. 696–711, 2022.

MALLMANN, L. P.; DE OLIVEIRA RIOS, A.; RODRIGUES, E. MS-FINDER and SIRIUS for phenolic compound identification from high-resolution mass spectrometry data. **Food Research International**, v. 163, p. 112315, 2023.

MARKOVIC, S.; GUTMAN, I. Spectral moments of the edge adjacency matrix in molecular graphs. Benzenoid hydrocarbons. **J. Chem. Inf. Comput. Sci.**, v. 39, n. 2, p. 289–293, 1999.

MARTÍNEZ-PINILLA, E.; OÑATIBIA-ASTIBIA, A.; FRANCO, R. The relevance of theobromine for the beneficial effects of cocoa consumption. **Frontiers in Pharmacology**, v. 6, p. 30, 2015.

MARTINON, D. et al. Potential fast COVID-19 containment with trehalose. **Frontiers in Immunology**, v. 11, p. 1623, 2020.

MATISSEK, R. Evaluation of xanthine derivatives in chocolate–nutritional and chemical aspects. **Zeitschrift für Lebensmitteluntersuchung und-Forschung A**, v. 205, p. 175–184, 1997.

MAUN, H. R. et al. Bivalent antibody pliers inhibit β-tryptase by an allosteric mechanism dependent on the IgG hinge. **Nature communications**, v. 11, n. 1, p. 1–12, 2020.

MCLAUGHLIN, C. A. et al. Regression of line-10 hepatocellular carcinomas following treatment with water-soluble, microbial extracts combined with trehalose or arabinose mycolates. **Cancer Immunology, Immunotherapy**, v. 4, n. 1, p. 61–68, 1978.

MEDEMA, M. H.; FISCHBACH, M. A. Computational approaches to natural product discovery. Nature Chemical Biology. **Nature**, v. 11, n. 9, p. 639-648, 2015.

MEUWLY, M. Machine learning for chemical reactions. **Chemical Reviews**, v. 121, n. 16, p. 10218-10239, 2021.

MIN, S.; LEE, B.; YOON, S. Deep learning in bioinformatics. **Briefings in bioinformatics**, v. 18, n. 5, p. 851-869, 2017.

MISHRA, C. B. et al. Identifying the natural polyphenol catechin as a multi-targeted agent against SARS-CoV-2 for the plausible therapy of COVID-19: an integrated computational approach. **Briefings in Bioinformatics**, v. 22, n. 2, p. 1346–1360, 2021.

MOHIMANI, H. et al. Dereplication of peptidic natural products through database search of mass spectra. **Nature chemical biology**, v. 13, n. 1, p. 30–37, 2017.

MONTGOMERY, D. C. The Use of Statistical Process Control and Design of Experiments in Product and Process Improvement. **IIE Transactions (Institute of Industrial Engineers)**, v. 24, n. 5, p. 4–17, 1992.

MOREIRA, I. M. DA V. et al. Microbial succession and the dynamics of metabolites and sugars during the fermentation of three different cocoa (Theobroma cacao L.) hybrids. **Food Research International**, v. 54, n. 1, p. 9–17, nov. 2013.

MORENO-ZAMBRANO, M. et al. A mathematical model of cocoa bean fermentation. **Royal Society Open Science**, v. 5, n. 10, 1 out. 2018.

MORIWAKI, H. et al. Mordred: a molecular descriptor calculator. **Journal of cheminformatics**, v. 10, n. 1, p. 1–14, 2018.

MORRONE XAVIER, M. et al. SAnDReS a computational tool for statistical analysis of docking results and development of scoring functions. **Combinatorial chemistry & high throughput screening**, v. 19, n. 10, p. 801–812, 2016.

MOTAMAYOR, J. C. et al. Cacao domestication I: The origin of the cacao cultivated by the Mayas. **Heredity**, v. 89, n. 5, p. 380–386, nov. 2002.

MOUSHUMI PRIYA, A.; JAYACHANDRAN, S. Induction of apoptosis and cell cycle arrest by Bis (2-ethylhexyl) phthalate produced by marine Bacillus pumilus MB 40. **Chemico-Biological Interactions**, v. 195, n. 2, p. 133–143, jan. 2012.

MOZZI, FERNANDA.; RAYA, R. R.; VIGNOLO, G. M. **Biotechnology of lactic acid bacteria. 2**. Singapore: Wiley-Blackwell, 2015.

MYERS, R. H. et al. Response Surface Methodology: A Retrospective and Literature Survey. **Journal of Quality Technology**, v. 36, n. 1, p. 53-77, 2004.

NABAVI, S. et al. Anti-Oxidative Polyphenolic Compounds of Cocoa. **Current Pharmaceutical Biotechnology**, v. 16, n. 10, p. 891–901, 2015.

NEHLIG, A. The neuroprotective effects of cocoa flavanol and its influence on cognitive performance. **British journal of clinical pharmacology**, v. 75, n. 3, p. 716–727, 2013.

NET, S. et al. Occurrence, fate, behavior and ecotoxicological state of phthalates in different environmental matrices. **Environmental Science and Technology**, v. 49, n. 7, p. 4019-4035, 2015.

NETTLES, J. H. et al. Bridging chemical and biological space:“target fishing” using 2D and 3D molecular descriptors. **Journal of medicinal chemistry**, v. 49, n. 23, p. 6802–6810, 2006.

NEWMAN, D. J.; CRAGG, G. M. Natural products as sources of new drugs over the 30 years from 1981 to 2010. **Journal of natural products**, v. 75, n. 3, p. 311–335, 2012.

NEWMAN, D. J.; CRAGG, G. M. Natural products as sources of new drugs from 1981 to 2014. **Journal of natural products**, v. 79, n. 3, p. 629–661, 2016.

NOBLE, W. S. What is a support vector machine? **Nature biotechnology**, v. 24, n. 12, p. 1565–1567, 2006.

NOLIBE, D. et al. Activation of rat alveolar macrophages and protection against iv injected tumor cells by intratracheal administration of trehalose dimycolate. **Cancer Immunology, Immunotherapy**, v. 23, n. 3, p. 200–206, 1986.

NOSRATI, M. Python: An appropriate language for real world programming. **World Applied Programming**, v. 1, n. 2, p. 110–117, 2011.

NOVAK, J. et al. Can natural products stop the SARS-CoV-2 virus? A docking and molecular dynamics study of a natural product database. **Future Medicinal Chemistry**, v. 13, n. 4, p. 363–378, 1 fev. 2021.

nsb0902-646. [s.d.].

OHTAKE, S.; WANG, Y. J. Trehalose: Current use and future applications. **Journal of Pharmaceutical Sciences**, v. 100, n. 6, p. 2020-2053, 2011.

ONG, S. P. et al. The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. **Computational Materials Science**, v. 97, p. 209–215, fev. 2015.

ONUFRIEV, A. V; CASE, D. A. Generalized Born Implicit Solvent Models for Biomolecules. **Annual review of biophysics**, v. 48, p. 275-296, 2019.

ORTIZ, A.; SANSINENEA, E. Di-2-ethylhexylphthalate may be a natural product, rather than a pollutant. **Journal of Chemistry**, v. 2018, 2018.

OUATTARA, H. G. et al. Molecular identification and pectate lyase production by Bacillus strains involved in cocoa fermentation. **Food Microbiology**, v. 28, n. 1, p. 1–8, fev. 2011.

OZATO, N. et al. Effect of catechins on upper respiratory tract infections in winter: A randomized, placebo-controlled, double-blinded trial. **Nutrients**, v. 14, n. 9, p. 1856, 2022.

ÖZDEMIR, V.; HEKIM, N. Birth of Industry 5.0: Making Sense of Big Data with Artificial Intelligence, “the Internet of Things” and Next-Generation Technology Policy. **OMICS A Journal of Integrative Biology**, v. 22, n. 1, p. 65–76, 2018.

PAGADALA, N. S.; SYED, K.; TUSZYNSKI, J. Software for molecular docking: a review. **Biophysical reviews**, v. 9, n. 2, p. 91–102, 2017.

PAJOKH, M.; POURFRIDONI, M. Proposing a nasal trehalose-induced autophagy approach against SARS-CoV 2. **Health Science Reports**, v. 4, n. 3, 2021.

PAN, G. et al. Decreased serum free testosterone in workers exposed to high levels of di-n-butyl phthalate (DBP) and di-2-ethylhexyl phthalate (DEHP): a cross-sectional study in China. **Environmental health perspectives**, v. 114, n. 11, p. 1643–1648, 2006.

PAPALEXANDRATOU, Z. et al. Hanseniaspora opuntiae, Saccharomyces cerevisiae, Lactobacillus fermentum, and Acetobacter pasteurianus predominate during well-performed Malaysian cocoa bean box fermentations, underlining the importance of these microbial species for a successful cocoa bean fermentation process. **Food Microbiology**, v. 35, n. 2, p. 73–85, set. 2013.

PARANT, M. et al. Enhancement of Nonspecific Immunity to Bacterial Infection by Cord Factor (6, 6′ MTrehalose Dimycolate). **Journal of Infectious Diseases**, v. 135, n. 5, p. 771–777, 1977.

PASSOS, F. M. L.; LOPEZ, A. S.; SILVA, D. O. Aeration and its influence on the microbial sequence in cacao fermentations in Bahia, with emphasis on lactic acid bacteria. **Journal of Food Science**, v. 49, n. 6, p. 1470–1474, 1984.

PAUL, D. et al. Artificial intelligence in drug discovery and development. **Drug Discovery Today**, v. 26, n. 1, p. 80, 2021.

PIEPEL, G. F.; CORNELL, J. A. Designs for Mixture-Amount Experiments. **Journal of Quality Technology**, v. 19, n. 1, p. 11–28, jan. 1987.

PIEPEL, G. F.; CORNELL, J. A. Mixture experiment approaches: examples, discussion, and recommendations. **Journal of Quality Technology**, v. 26, n. 3, p. 177–196, 1994.

PILON, A. C. et al. NuBBEDB: an updated database to uncover chemical and biological information from Brazilian biodiversity. **Scientific Reports**, v. 7, n. 1, p. 1–12, 2017.

POLITIS, S. N. et al. Design of experiments (DoE) in pharmaceutical development. **Drug Development and Industrial Pharmacy**, v. 43, n. 6, p. 889-901, 2017.

PUERARI, C.; MAGALHÃES, K. T.; SCHWAN, R. F. New cocoa pulp-based kefir beverages: Microbiological, chemical composition and sensory analysis. **Food Research International**, v. 48, n. 2, p. 634–640, out. 2012.

QIAN, Z.-J.; KANG, K.-H.; KIM, S.-K. Isolation and antioxidant activity evaluation of two new phthalate derivatives from seahorse, Hippocampus kuda Bleeler. **Biotechnology and Bioprocess Engineering**, v. 17, n. 5, p. 1031–1040, 2012.

QIN, G. F. et al. MS/MS-Based Molecular Networking: An Efficient Approach for Natural Products Dereplication. **Molecules**, v. 28, n. 1, p. 157, 2022.

QIU, J. et al. A survey of machine learning for big data processing. **Eurasip Journal on Advances in Signal Processing**, v. 2016, p. 1-16, 2016.

RAHARDJO, Y. P. et al. Impact of controlled fermentation on the volatile aroma of roasted cocoa. **Brazilian Journal of Food Technology**, v. 25, 2022.

RAJASEKARAN, S.; RAJASEKAR, N.; SIVANANTHAM, A. Therapeutic potential of plant-derived tannins in non-malignant respiratory diseases. **The Journal of Nutritional Biochemistry**, v. 94, p. 108632, 2021.

RICHARDS, A. B. et al. Trehalose: a review of properties, history of use and human tolerance, and results of multiple safety studies. **Food and Chemical Toxicology**, v. 40, n. 7, p. 871-898, 2002.

RIED, K. et al. Effect of cocoa on blood pressure. **Cochrane Database of Systematic Reviews**, n. 8, 2012.

ROMANO, J. D.; TATONETTI, N. P. Informatics and computational methods in natural product drug discovery: A review and perspectives. **Frontiers in Genetics**, v. 10, p. 368, 2019..

ROSSETTI, G. G. et al. Non-covalent SARS-CoV-2 Mpro inhibitors developed from in silico screen hits. **Scientific reports**, v. 12, n. 1, p. 1–9, 2022.

ROTTIERS, H. et al. Dynamics of volatile compounds and flavor precursors during spontaneous fermentation of fine flavor Trinitario cocoa beans. **European Food Research and Technology**, v. 245, p. 1917–1937, 2019.

ROY, R. N. Bioactive natural derivatives of phthalate ester. **Critical reviews in biotechnology**, v. 40, n. 7, p. 913–929, 2020a.

ROY, R. N.; SEN, S. K. Fermentation studies for the production of dibutyl phthalate, an ester bioactive compound from Streptomyces albidoflavus MTCC 3662 using low-priced substrates. **Jordan J Biol Sci**, v. 6, p. 177–181, 2013.

RUDIN, C.; WAGSTAFF, K. L. Machine learning for science and society. Machine Learning. **Machine Learning**, v. 95, p. 1-9, 2014.

RUTZ, A. et al. The LOTUS initiative for open knowledge management in natural products research. **Elife**, v. 11, p. e70780, 2022.

SAITA, N. et al. Trehalose 6, 6′-dimycolate (cord factor) of Mycobacterium tuberculosis induces corneal angiogenesis in rats. **Infection and immunity**, v. 68, n. 10, p. 5991–5997, 2000.

SALDIVAR-GONZALEZ, F. I. et al. Chemical space and diversity of the NuBBE database: a chemoinformatic characterization. **Journal of Chemical Information and Modeling**, v. 59, n. 1, p. 74–85, 2018.

SALMASO, L. et al. Design of experiments and machine learning to improve robustness of predictive maintenance with application to a real case study. **Communications in Statistics: Simulation and Computation**, v. 51, n. 2, p. 570–582, 2022.

SANTANDER MUÑOZ, M. et al. An overview of the physical and biochemical transformation of cocoa seeds to beans and to chocolate: Flavor formation. **Critical Reviews in Food Science and Nutrition,** v. 60, n. 10, p. 1593-1613, 2020.

SARBU, I.; CSUTAK, O. The microbiology of cocoa fermentation. In: **Caffeinated and cocoa based beverages**. Woodhead Publishing, 2019. p. 423-446.

SARKER, I. H. Machine Learning: Algorithms, Real-World Applications and Research Directions. **SN Computer Science**, v. 2, n. 3, p. 160, 2021.

SARUMATHI, S. et al. Statistica software: a state of the art review. **International Journal of Computer and Information Engineering**, v. 9, n. 2, p. 473–480, 2015.

SCHIRALDI, C.; DI LERNIA, I.; DE ROSA, M. Trehalose production: exploiting novel approaches. **TRENDS in Biotechnology**, v. 20, n. 10, p. 420–425, 2002.

SCHMIDT, M.; LIPSON, H. Distilling free-form natural laws from experimental data. **Science**, v. 324, n. 5923, p. 81–85, abr. 2009.

SCHROTH, G.; HARVEY, C. A. Biodiversity conservation in cocoa production landscapes: An overview. **Biodiversity and Conservation**, v. 16, p. 2237-2244, 2007.

SCHWAN, R. F. Cocoa fermentations conducted with a defined microbial cocktail inoculum. **Applied and Environmental Microbiology**, v. 64, n. 4, p. 1477–1483, 1998.

SCHWAN, R. F.; WHEALS, A. E. The microbiology of cocoa fermentation and its role in chocolate quality. **Critical Reviews in Food Science and Nutrition**, v. 44, n. 4, p. 205–221, 2004.

SHI, Y.-F. et al. Machine Learning for Chemistry: Basics and Applications. **Engineering**, jul. 2023.

SIMONS, F. E. R. et al. The bronchodilator effect and pharmacokinetics of theobromine in young patients with asthma. **Journal of allergy and clinical immunology**, v. 76, n. 5, p. 703–707, 1985.

SINGH, P.; RAJ, P.; NAMBOODIRI, V. P. EDS pooling layer. **Image and Vision Computing**, v. 98, p. 103923, jun. 2020.

ŚLEDŹ, P.; CAFLISCH, A. Protein structure-based drug design: from docking to molecular dynamics. **Current Opinion in Structural Biology**, v. 48, p. 93–102, 2018.

SONG, X. et al. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. **International Journal of Medical Informatics**, v. 151, p. 104484, 2021.

SOORI, M.; AREZOO, B.; DASTRES, R. Artificial intelligence, machine learning and deep learning in advanced robotics, a review. **Cognitive Robotics** jan. 2023.

SOROKINA, M. et al. COCONUT online: Collection of Open Natural Products database. **Journal of Cheminformatics**, v. 13, n. 1, p. 1-13, 2021.

SPARKS, T. C. et al. The new age of insecticide discovery-the crop protection industry and the impact of natural products. **Pesticide biochemistry and physiology**, v. 161, p. 12–22, 2019.

SQUEO, G. et al. Background, applications and issues of the experimental designs for mixture in the food sector. **Foods**, v. 10, n. 5, p. 1128, 2021.

SRINATH, K. R. Python–the fastest growing programming language. **International Research Journal of Engineering and Technology**, v. 4, n. 12, p. 354–357, 2017.

STANTON, D. T.; JURS, P. C. Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies. **Analytical Chemistry**, v. 62, n. 21, p. 2323–2329, 1990.

STENTA, M. **Chemistry 4.0: How the digital revolution is changing chemical research**. **Chimia**, v. 75, n. 3, p. 211-211, 2021.

SUD, M. et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. **Nucleic acids research**, v. 44, n. D1, p. D463–D470, 2016.

SUN, R. Y. Optimization for Deep Learning: An Overview. **Journal of the Operations Research Society of China**, v. 8, n. 2, p. 249–294, 1 jun. 2020.

SUSSMAN, J. L. et al. Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. **Acta Crystallographica Section D: Biological Crystallography**, v. 54, n. 6, p. 1078–1084, 1998.

SUZUKI, T.; ASHIHARA, H.; WALLER, G. R. Purine and purine alkaloid metabolism in camellia and coffea plants**. Phytochemistry**, v. 31, n. 8, p. 2575-2584, 1992.

SVATIKOVA, A. et al. Circulating free nitrotyrosine in obstructive sleep apnea. **American Journal of Physiology-Regulatory, Integrative and Comparative Physiology**, v. 287, n. 2, p. R284-R287, 2004.

SVETNIK, V. et al. Random forest: a classification and regression tool for compound classification and QSAR modeling. **Journal of chemical information and computer sciences**, v. 43, n. 6, p. 1947–1958, 2003.

SWAINSTON, N. et al. LibChEBI: An API for accessing the ChEBI database. **Journal of Cheminformatics**, v. 8, n. 1, mar. 2016.

TAMMINA, S. Transfer learning using vgg-16 with deep convolutional neural network for classifying images. **International Journal of Scientific and Research Publications (IJSRP)**, v. 9, n. 10, p. 143–150, 2019.

TARKA, S. M.; CORNISH, H. H. The toxicology of cocoa and methylxanthines: a review of the literature. **CRC critical Reviews in Toxicology**, v. 9, n. 4, p. 275–312, 1982.

TARTAGLIONE, L. et al. Dereplication of Gambierdiscus balechii extract by LC-HRMS and in vitro assay: First description of a putative ciguatoxin and confirmation of 44-methylgambierone. **Chemosphere**, v. 319, 1 abr. 2023.

TAYLOR, A. J. et al. Microbes associated with spontaneous cacao fermentations - A systematic review and meta-analysis. **Current Research in Food Science**, v. 5, p. 1452–1464, jan. 2022.

THAKARE, R. et al. Antibiotics: past, present, and future. **Current opinion in microbiology**, v. 51, p. 72-80, 2019.

THIEMANN, T. Isolation of Phthalates and Terephthalates from Plant Material–Natural Products or Contaminants? **Open Chemistry Journal**, v. 8, n. 1, 2021.

TODESCHINI, R.; CONSONNI, V. **Handbook of molecular descriptors**. [s.l.] John Wiley & Sons, 2008.

TROTT, O.; OLSON, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. **Journal of computational chemistry**, v. 31, n. 2, p. 455–461, 2010.

TSAVKELOVA, E. et al. Identification and functional characterization of indole-3-acetamide-mediated IAA biosynthesis in plant-associated Fusarium species. **Fungal Genetics and Biology**, v. 49, n. 1, p. 48–57, 2012.

TSLJCHIE, H. Effect of cacao husk extract on human immunodeficiency virus infection. **Letters in Applied Microbiology**, v. 13, n. 6, p. 251-254, 1991.

TSUGAWA, H. et al. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. **Nature Methods**, v. 12, n. 6, p. 523–526, 28 maio 2015.

UMEDA, M. et al. Preventive effects of tea and tea catechins against influenza and acute upper respiratory tract infections: A systematic review and meta-analysis. **European Journal of Nutrition**, v. 60, n. 8, p. 4189–4202, 2021a.

VALLI, M.; BOLZANI, V. S. Natural products: perspectives and challenges for use of Brazilian plant species in the bioeconomy. **Anais da Academia Brasileira de Ciências**, v. 91, 2019.

VAN DER SPOEL, D. et al. **GROMACS: Fast, flexible, and free**. **Journal of Computational Chemistry**, dez. 2005.

VERDONK, M. L. et al. Improved protein–ligand docking using GOLD. **Proteins: Structure, Function, and Bioinformatics**, v. 52, n. 4, p. 609–623, 2003.

VIEIRA, R. et al. Induction of metabolic variability of the endophytic fungus Xylaria sp. by OSMAC approach and experimental design. **Archives of Microbiology**, v. 203, n. 6, p. 3025–3032, 1 ago. 2021.

VIEIRA, R. et al. CHEIC: Chemical Image Classificator. An intelligent system for identification of volatiles compounds with potential for respiratory diseases using Deep Learning. **Expert Systems with Applications**, v. 234, p. 121178, dez. 2023.

VIEIRA, R.; ALVES DE SOUSA, K.; CASTRO-GAMBOA, I. **LUMIOS: Label Using Machine In Organic Samples-a software for dereplication, molecular docking, and combined machine and deep learning**. [s.l: s.n.].

VONRANKE, N. L. et al. Structure-activity relationship, molecular docking, and molecular dynamic studies of diterpenes from marine natural products with anti-HIV activity. **Journal of Biomolecular Structure and Dynamics**, v. 40, n. 7, p. 3185–3195, 2022.

WANG, G. et al. Trehalose and glucose levels regulate feeding behavior of the phloem-feeding insect, the pea aphid Acyrthosiphon pisum Harris. **Scientific Reports**, v. 11, n. 1, p. 15864, 2021.

WANG, W.; WANG, J.; KOLLMAN, P. A. What determines the van der Waals coefficient β in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations? **Proteins: Structure, Function and Genetics**, v. 34, n. 3, p. 395–402, 15 fev. 1999.

WOSCHANK, M.; RAUCH, E.; ZSIFKOVITS, H. A review of further directions for artificial intelligence, machine learning, and deep learning in smart logistics. **Sustainability (Switzerland)**, v. 12, n. 9, maio 2020.

WYNER, A. J. et al. Explaining the success of adaboost and random forests as interpolating classifiers. **The Journal of Machine Learning Research**, v. 18, n. 1, p. 1558–1590, 2017.

YAHYA, M.; GINTING, B.; SAIDI, N. In-Vitro Screenings for Biological and Antioxidant Activities of Water Extract from Theobroma cacao L. Pod Husk: Potential Utilization in Foods. **Molecules**, v. 26, n. 22, p. 6915, 2021.

YAMASHITA, R. et al. Convolutional neural networks: an overview and application in radiology. **Insights into Imaging**, v. 9, p. 611-629, 2018.

YAÑEZ, O. et al. Theobroma cacao L. compounds: Theoretical study and molecular modeling as inhibitors of main SARS-CoV-2 protease. **Biomedicine & Pharmacotherapy**, v. 140, p. 111764, 2021.

YANG, Z.; FANG, Y.; JI, H. Controlled release and enhanced antibacterial activity of salicylic acid by hydrogen bonding with chitosan. **Chinese Journal of Chemical Engineering**, v. 24, n. 3, p. 421–426, mar. 2016.

YARKONI, E.; BEKIERKUNST, A. Nonspecific resistance against infection with Salmonella typhi and Salmonella typhimurium induced in mice by cord factor (trehalose-6, 6'-dimycolate) and its analogues. **Infection and immunity**, v. 14, n. 5, p. 1125–1129, 1976.

YUAN, H. et al. The traditional medicine and modern medicine from natural products. **Molecules**, v. 21, n. 5, p. 559, 2016.

ZHANG, C. et al. Comparative Research on Network Intrusion Detection Methods Based on Machine Learning. **Computers & Security**, p. 102861, 2022.

ZHANG, H. et al. Organism-derived phthalate derivatives as bioactive natural products. **Journal of Environmental Science and Health, Part C**, v. 36, n. 3, p. 125–144, 2018.

ZHANG, J. et al. New lignans and their biological activities. **Chemistry & Biodiversity**, v. 11, n. 1, p. 1–54, 2014.

ZHANG, L. et al. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. **Drug discovery today**, v. 22, n. 11, p. 1680–1685, 2017.

ZHAO, J. et al. Plant-derived bioactive compounds produced by endophytic fungi. **Mini reviews in medicinal chemistry**, v. 11, n. 2, p. 159–168, 2011.

ZHENG, X.-Q. et al. Biosynthesis, accumulation and degradation of theobromine in developing Theobroma cacao fruits. **Journal of plant physiology**, v. 161, n. 4, p. 363-369, 2004.

ZHOU, D.-Y. et al. Proanthocyanidin from grape seed extract inhibits airway inflammation and remodeling in a murine model of chronic asthma. **Natural Product Communications**, v. 10, n. 2, p. 1934578X1501000210, 2015.

ZHU, W. et al. Anti-inflammatory and immunomodulatory effects of iridoid glycosides from Paederia scandens (LOUR.) MERRILL (Rubiaceae) on uric acid nephropathy rats. **Life Sciences**, v. 91, n. 11–12, p. 369–376, 5 out. 2012.

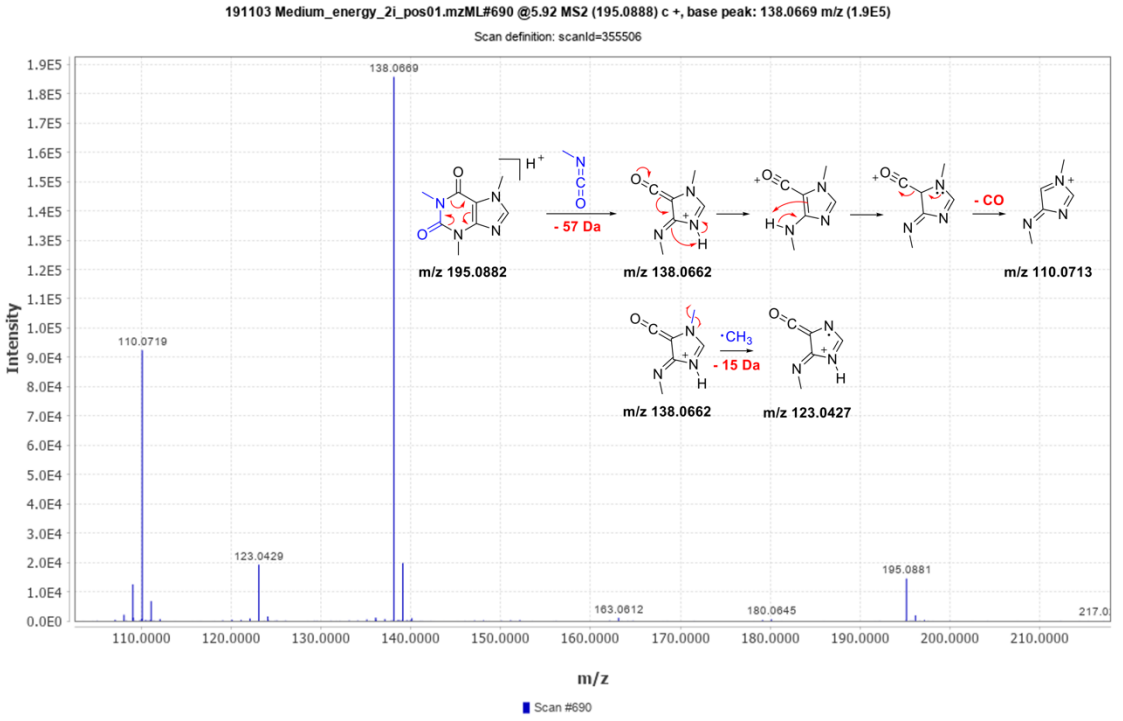
ZIĘBA, K.; MAKAREWICZ-WUJEC, M.; KOZŁOWSKA-WOJCIECHOWSKA, M. Cardioprotective Mechanisms of Cocoa. **Journal of the American College of Nutrition**, v. 38, n. 6, p. 564-575, 2019.

ZUMAETA, C. R. B. et al. Metabolomics during the spontaneous fermentation in cocoa (Theobroma cacao L.): An exploraty review. **Food Research International**, p. 112190, 2022.

1 MATERIAL SUPLEMENTAR A – CAPÍTULO 2

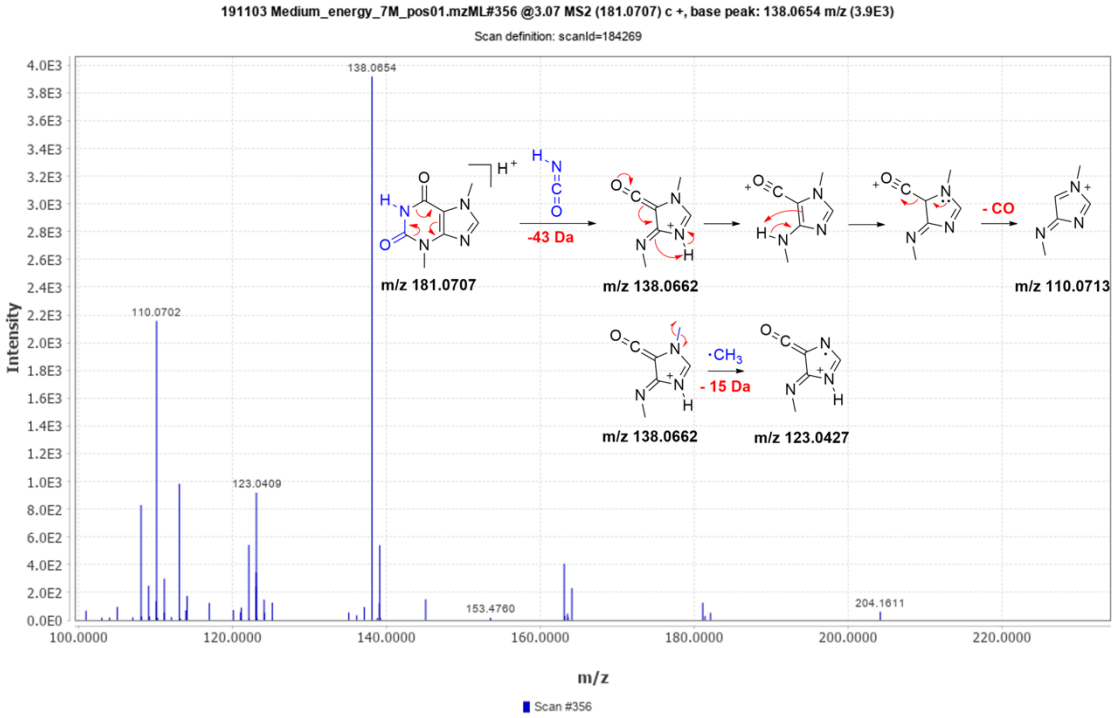
1.1 A – CAFEÍNA

Figura Suplementar 1 - Espectro de massa (MS2) atribuído à anotação molecular da cafeína, bem como os mecanismos propostas para justificativa dos sinais principais



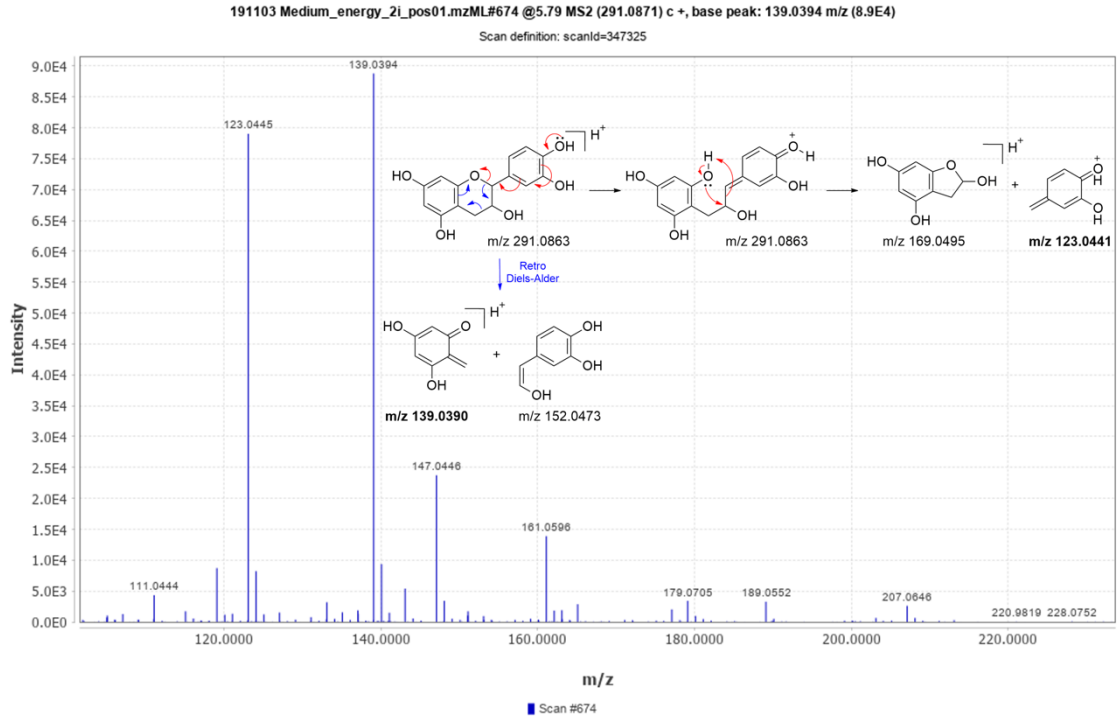
1.2 B – TEOBROMINA

Figura Suplementar 2 - Espectro de massa (MS2) atribuído à anotação molecular da teobromina, bem como os mecanismos propostas para justificativa dos sinais principais



1.3 C – CATEQUINA

Figura Suplementar 3 - Espectro de massa (MS2) atribuído à anotação molecular da catequina, bem como os mecanismos propostas para justificativa dos sinais principais.



1.4 D – PROCIANIDINA

Figura Suplementar 4 - Espectro de massa (MS2) atribuído à anotação molecular da procianidina.

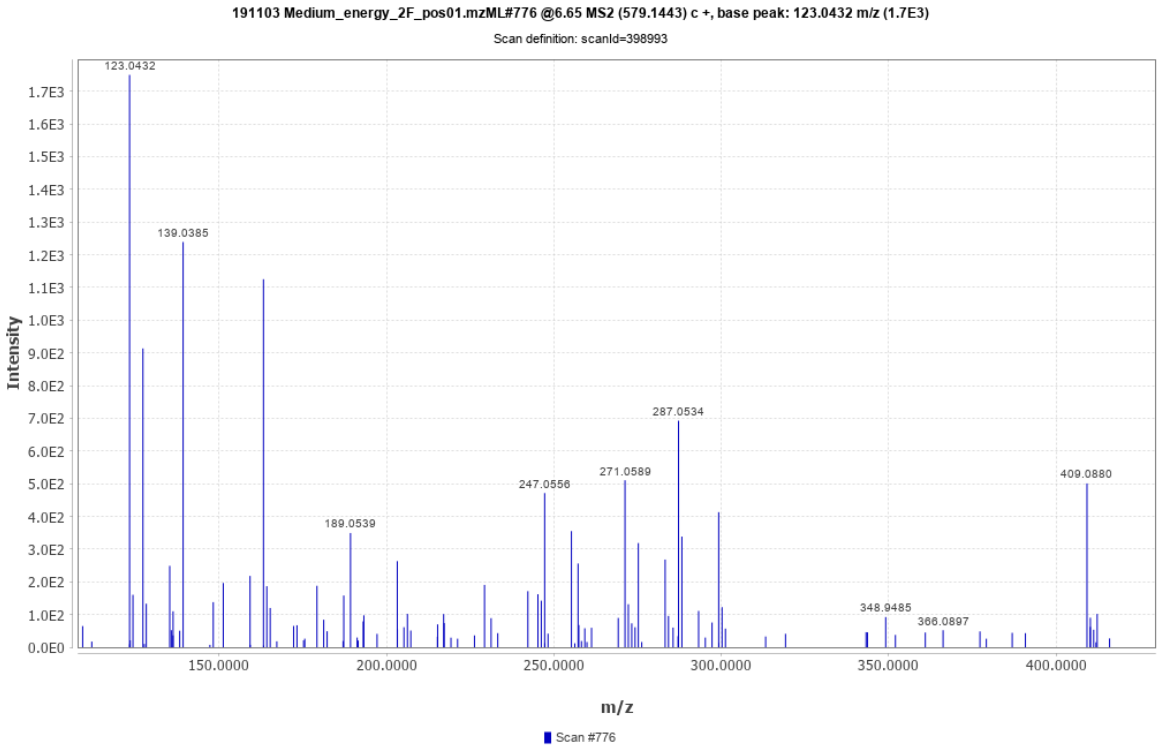
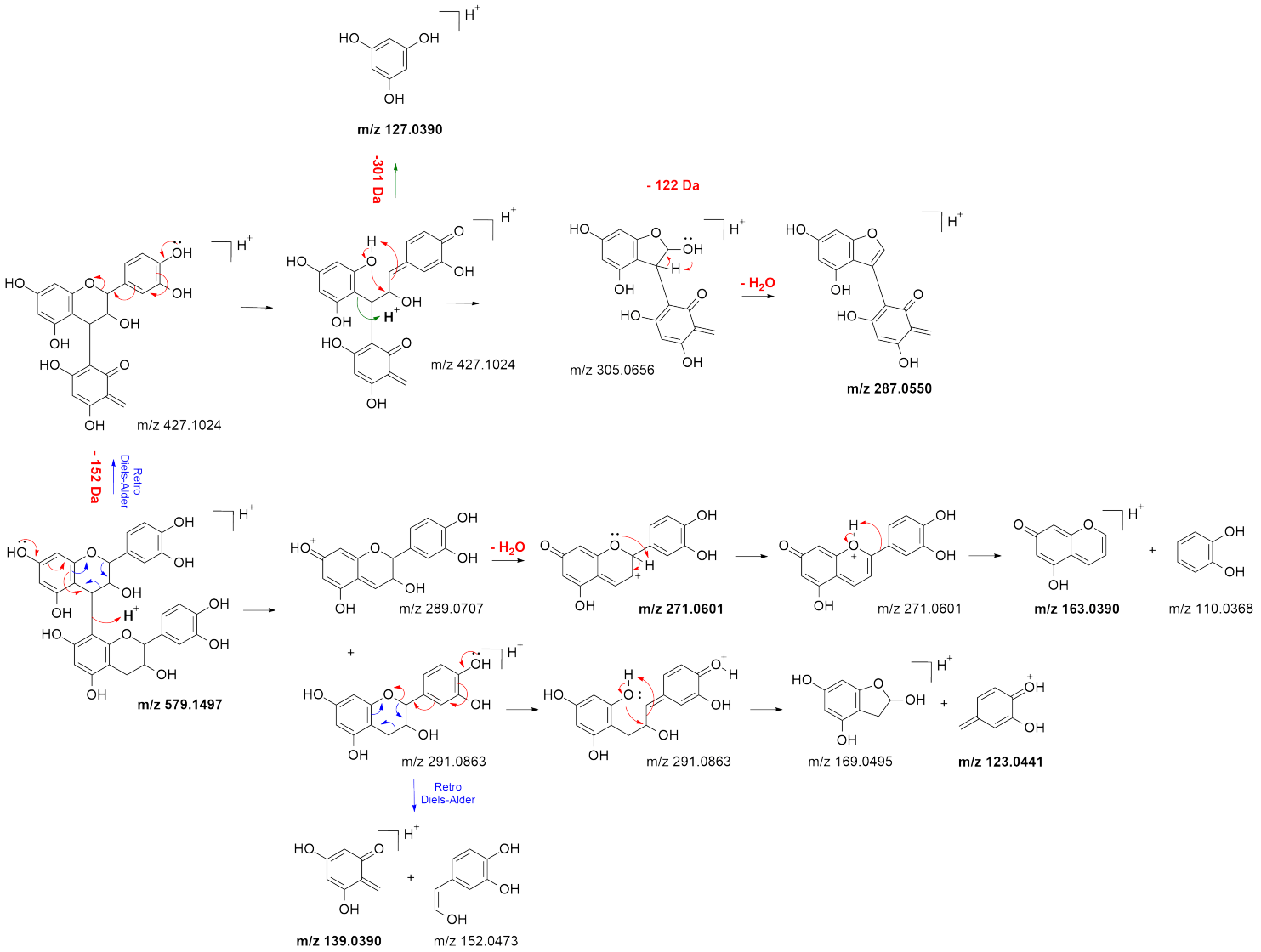
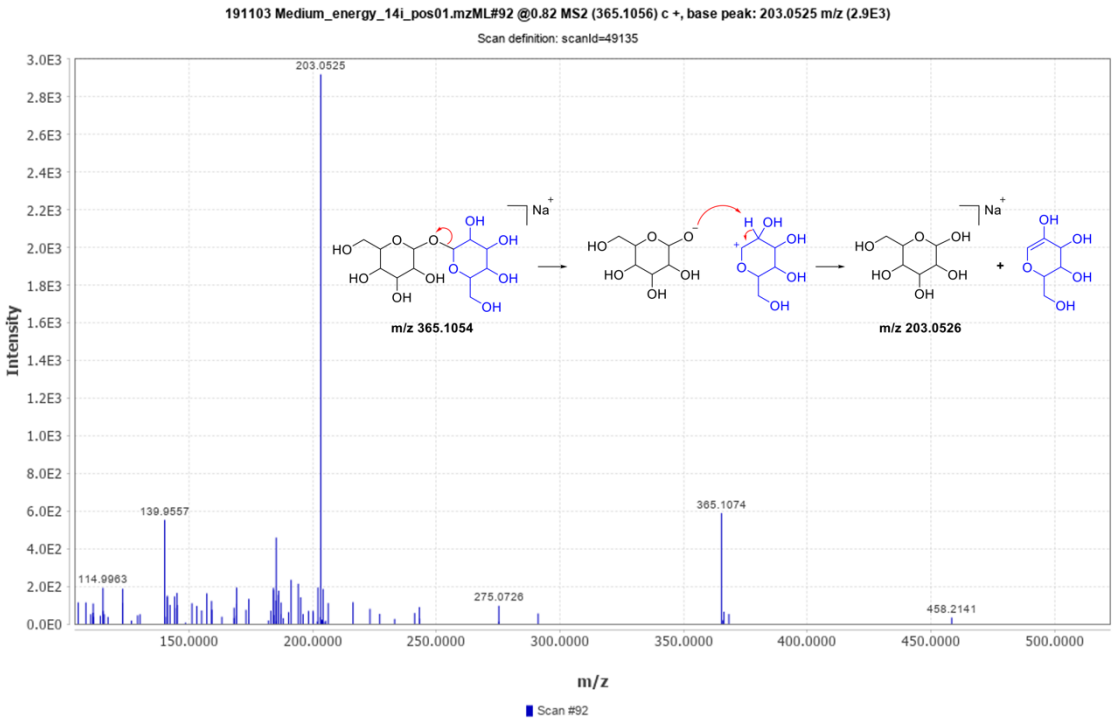


Figura Suplementar 5 - Mecanismos propostos para justificativa dos sinais principais da procianidina.



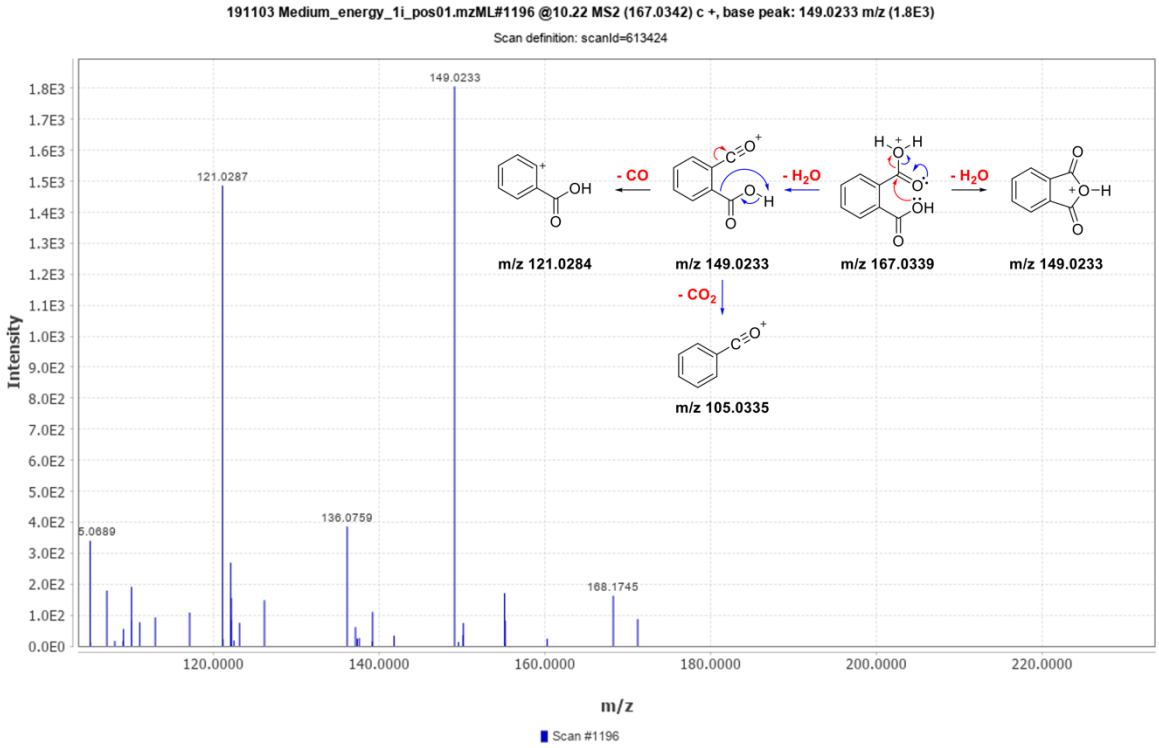
1.5 E – TREALOSE

Figura Suplementar 6 - Espectro de massa (MS2) atribuído à anotação molecular da trealose, bem como os mecanismos propostas para justificativa dos sinais majoritários



1.6 F – ÁCIDO FTÁLICO[8]

Figura Suplementar 7 - Espectro de massa (MS2) atribuído à anotação molecular do ácido ftálico, bem como os mecanismos propostos para justificativa dos sinais majoritários.



[8] Os ftalatos apresentam os mesmos padrões de fragmentação, com a presença do sinal com razão massa/carga de 149,0233 e 121,0287.

1.7 J – TIROSINA

Figura Suplementar 8 - Espectro de massa (MS2) atribuído à anotação molecular da tirosina, bem como os mecanismos propostas para justificativa dos sinais majoritários.

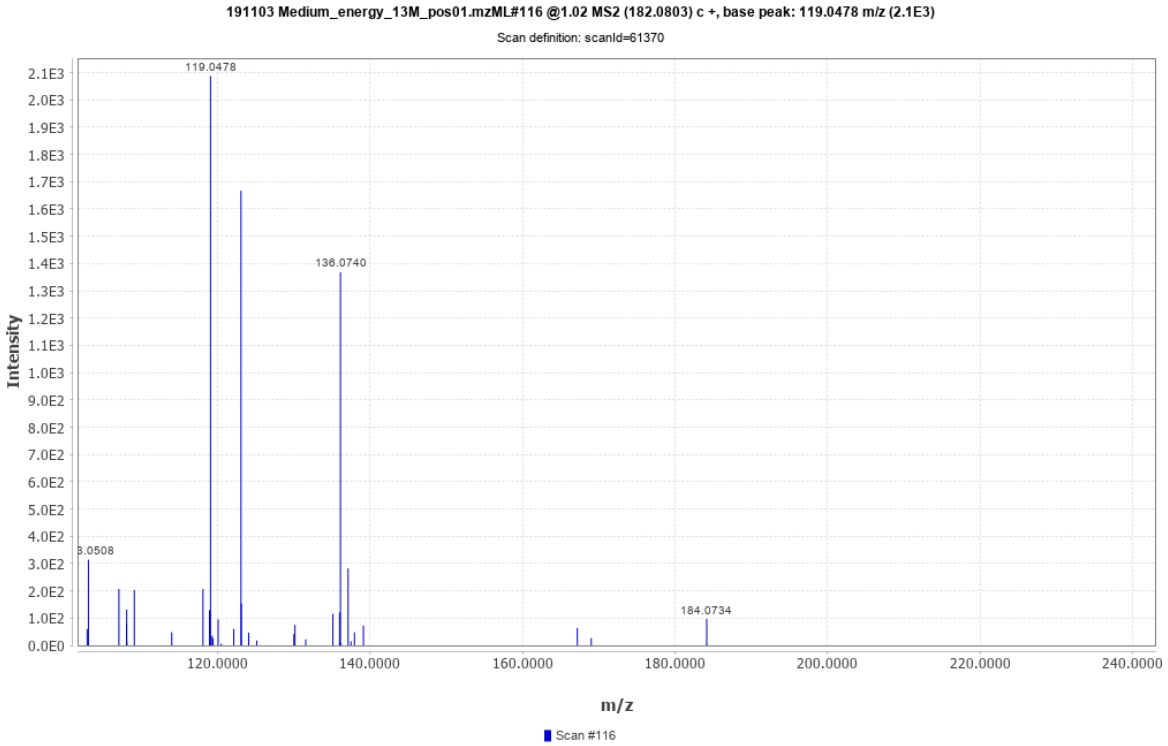
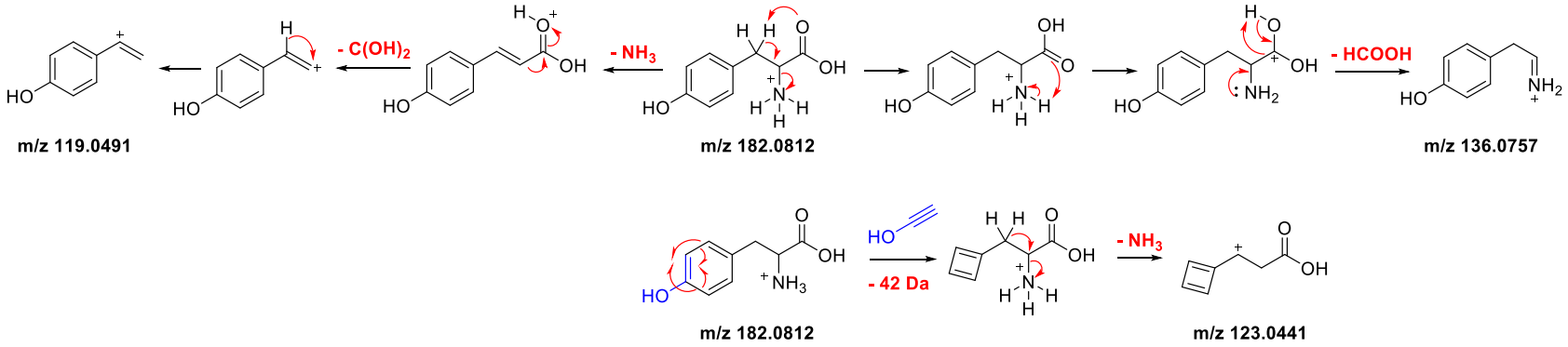


Figura Suplementar 9 - Mecanismos propostos para a anotação molecular da tirosina.



1.8 K – FENILALANINA

Figura Suplementar 10 - Espectro de massa (MS2) atribuído à anotação molecular da fenilalanina.

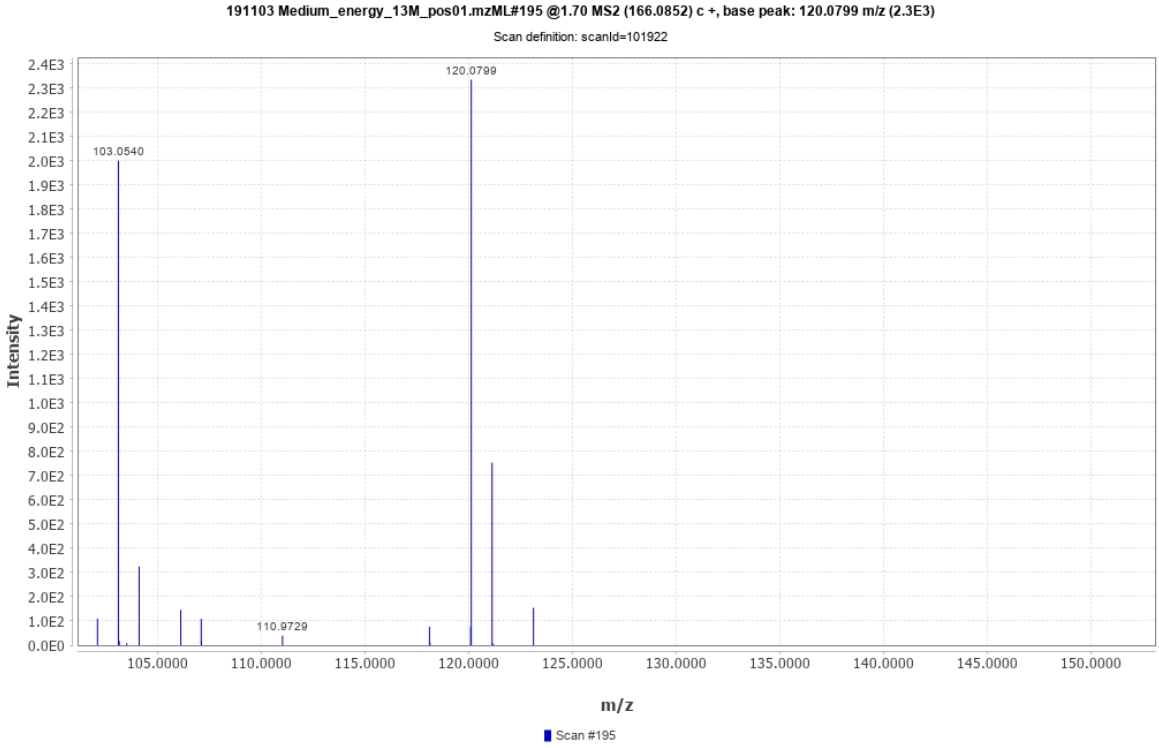
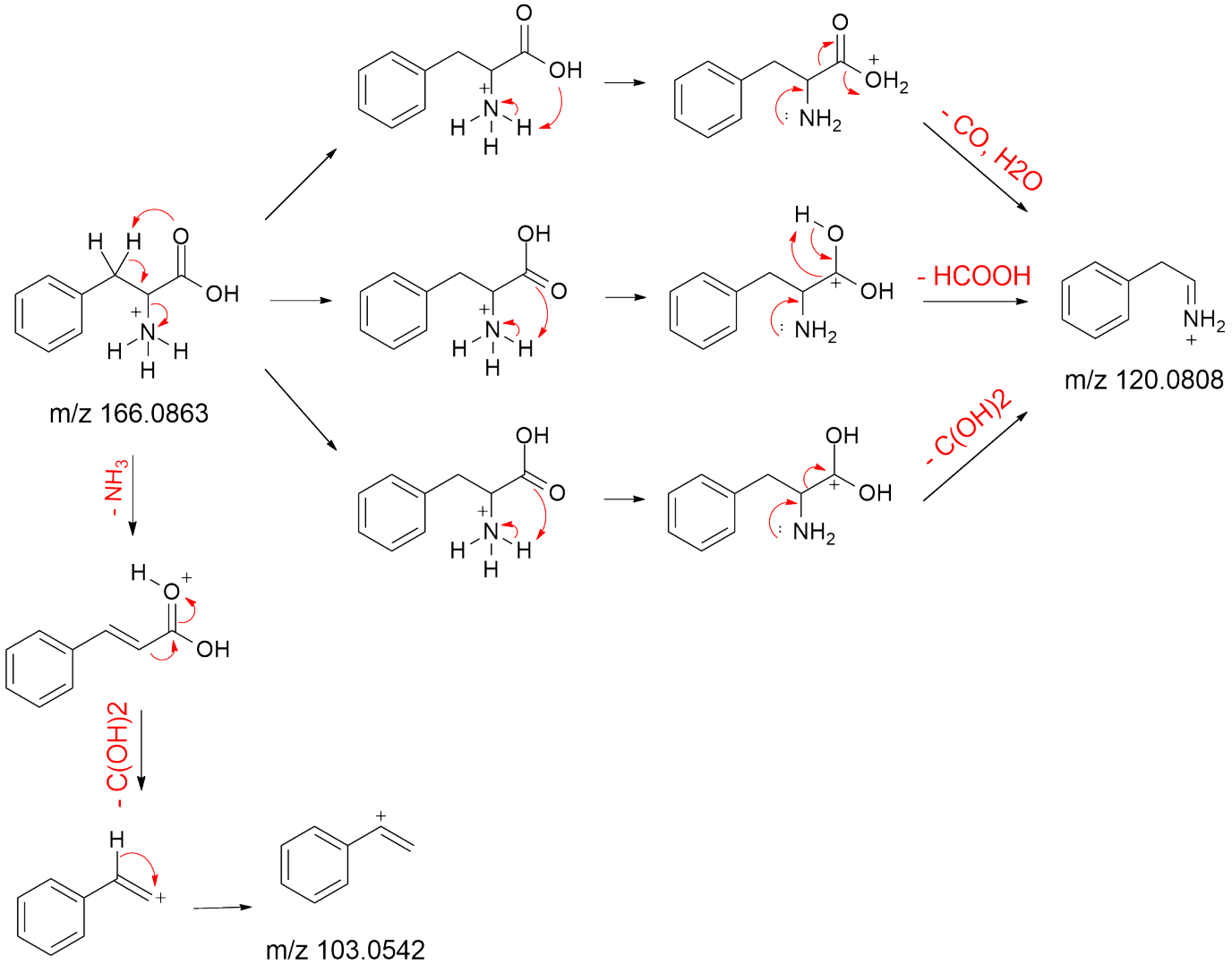
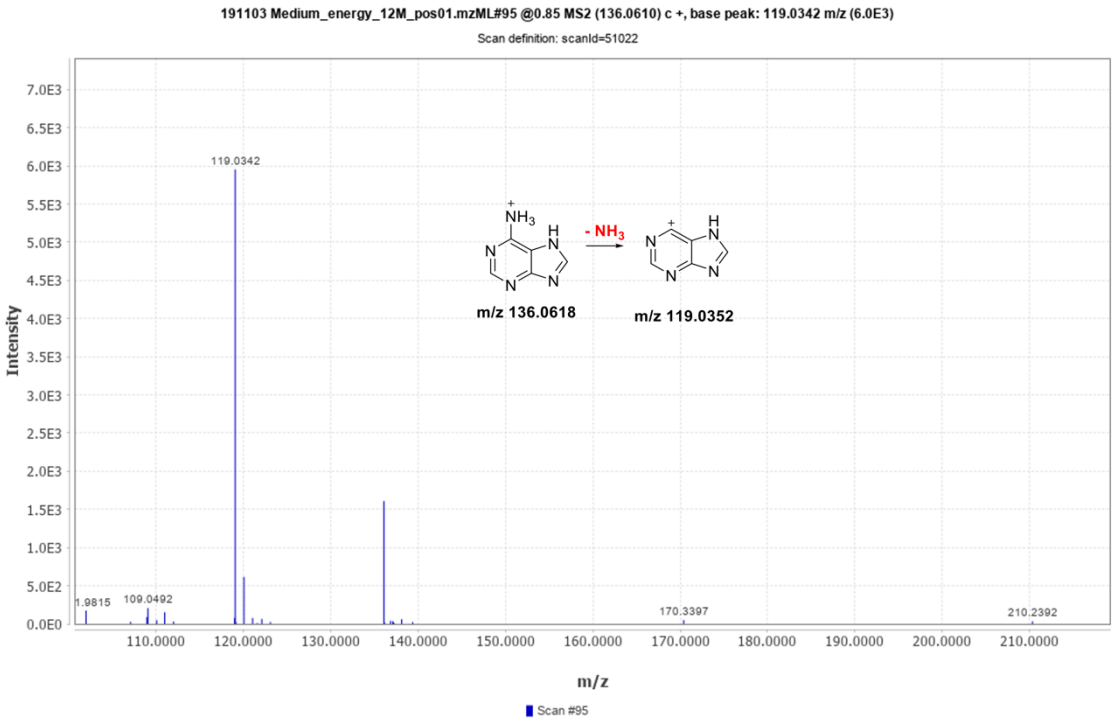


Figura Suplementar 11 - Proposta de mecanismos de fragmentação da anotação molecular fenilalanina.



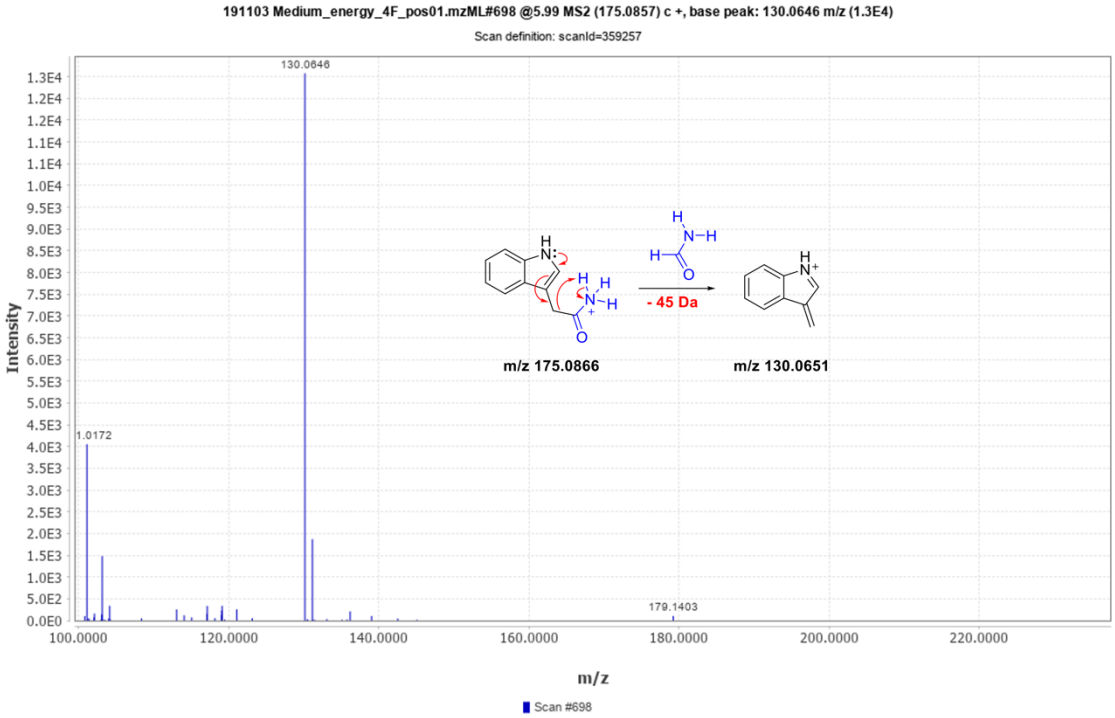
1.9 L – ADENINA

Figura Suplementar 12 - Espectro de massa (MS2) atribuído à anotação molecular da adenina, bem como os mecanismos propostas para justificativa dos sinais majoritários.



1.10 M – INDOL-3-ACETAMIDA

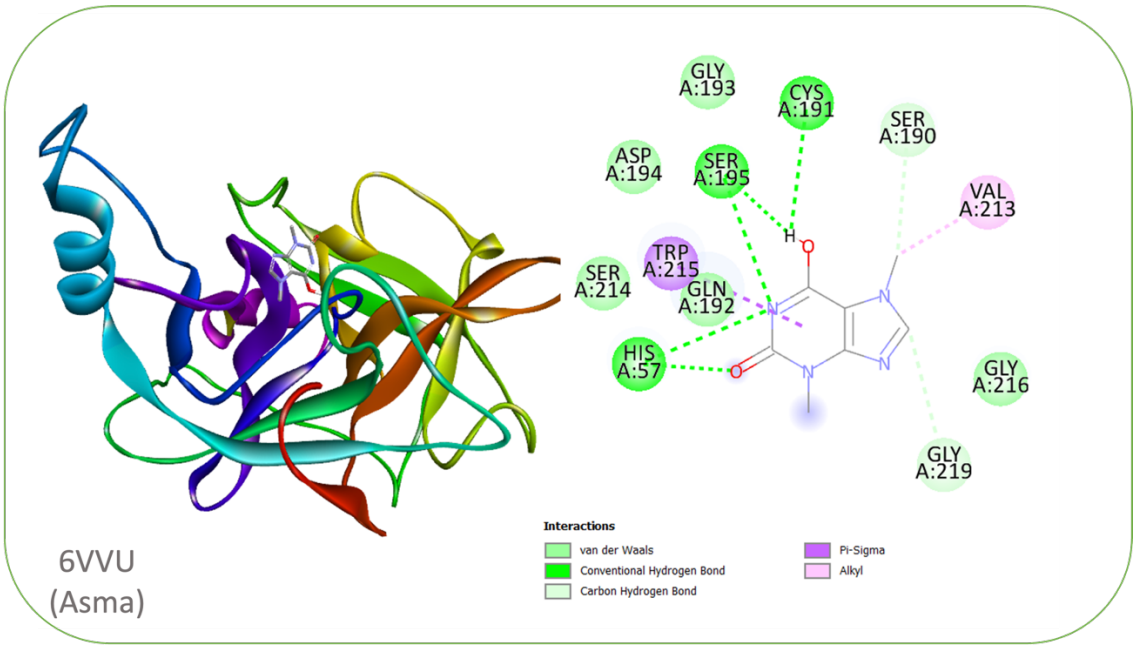
Figura Suplementar 13 - Espectro de massa (MS2) atribuído à anotação molecular da indol-3-acetamida, bem como os mecanismos propostas para justificativa dos sinais majoritários.



2 VISUALIZAÇÕES OBTIDAS DOS RESULTADOS DE DOCAGEM MOLECULAR[9]

2.1 Docking Teobromina

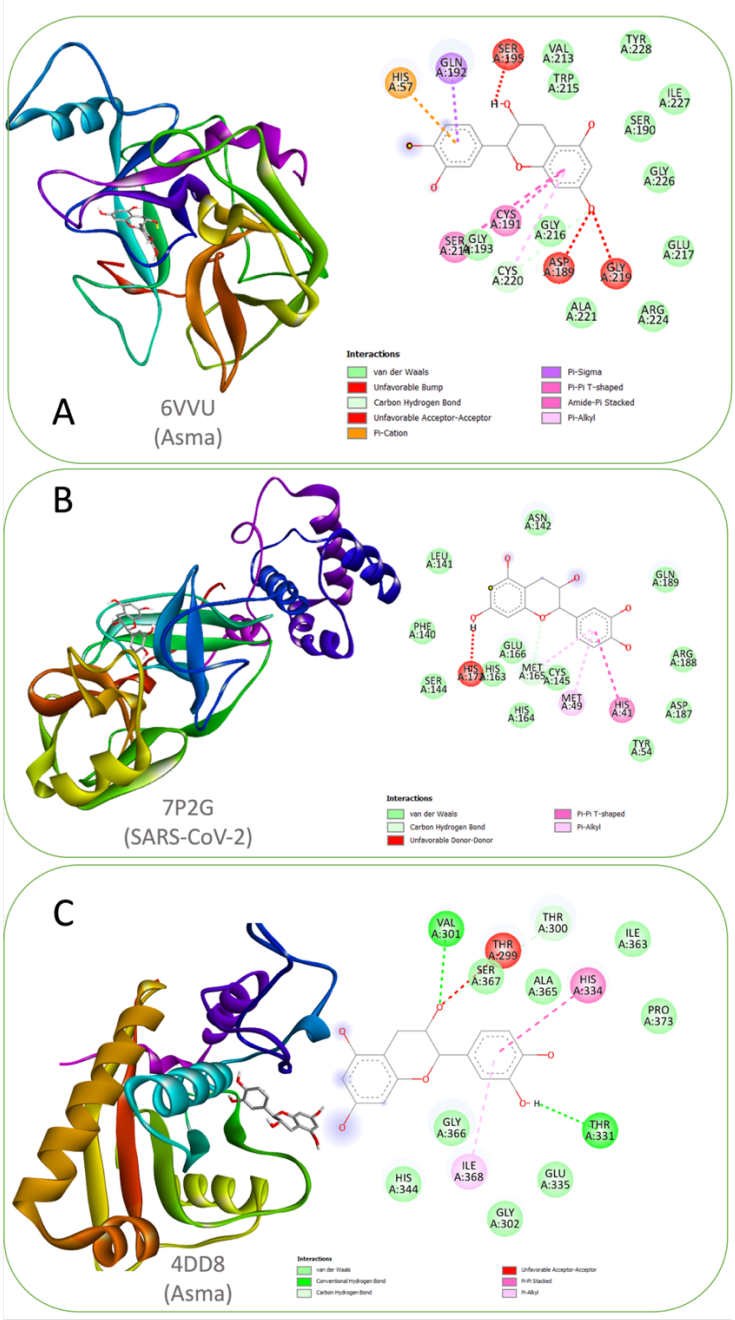
Figura Suplementar 14 - Resultados de docagem molecular para a molécula de teobromina no alvo 6VVU (-4,6 kcal/mol).



[9] As visualizações da docagem foram criadas apenas para as anotações moleculares que apresentaram afinidade melhor do que o ligante padrão associado às respectivas proteínas.

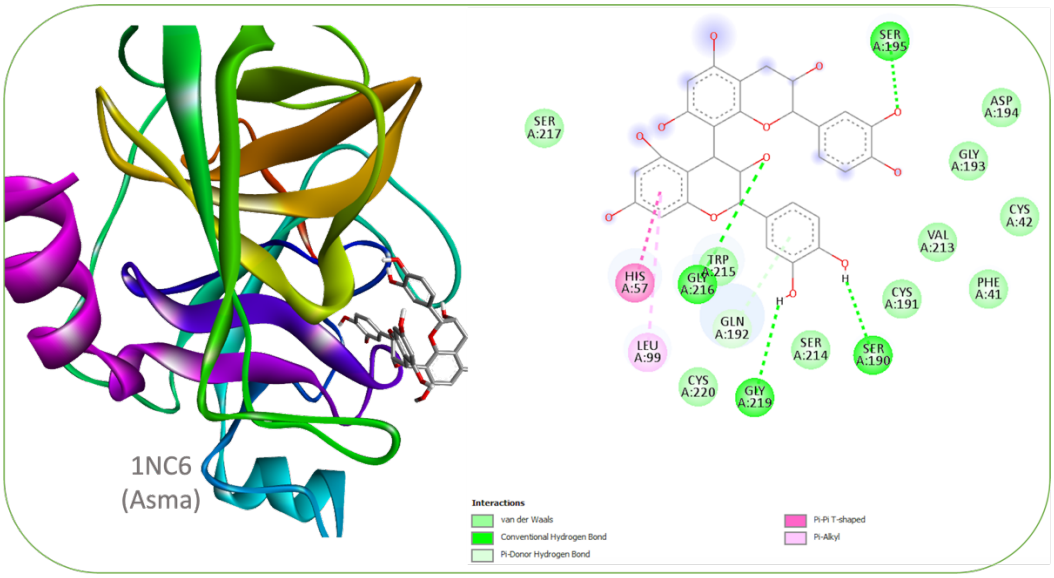
2.2 Docking Catequina

Figura Suplementar 15 - Resultados de docagem molecular para a molécula de catequina no alvo 6VVU (-4,3 kcal/mol)(A), 7P2G (-6,7 kcal/mol) (B) e 4DD8 (-6,4 kcal/mol) (C)



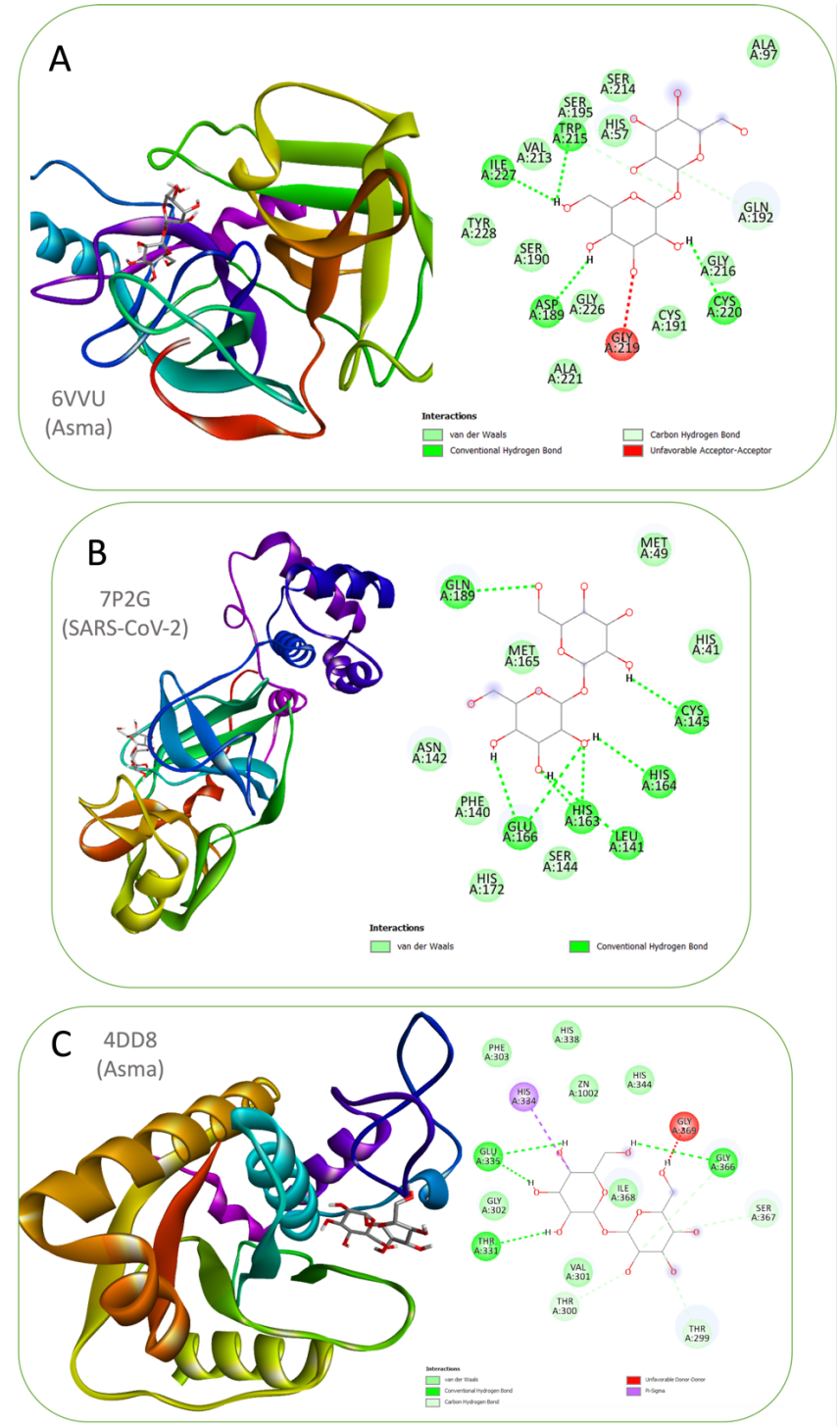
2.3 Docking Procianidina

Figura Suplementar 16 - Resultados de docagem molecular para a molécula de procianidina no alvo 1NC6 (-6,1 kcal/mol).



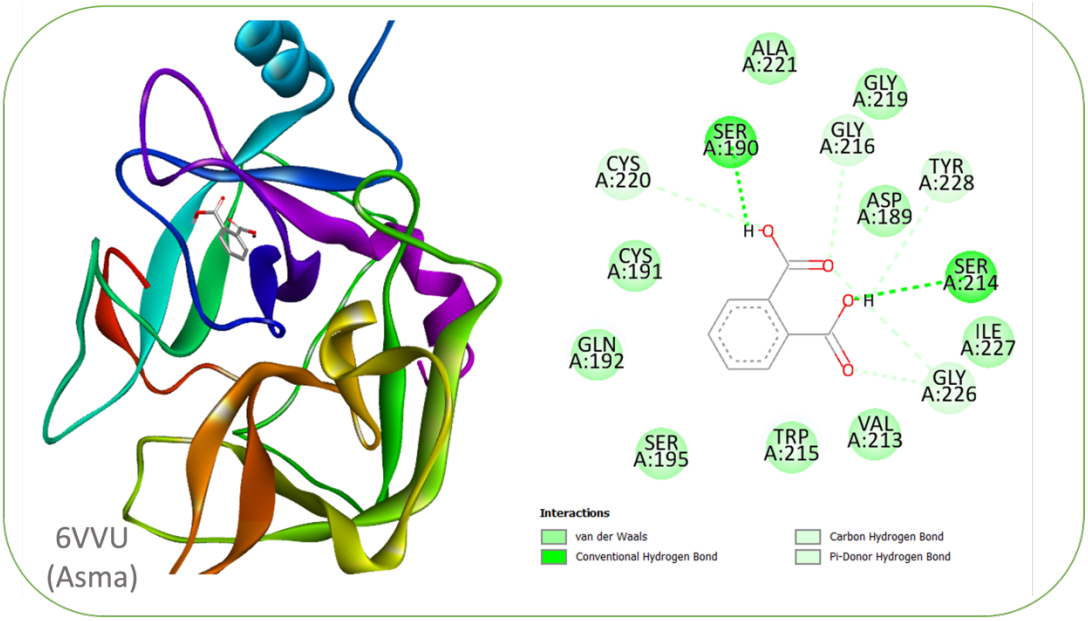
2.4 Docking trealose

Figura Suplementar 17 - Resultados de docagem molecular para a molécula de trealose no alvo 6VVU (-5,5 kcal/mol)(A), 7P2G (-6,3 kcal/mol) (B) e 4DD8 (-6,4 kcal/mol) (C).



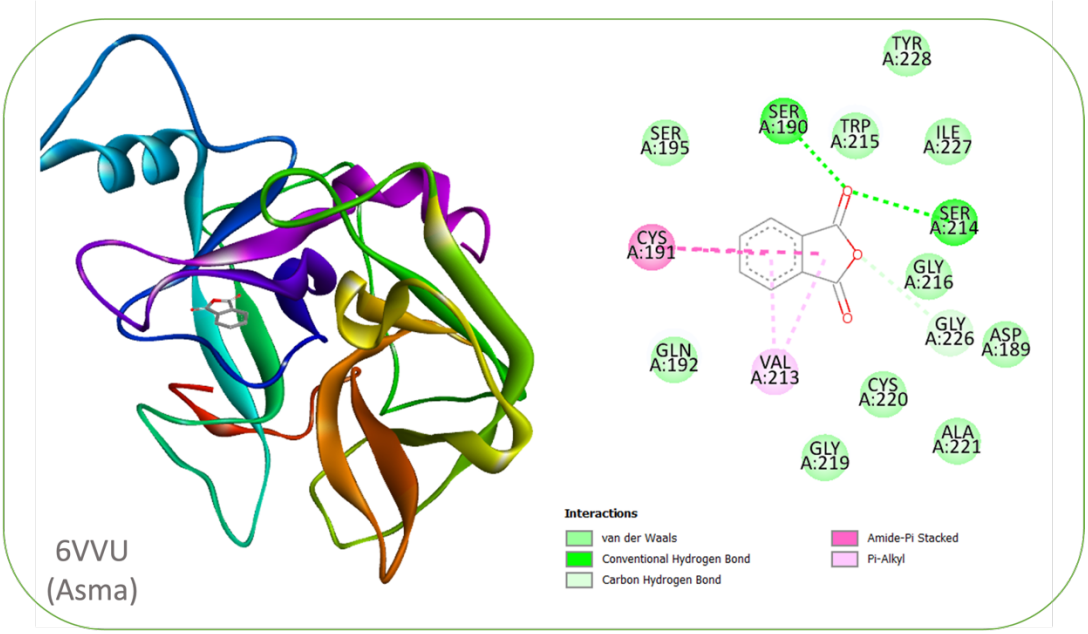
2.5 Docking ácido ftálico

Figura Suplementar 18 - Resultados de docagem molecular para a molécula de ácido ftálico no alvo 6VVU (-4,9 kcal/mol).



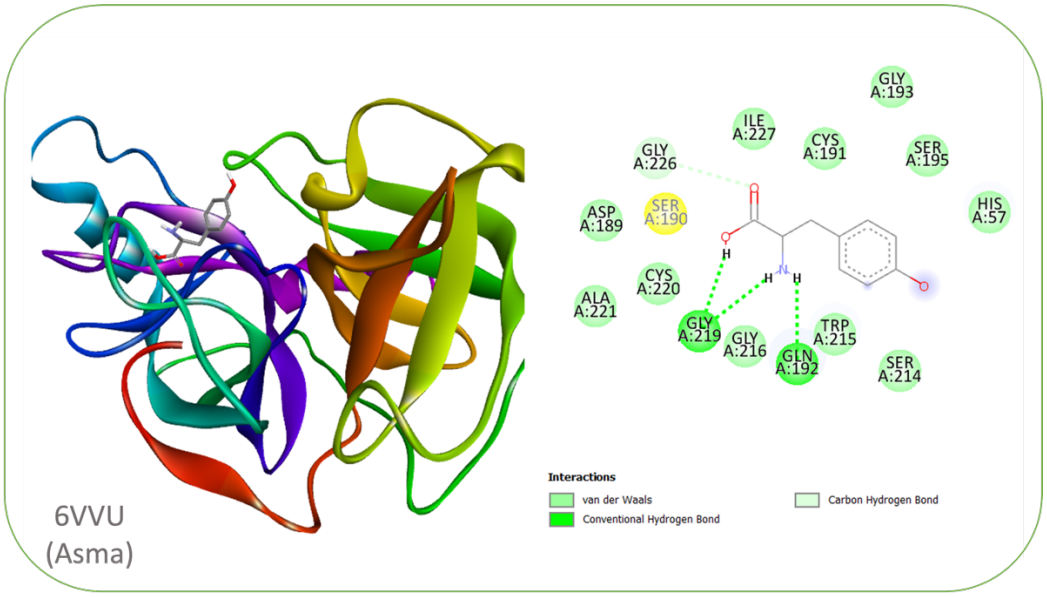
2.6 Docking anidrido ftálico

Figura Suplementar 19 - Resultados de docagem molecular para a molécula de anidrido ftálico no alvo 6VVU (-5,0 kcal/mol).



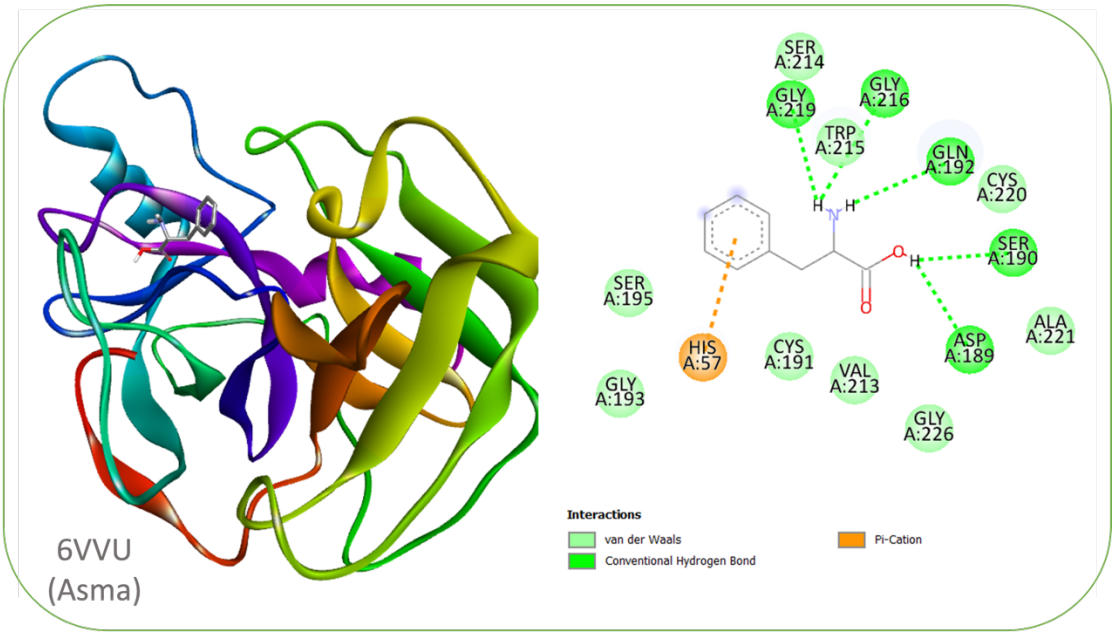
2.7 Docking tirosina

Figura Suplementar 20 - Resultados de docagem molecular para a molécula de tirosina no alvo 6VVU (-5,5 kcal/mol).



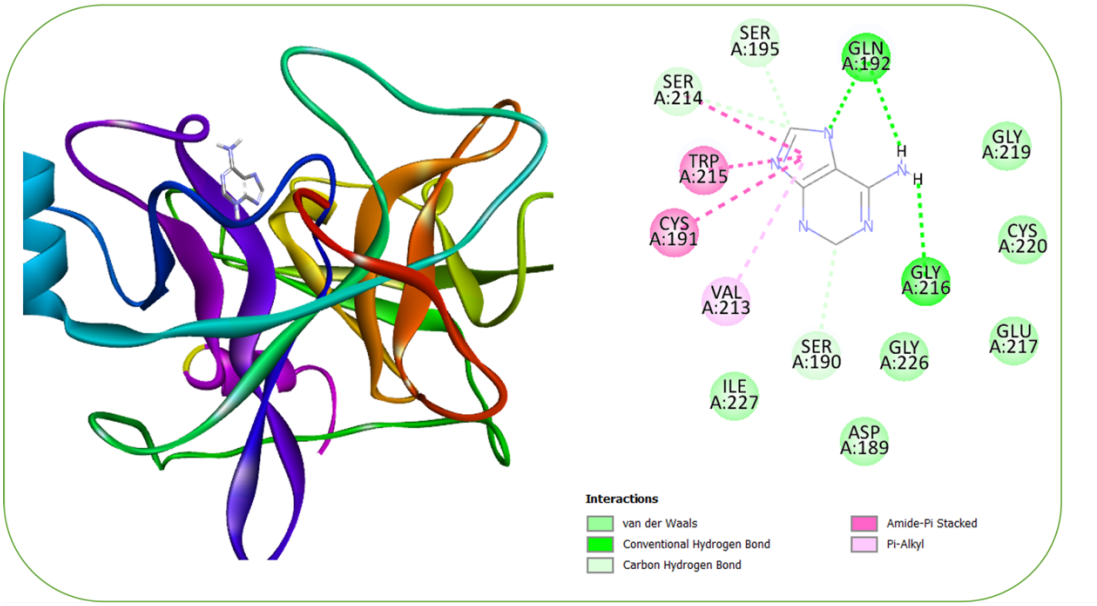
2.8 Docking fenilalanina

Figura Suplementar 21 - Resultados de docagem molecular para a molécula de fenilalanina no alvo 6VVU (-5,3 kcal/mol).



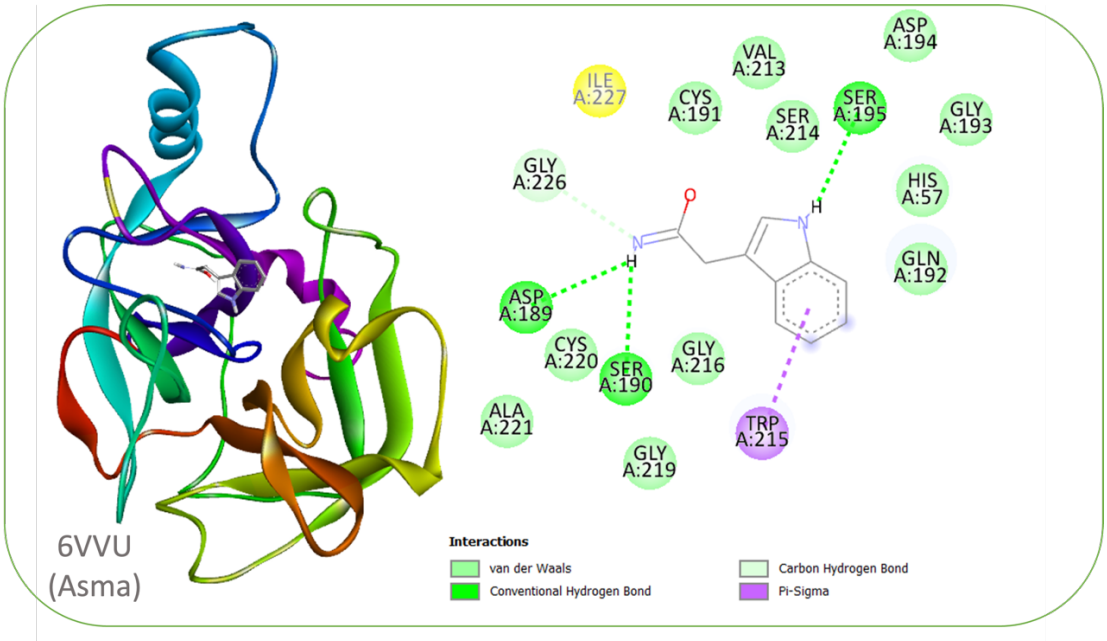
2.9 Docking adenina

Figura Suplementar 22 - Resultados de docagem molecular para a molécula de adenina no alvo 6VVU (-5,6 kcal/mol).



2.10 Docking Indol-3-acetamida

Figura Suplementar 23 - Resultados de docagem molecular para a molécula de indol-3-acetamida no alvo 6VVU (-5,8 kcal/mol).



MATERIAL SUPLEMENTAR B – CAPÍTULO 3

Tabela Suplementar 1 - Planejamento experimental da trealose, usada como exemplo de funcionamento do API Chemistika (0 hora de fermentação) usando a intensidade dos sinais como resposta.

EXP	A	B	C	TREALOSE
1	1,00	0,00	0,00	35744
2	0,00	1,00	0,00	2980
3	0,00	0,00	1,00	12708
4	0,33	0,67	0,00	1751
5	0,33	0,00	0,67	54977
6	0,00	0,33	0,67	51022
7	0,67	0,33	0,00	50382
8	0,67	0,00	0,33	34544
9	0,00	0,67	0,33	55645
10	0,33	0,33	0,33	41943
11	0,67	0,17	0,17	37795
12	0,17	0,67	0,17	20012
13	0,17	0,17	0,67	59771
14	0,33	0,33	0,33	76506
1	1,00	0,00	0,00	36101
2	0,00	1,00	0,00	3010
3	0,00	0,00	1,00	12835
4	0,33	0,67	0,00	1769
5	0,33	0,00	0,67	55527
6	0,00	0,33	0,67	51532
7	0,67	0,33	0,00	50886
8	0,67	0,00	0,33	34889
9	0,00	0,67	0,33	56201
10	0,33	0,33	0,33	42362
11	0,67	0,17	0,17	38173
12	0,17	0,67	0,17	20212
13	0,17	0,17	0,67	60369
14	0,33	0,33	0,33	77271

MATERIAL SUPLEMENTAR C – CAPÍTULO 4

Tabela Suplementar 2 - Planejamento experimental para anotações moleculares (0 hora de fermentação) usando a intensidade dos sinais como resposta.

EXP	A	B	C	ANIDRIDO FTÁLICO.	ÁCIDO FTÁLICO	THEOB	CATEC	TREAL	PROCIAN.
1	1,00	0,00	0,00	475617	44004	1931	1161	35744	0
2	0,00	1,00	0,00	43573	1275	146977	824828	2980	928
3	0,00	0,00	1,00	11237	1352	4887	10155	12708	3
4	0,33	0,67	0,00	8809	1888	59146	489965	1751	527
5	0,33	0,00	0,67	5241	2213	13795	917	54977	0
6	0,00	0,33	0,67	4955	2519	8223	783	51022	14
7	0,67	0,33	0,00	3820	2341	377566	14964	50382	13633
8	0,67	0,00	0,33	3867	2213	32430	60	34544	76
9	0,00	0,67	0,33	3417	1935	17627	4071	55645	73
10	0,33	0,33	0,33	3569	2559	17634	253	41943	17
11	0,67	0,17	0,17	3715	2180	14037	122	37795	0
12	0,17	0,67	0,17	4052	2283	59045	196593	20012	541
13	0,17	0,17	0,67	3476	1938	8262	332	59771	16
14	0,33	0,33	0,33	3526	2136	11237	371	76506	11
1	1,00	0,00	0,00	480373	44444	1950	1173	36101	0
2	0,00	1,00	0,00	44009	1288	148447	833076	3010	937
3	0,00	0,00	1,00	11349	1366	4936	10257	12835	3
4	0,33	0,67	0,00	8897	1907	59737	494865	1769	532
5	0,33	0,00	0,67	5293	2235	13933	926	55527	0
6	0,00	0,33	0,67	5005	2544	8305	791	51532	14
7	0,67	0,33	0,00	3858	2364	381342	15114	50886	13769
8	0,67	0,00	0,33	3906	2235	32754	61	34889	77
9	0,00	0,67	0,33	3451	1954	17803	4112	56201	74
10	0,33	0,33	0,33	3605	2585	17810	256	42362	17
11	0,67	0,17	0,17	3752	2202	14177	123	38173	0
12	0,17	0,67	0,17	4093	2306	59635	198559	20212	546
13	0,17	0,17	0,67	3511	1957	8345	335	60369	16
14	0,33	0,33	0,33	3561	2157	11349	375	77271	11

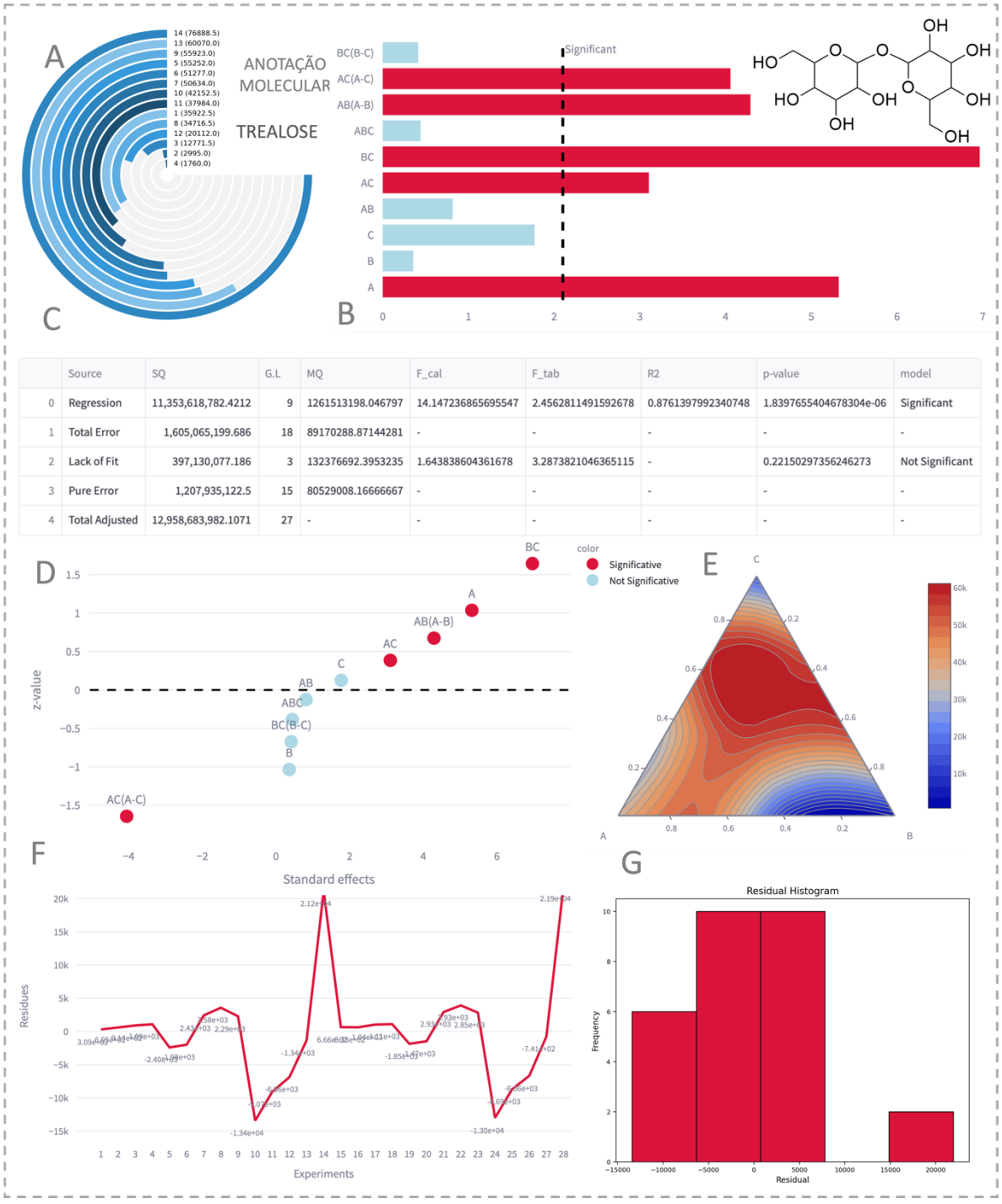
Tabela Suplementar 3 - Planejamento experimental para anotações moleculares (84 horas de fermentação) usando a intensidade dos sinais como resposta.

EXP	A	B	C	ADEN.	ANIDRIDO FTÁLICO	FENILA	TEOBR	TIROS	CATEQ	PROCI.
1	1,00	0,00	0,00	9262	1488	15979	669919	20961	0	53
2	0,00	1,00	0,00	47395	4628	7618	476444	9950	79549	3742
3	0,00	0,00	1,00	12816	6051	9966	437800	21597	1095	134
4	0,33	0,67	0,00	32681	6038	5072	1058946	11568	47097	968
5	0,33	0,00	0,67	13261	6820	13540	130494	27721	214	26
6	0,00	0,33	0,67	17214	7083	17748	385004	31062	430	64
7	0,67	0,33	0,00	30155	109828	701	1443	2169	76	23
8	0,67	0,00	0,33	15782	6892	15215	287788	31067	453	43
9	0,00	0,67	0,33	25111	9086	13782	449417	24294	9152	380
10	0,33	0,33	0,33	28276	7211	11415	527718	26022	7787	372
11	0,67	0,17	0,17	12561	5622	10348	53890	24548	5462	254
12	0,17	0,67	0,17	68130	5699	11210	577528	17933	369244	30761
13	0,17	0,17	0,67	22858	6104	25530	279350	44454	1963	122
14	0,33	0,33	0,33	12569	5783	11755	136039	37019	1642	73
1	1,00	0,00	0,00	9355	1503	16139	676618	21171	0	54
2	0,00	1,00	0,00	47869	4674	7694	481208	10050	80344	3779
3	0,00	0,00	1,00	12944	6112	10066	442178	21813	1106	135
4	0,33	0,67	0,00	33008	6098	5123	1069535	11684	47568	978
5	0,33	0,00	0,67	13394	6888	13675	131799	27998	216	26
6	0,00	0,33	0,67	17386	7154	17925	388854	31373	434	65
7	0,67	0,33	0,00	30457	110926	708	1457	2191	77	23
8	0,67	0,00	0,33	15940	6961	15367	290666	31378	458	43
9	0,00	0,67	0,33	25362	9177	13920	453911	24537	9244	384
10	0,33	0,33	0,33	28559	7283	11529	532995	26282	7865	376
11	0,67	0,17	0,17	12687	5678	10451	54429	24793	5517	257
12	0,17	0,67	0,17	68811	5756	11322	583303	18112	372936	31069
13	0,17	0,17	0,67	23087	6165	25785	282144	44899	1983	123
14	0,33	0,33	0,33	12695	5841	11873	137399	37389	1658	74

Tabela Suplementar 4 - Dados oriundos do planejamento experimental para anotações moleculares (168 horas de fermentação) usando a intensidade dos sinais como resposta.

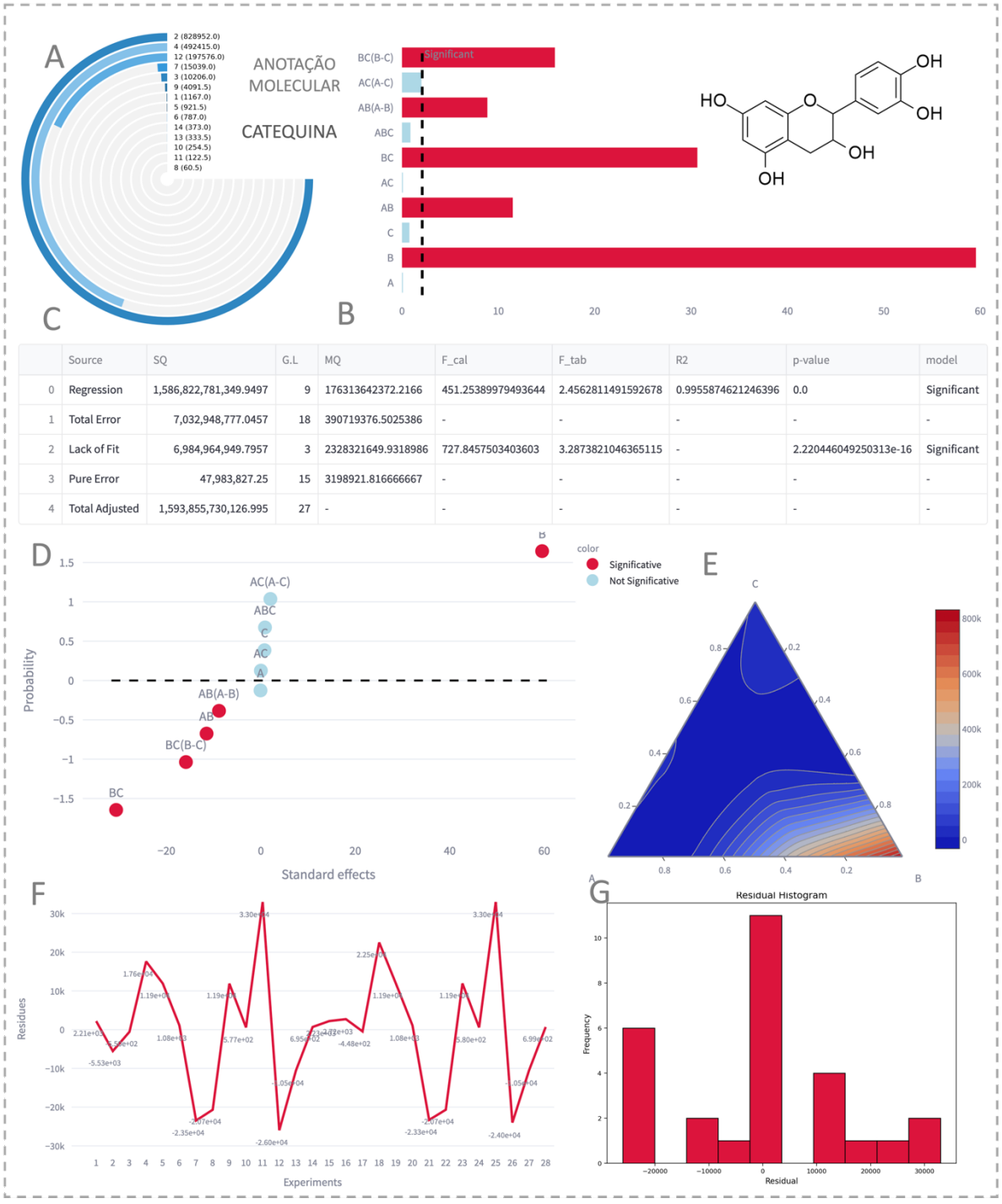
EXP	A	B	C	ADEN.	FENILA	TEOBR	CATEQ	PROCI.
1	1,00	0,00	0,00	4887	939	1174	785442	887
2	0,00	1,00	0,00	121950	15596	14801	1371960	54696
3	0,00	0,00	1,00	19053	11385	2476	299639	906
4	0,33	0,67	0,00	89757	8661	36771	1595248	58848
5	0,33	0,00	0,67	17882	11058	2380	593559	795
6	0,00	0,33	0,67	15579	7640	1616	140792	316
7	0,67	0,33	0,00	7821	2372	4334	485372	112
8	0,67	0,00	0,33	19318	9647	2406	762071	1261
9	0,00	0,67	0,33	30485	15846	2640	643362	522
10	0,33	0,33	0,33	22973	11761	2677	874445	1010
11	0,67	0,17	0,17	15488	5669	3140	591596	921
12	0,17	0,67	0,17	90077	15983	8599	1224443	35281
13	0,17	0,17	0,67	20823	10379	1899	609039	604
14	0,33	0,33	0,33	33061	12568	3562	903423	437
1	1,00	0,00	0,00	4936	948	1186	793296	896
2	0,00	1,00	0,00	123170	15752	14949	1385680	55243
3	0,00	0,00	1,00	19244	11499	2501	302635	915
4	0,33	0,67	0,00	90655	8748	37139	1611200	59436
5	0,33	0,00	0,67	18061	11169	2404	599495	803
6	0,00	0,33	0,67	15735	7716	1632	142200	319
7	0,67	0,33	0,00	7899	2396	4377	490226	113
8	0,67	0,00	0,33	19511	9743	2430	769692	1274
9	0,00	0,67	0,33	30790	16004	2666	649796	527
10	0,33	0,33	0,33	23203	11879	2704	883189	1020
11	0,67	0,17	0,17	15643	5726	3171	597512	930
12	0,17	0,67	0,17	90978	16143	8685	1236687	35634
13	0,17	0,17	0,67	21031	10483	1918	615129	610
14	0,33	0,33	0,33	33392	12694	3598	912457	441

Figura Suplementar 24 - Análise estatística (Trealose). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G - Histograma residual.



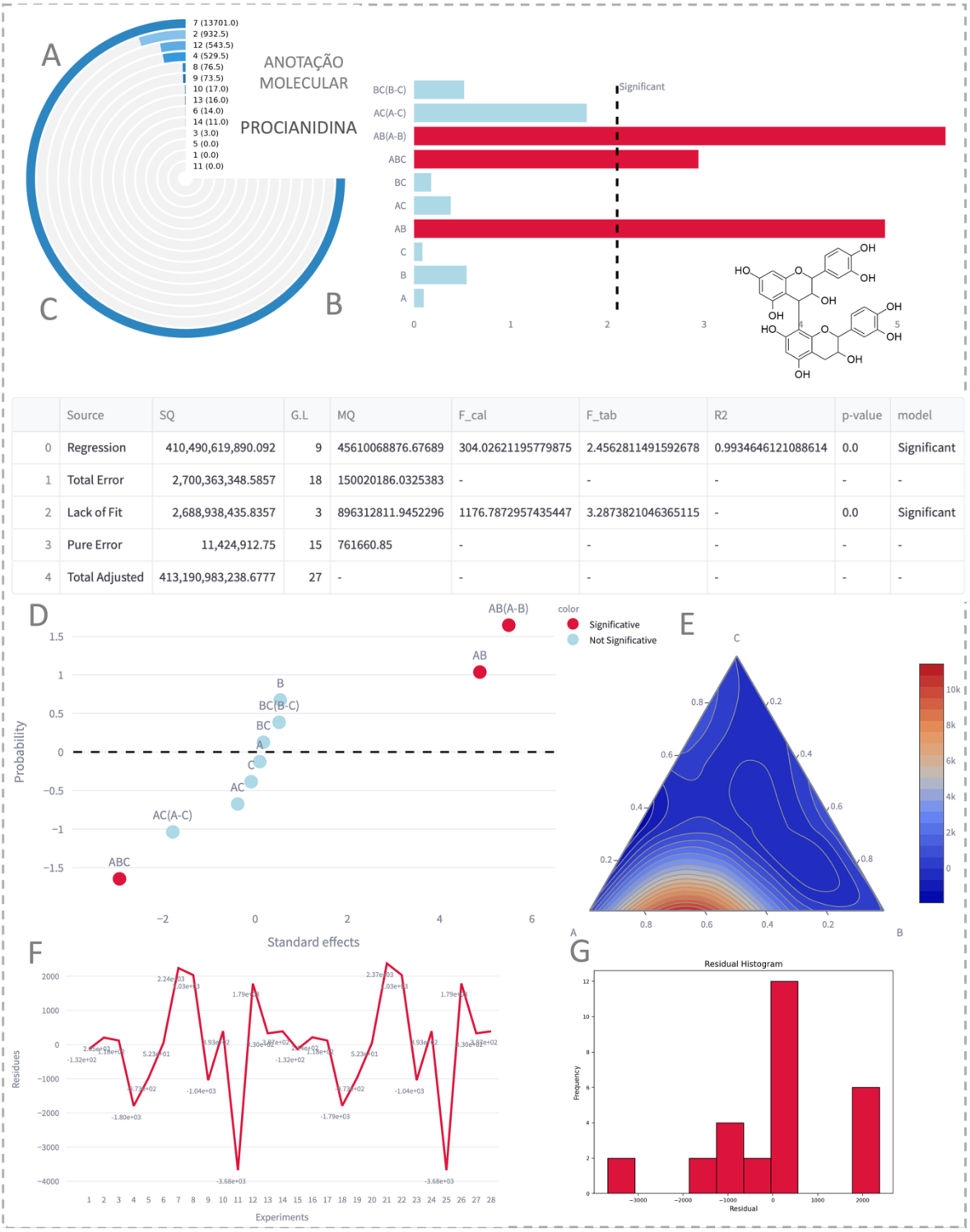
	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	11,353,618,782.4212	9	1261513198.046797	14.147236865695547	2.4562811491592678	0.8761397992340748	1.8397655404678304e-06	Significant
1	Total Error	1,605,065,199.686	18	89170288.87144281	-	-	-	-	-
2	Lack of Fit	397,130,077.186	3	132376692.3953235	1.643838604361678	3.2873821046365115	-	0.22150297356246273	Not Significant
3	Pure Error	1,207,935,122.5	15	80529008.16666667	-	-	-	-	-
4	Total Adjusted	12,958,683,982.1071	27	-	-	-	-	-	-

Figura Suplementar 25 - Análise estatística (Catequina). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G - Histograma residual.



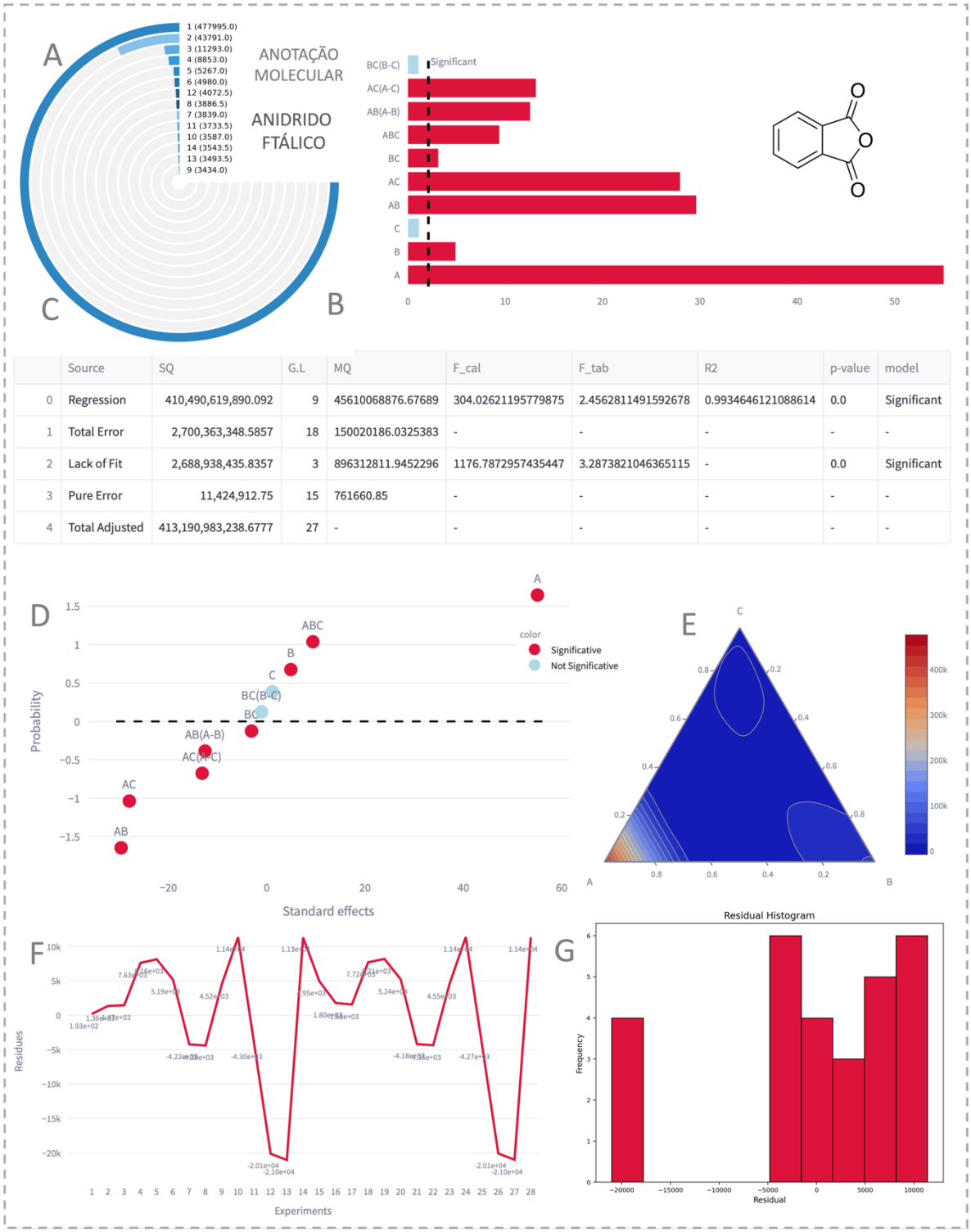
	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	1,586,822,781,349.9497	9	176313642372.2166	451.25389979493644	2.4562811491592678	0.9955874621246396	0.0	Significant
1	Total Error	7,032,948,777.0457	18	390719376.5025386	-	-	-	-	-
2	Lack of Fit	6,984,964,949.7957	3	2328321649.9318986	727.8457503403603	3.2873821046365115	-	2.220446049250313e-16	Significant
3	Pure Error	47,983,827.25	15	3198921.816666667	-	-	-	-	-
4	Total Adjusted	1,593,855,730,126.995	27	-	-	-	-	-	-

Figura Suplementar 26 - Análise estatística (Procianidina). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G - Histograma residual.



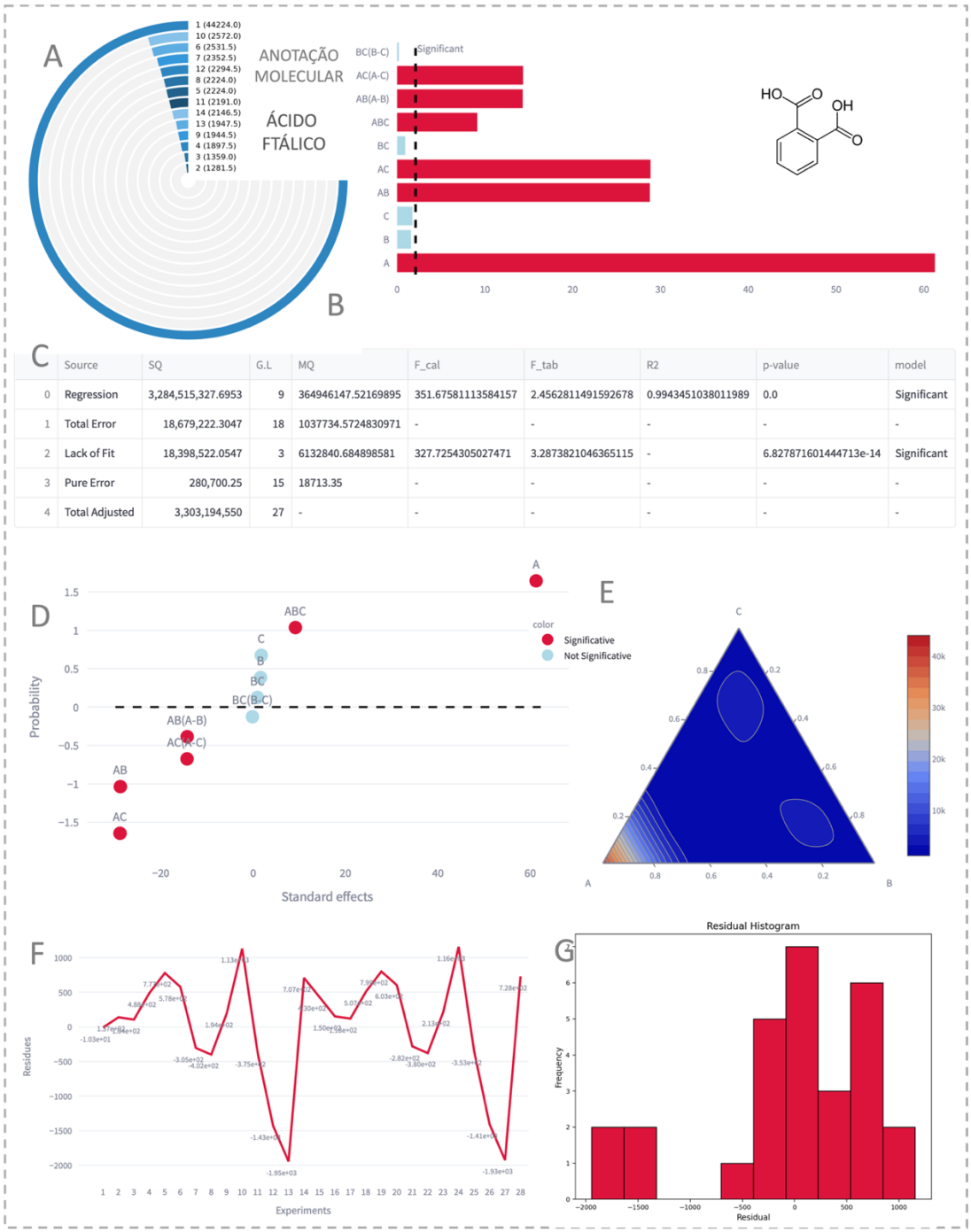
	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	410,490,619,890.092	9	45610068876.67689	304.02621195779875	2.4562811491592678	0.9934646121088614	0.0	Significant
1	Total Error	2,700,363,348.5857	18	150020186.0325383	-	-	-	-	-
2	Lack of Fit	2,688,938,435.8357	3	896312811.9452296	1176.7872957435447	3.2873821046365115	-	0.0	Significant
3	Pure Error	11,424,912.75	15	761660.85	-	-	-	-	-
4	Total Adjusted	413,190,983,238.6777	27	-	-	-	-	-	-

Figura Suplementar 27 - Análise estatística (Anidrido Ftálico). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G - Histograma residual.



	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	410,490,619,890.092	9	45610068876.67689	304.02621195779875	2.4562811491592678	0.9934646121088614	0.0	Significant
1	Total Error	2,700,363,348.5857	18	150020186.0325383	-	-	-	-	-
2	Lack of Fit	2,688,938,435.8357	3	896312811.9452296	1176.7872957435447	3.2873821046365115	-	0.0	Significant
3	Pure Error	11,424,912.75	15	761660.85	-	-	-	-	-
4	Total Adjusted	413,190,983,238.6777	27	-	-	-	-	-	-

Figura Suplementar 28 - Análise estatística (Ácido Ftálico). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G - Histograma residual.



	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	3,284,515,327.6953	9	364946147.52169895	351.67581113584157	2.4562811491592678	0.9943451038011989	0.0	Significant
1	Total Error	18,679,222.3047	18	1037734.5724830971	-	-	-	-	-
2	Lack of Fit	18,398,522.0547	3	6132840.684898581	327.7254305027471	3.2873821046365115	-	6.827871601444713e-14	Significant
3	Pure Error	280,700.25	15	18713.35	-	-	-	-	-
4	Total Adjusted	3,303,194,550	27	-	-	-	-	-	-

Figura Suplementar 29 - Análise estatística (Teobromina – 84 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G - Histograma residual.

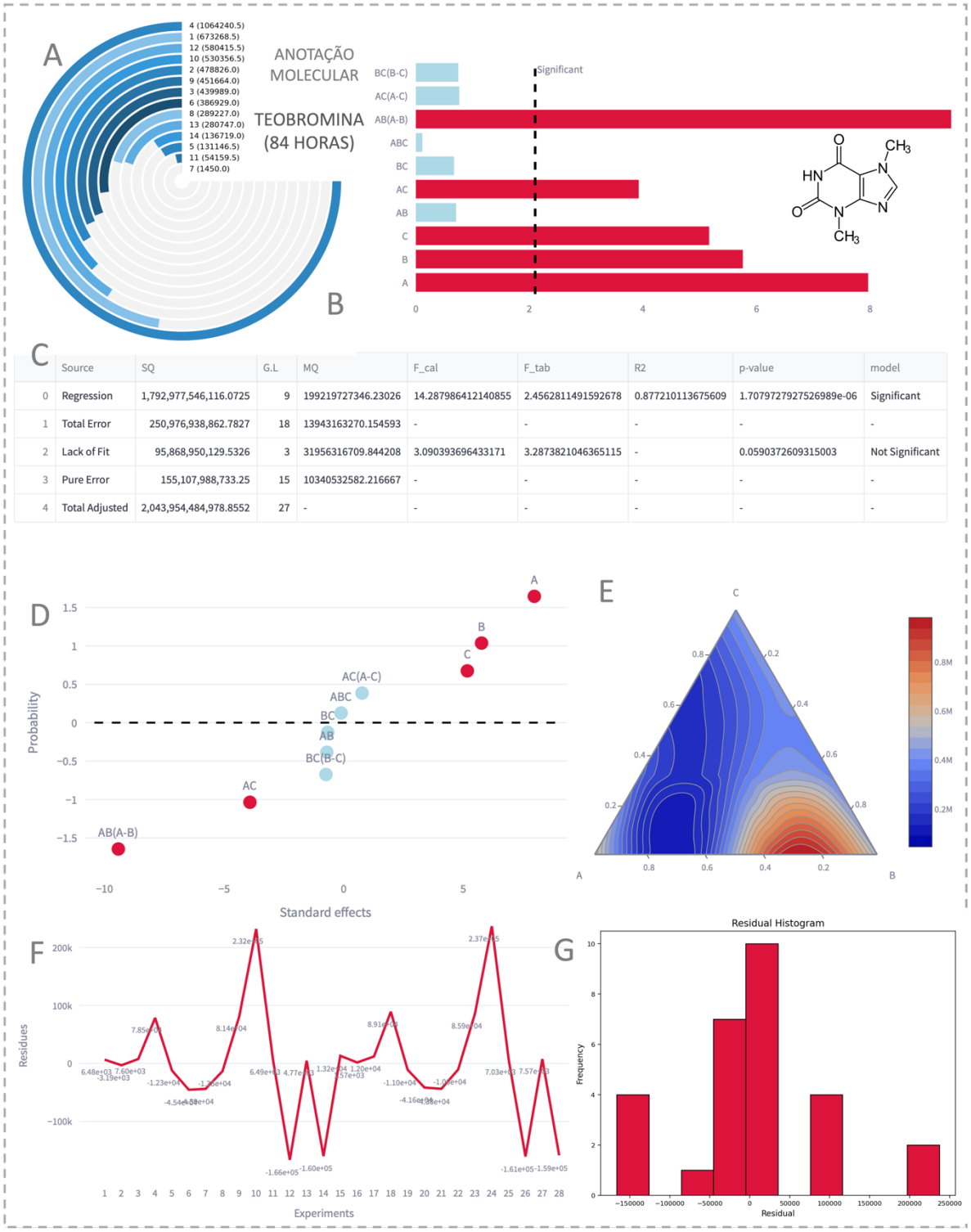
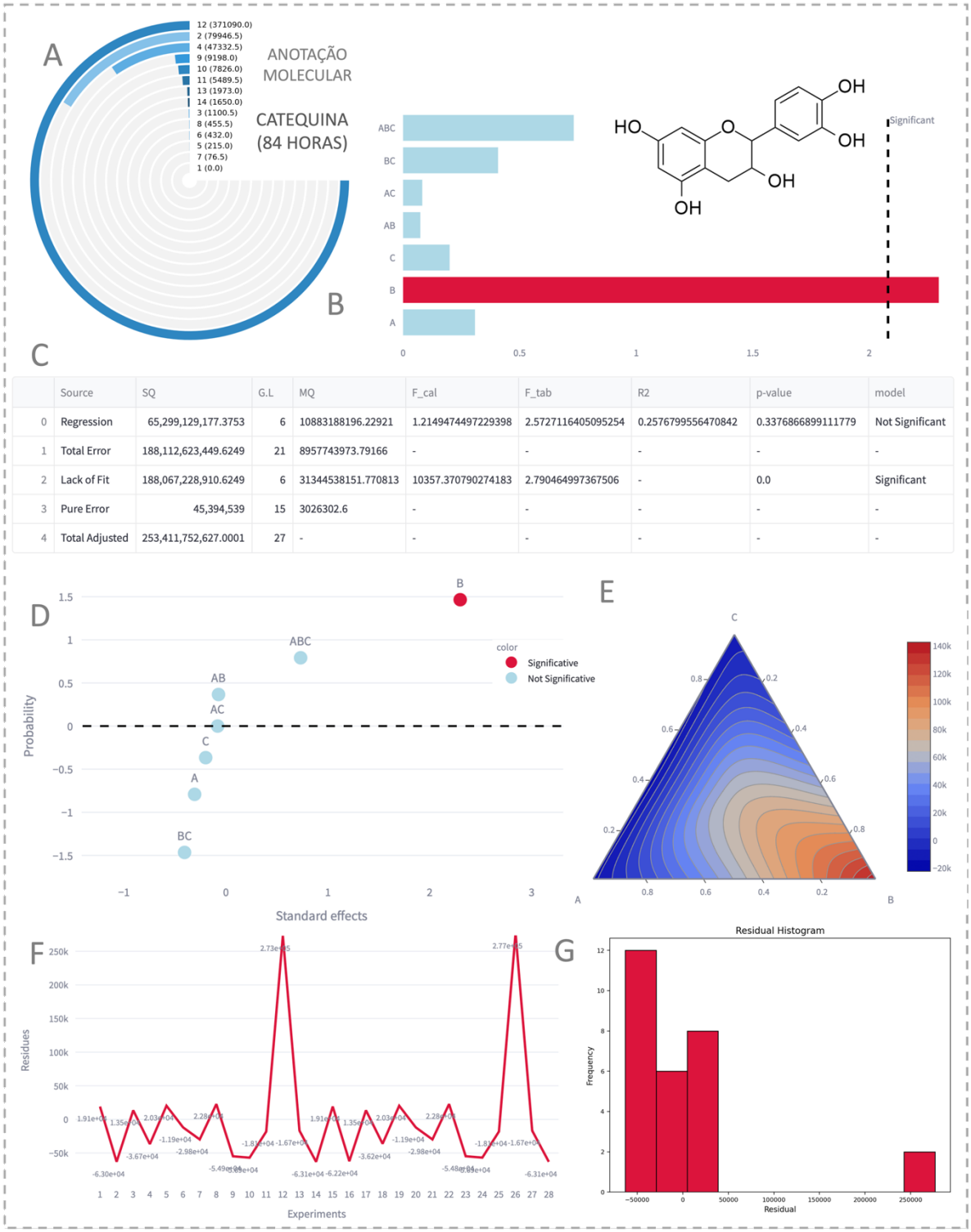
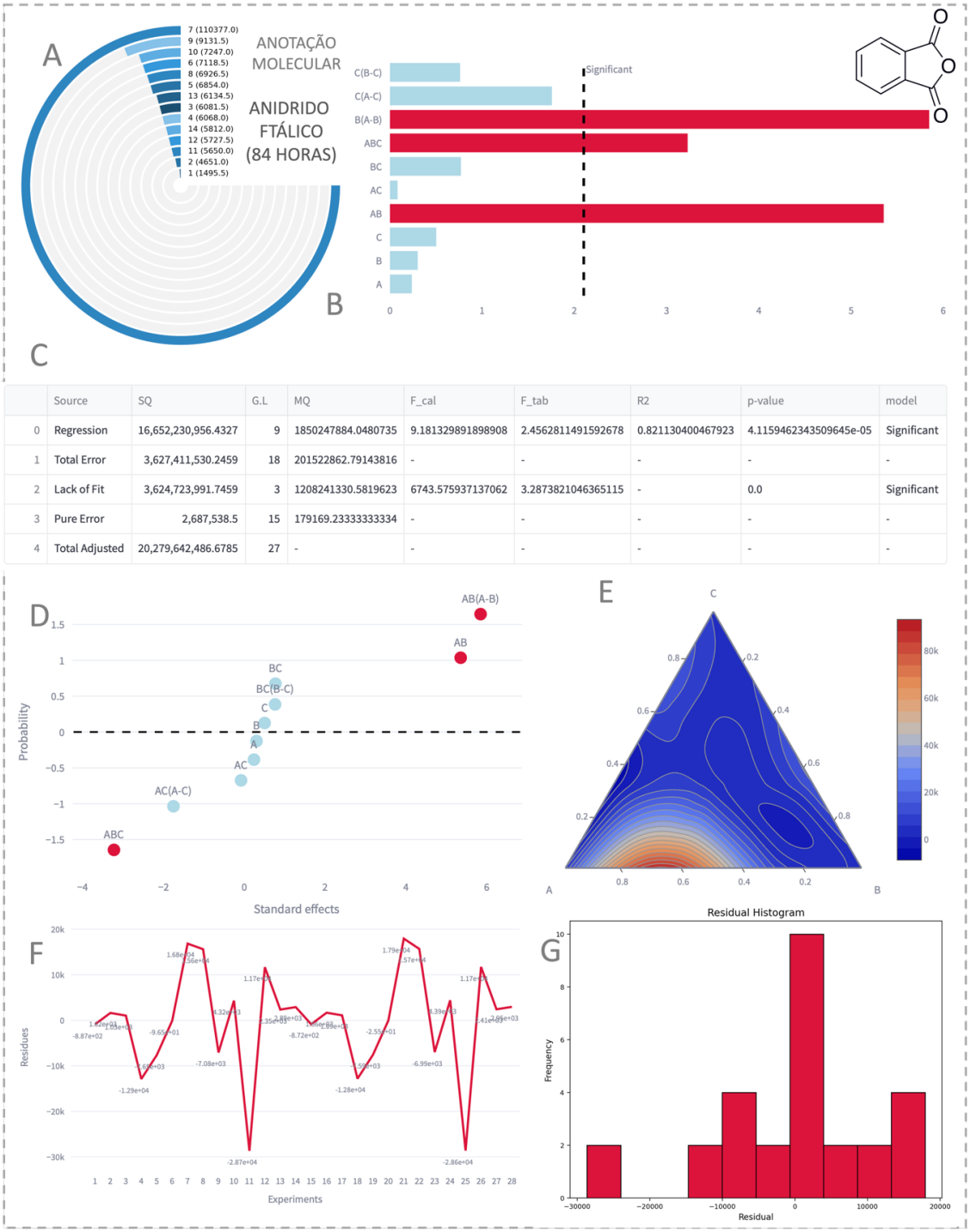


Figura Suplementar 30 - Análise estatística (Catequina – 84 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.



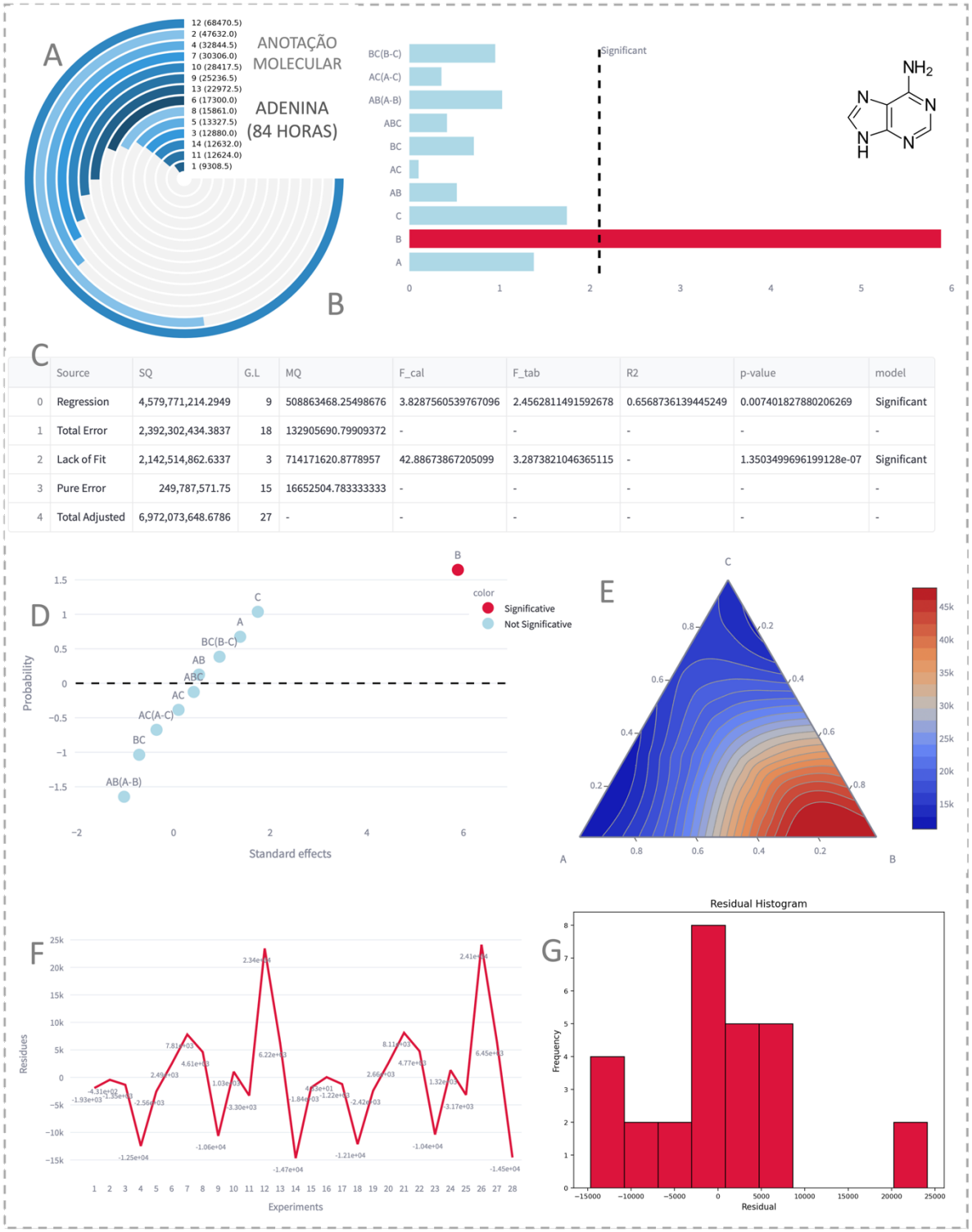
	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	65,299,129,177.3753	6	10883188196.22921	1.2149474497229398	2.5727116405095254	0.2576799556470842	0.3376866899111779	Not Significant
1	Total Error	188,112,623,449.6249	21	8957743973.79166	-	-	-	-	-
2	Lack of Fit	188,067,228,910.6249	6	31344538151.770813	10357.370790274183	2.790464997367506	-	0.0	Significant
3	Pure Error	45,394,539	15	3026302.6	-	-	-	-	-
4	Total Adjusted	253,411,752,627.0001	27	-	-	-	-	-	-

Figura Suplementar 31 - Análise estatística (Anidrido Ftálico – 84 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.



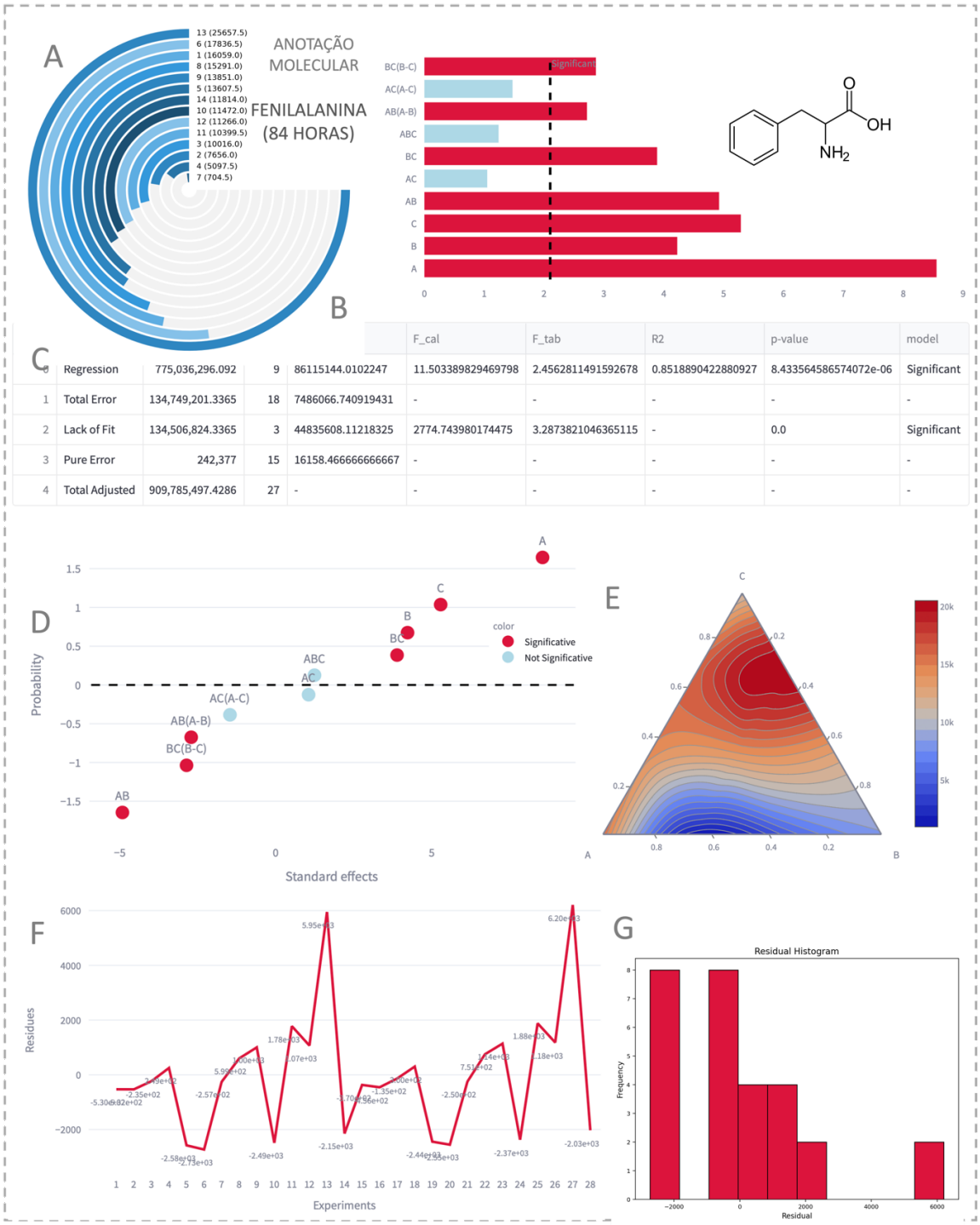
	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	16,652,230,956.4327	9	1850247884.0480735	9.181329891898908	2.4562811491592678	0.821130400467923	4.1159462343509645e-05	Significant
1	Total Error	3,627,411,530.2459	18	201522862.79143816	-	-	-	-	-
2	Lack of Fit	3,624,723,991.7459	3	1208241330.5819623	6743.575937137062	3.2873821046365115	-	0.0	Significant
3	Pure Error	2,687,538.5	15	179169.23333333334	-	-	-	-	-
4	Total Adjusted	20,279,642,486.6785	27	-	-	-	-	-	-

Figura Suplementar 32 - Análise estatística (Adenina – 84 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.



	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	4,579,771,214.2949	9	508863468.25498676	3.8287560539767096	2.4562811491592678	0.6568736139445249	0.007401827880206269	Significant
1	Total Error	2,392,302,434.3837	18	132905690.79909372	-	-	-	-	-
2	Lack of Fit	2,142,514,862.6337	3	714171620.8778957	42.88673867205099	3.2873821046365115	-	1.3503499696199128e-07	Significant
3	Pure Error	249,787,571.75	15	16652504.783333333	-	-	-	-	-
4	Total Adjusted	6,972,073,648.6786	27	-	-	-	-	-	-

Figura Suplementar 33 - Análise estatística (Fenilalanina – 84 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.



		F_cal	F_tab	R2	p-value	model
0	Regression 775,036,296.092 9 86115144.0102247	11.503389829469798	2.4562811491592678	0.8518890422880927	8.433564586574072e-06	Significant
1	Total Error 134,749,201.3365 18 7486066.740919431	-	-	-	-	-
2	Lack of Fit 134,506,824.3365 3 44835608.11218325	2774.743980174475	3.2873821046365115	-	0.0	Significant
3	Pure Error 242,377 15 16158.466666666667	-	-	-	-	-
4	Total Adjusted 909,785,497.4286 27 -	-	-	-	-	-

Figura Suplementar 34 - Análise estatística (Tirosina – 84 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.

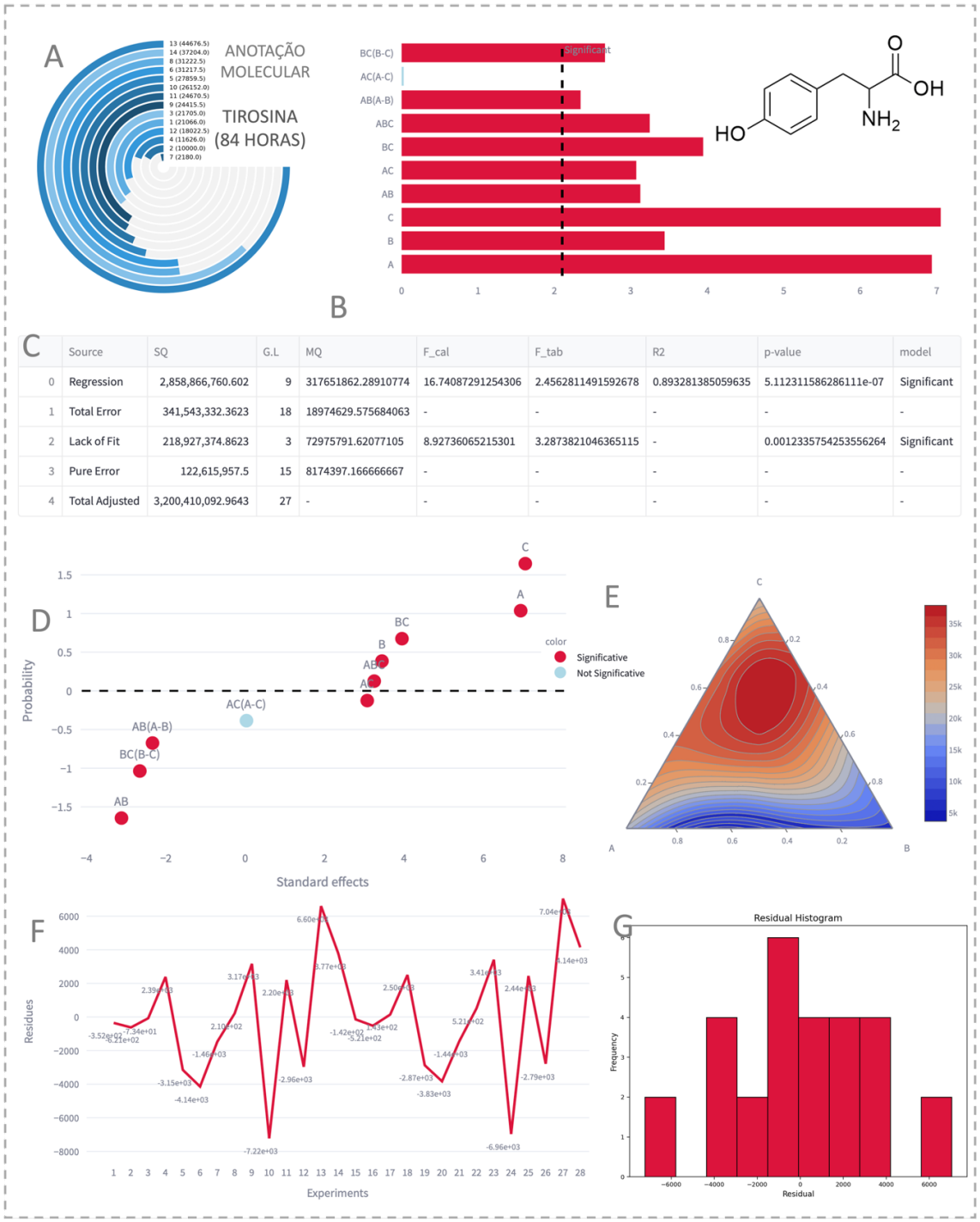
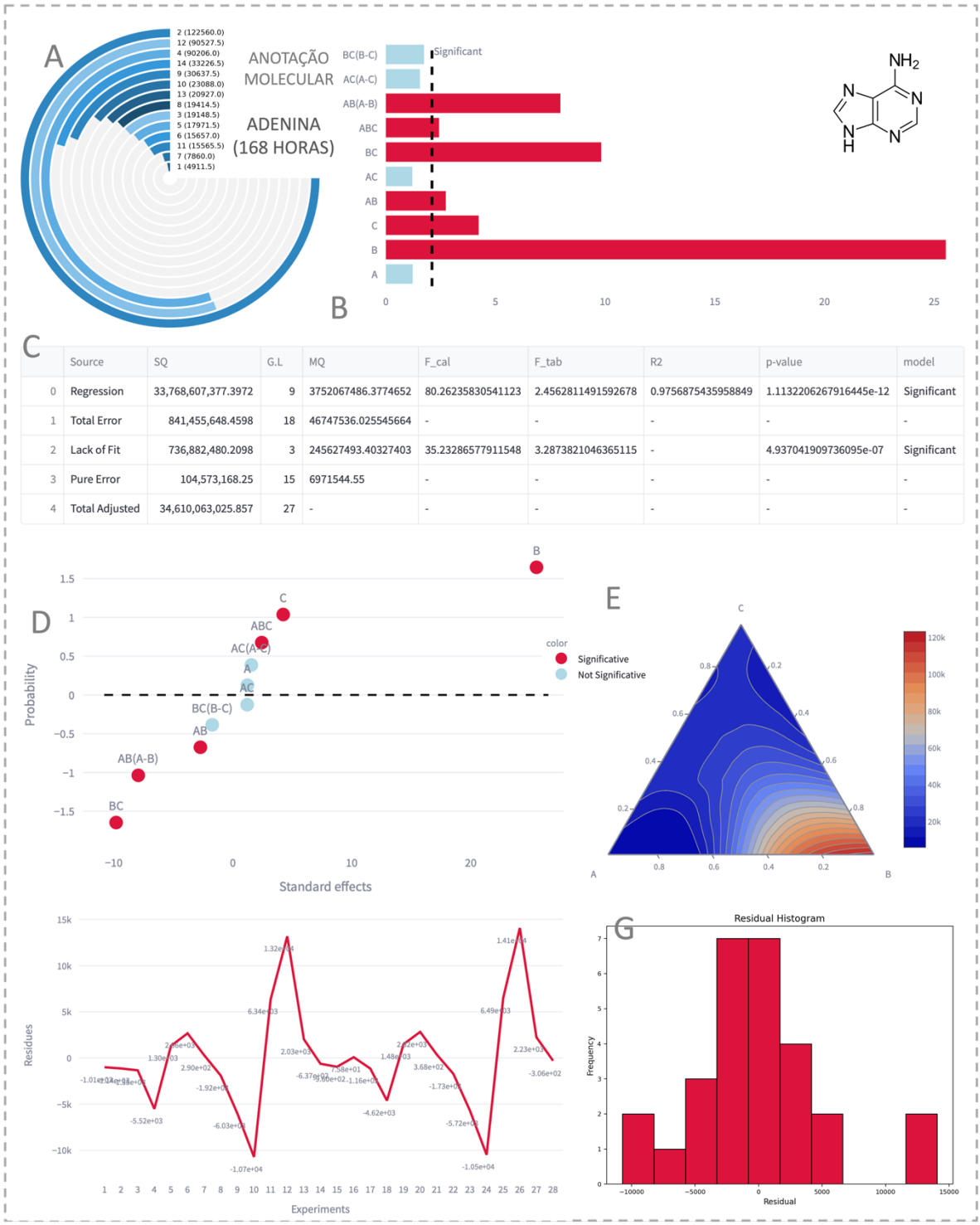
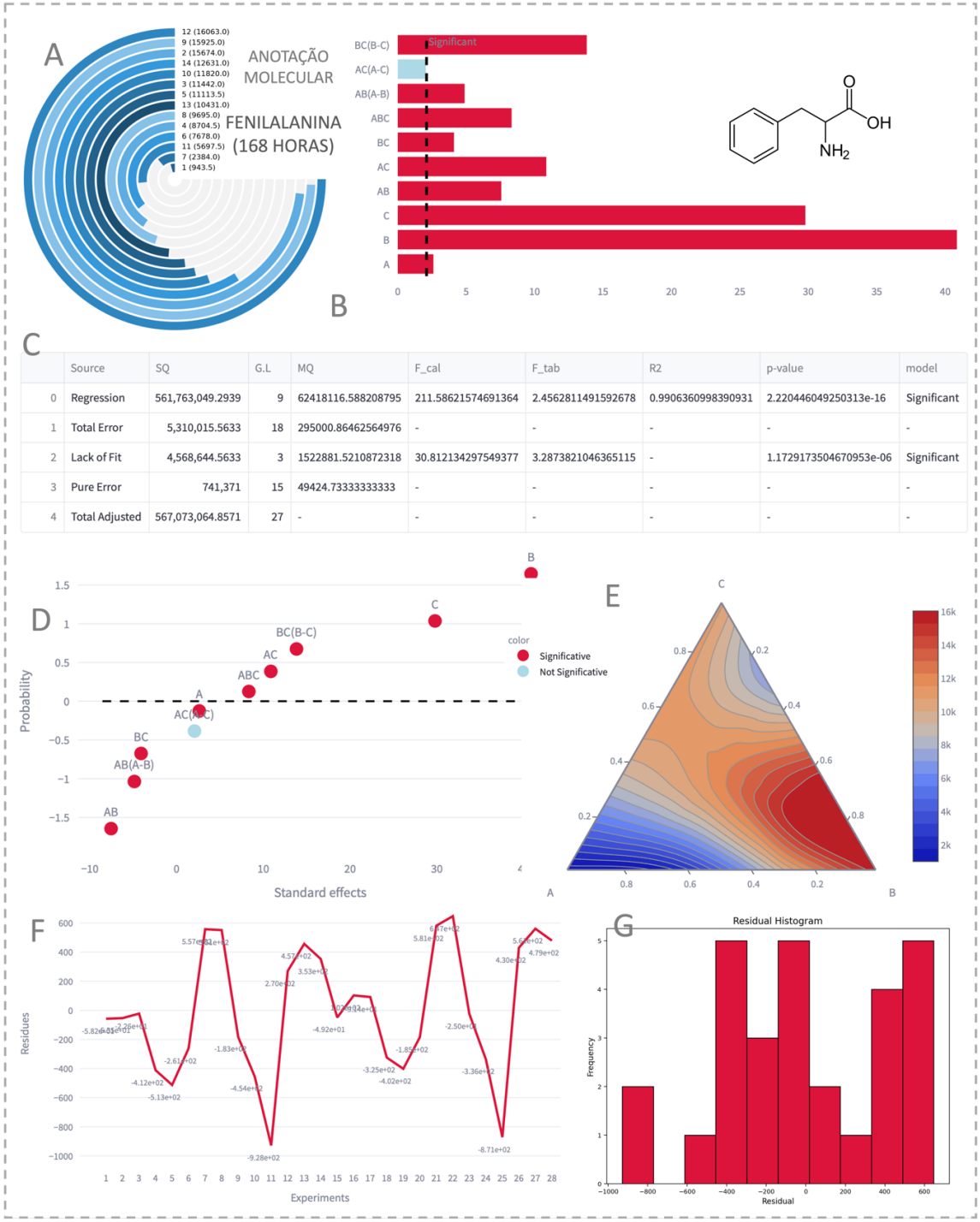


Figura Suplementar 35 - Análise estatística (Adenina – 168 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.



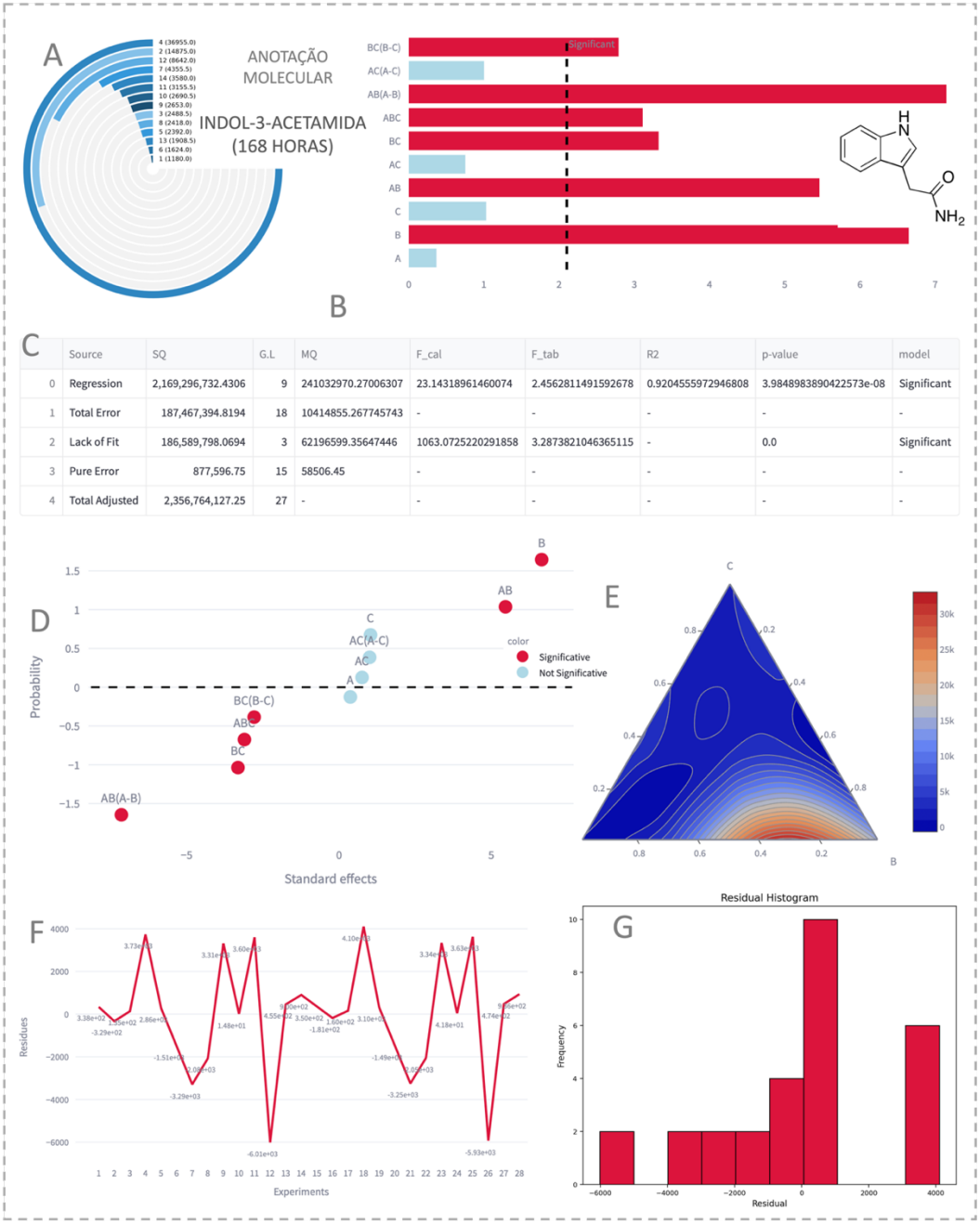
	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	33,768,607,377.3972	9	3752067486.3774652	80.26235830541123	2.4562811491592678	0.9756875435958849	1.1132206267916445e-12	Significant
1	Total Error	841,455,648.4598	18	46747536.025545664	-	-	-	-	-
2	Lack of Fit	736,882,480.2098	3	245627493.40327403	35.23286577911548	3.2873821046365115	-	4.937041909736095e-07	Significant
3	Pure Error	104,573,168.25	15	6971544.55	-	-	-	-	-
4	Total Adjusted	34,610,063,025.857	27	-	-	-	-	-	-

Figura Suplementar 36 - Análise estatística (Fenilalanina – 168 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.



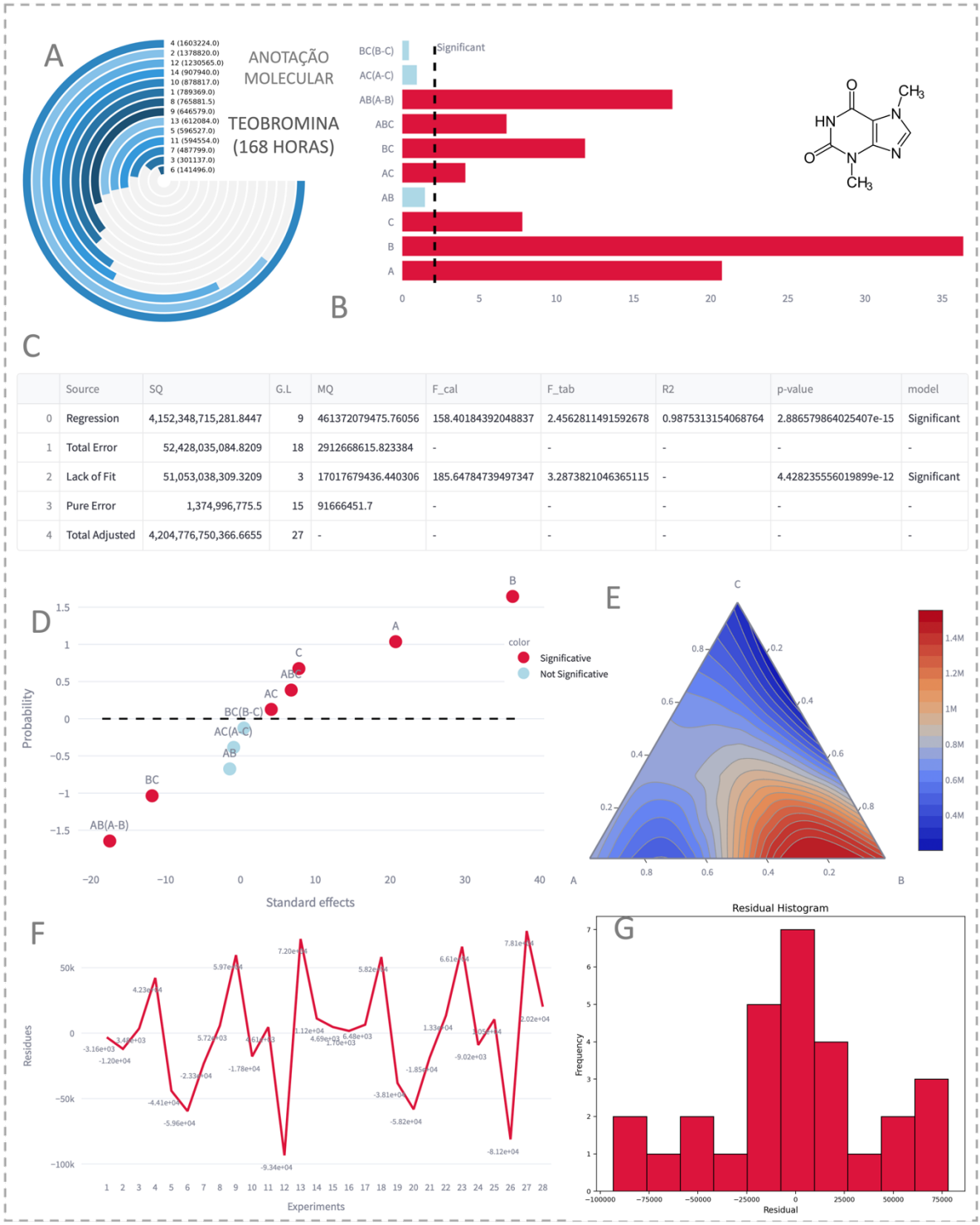
	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	561,763,049.2939	9	62418116.588208795	211.58621574691364	2.4562811491592678	0.9906360998390931	2.220446049250313e-16	Significant
1	Total Error	5,310,015.5633	18	295000.86462564976	-	-	-	-	-
2	Lack of Fit	4,568,644.5633	3	1522881.5210872318	30.812134297549377	3.2873821046365115	-	1.1729173504670953e-06	Significant
3	Pure Error	741,371	15	49424.73333333333	-	-	-	-	-
4	Total Adjusted	567,073,064.8571	27	-	-	-	-	-	-

Figura Suplementar 37 - Análise estatística (Indol-3-acetamida – 168 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.



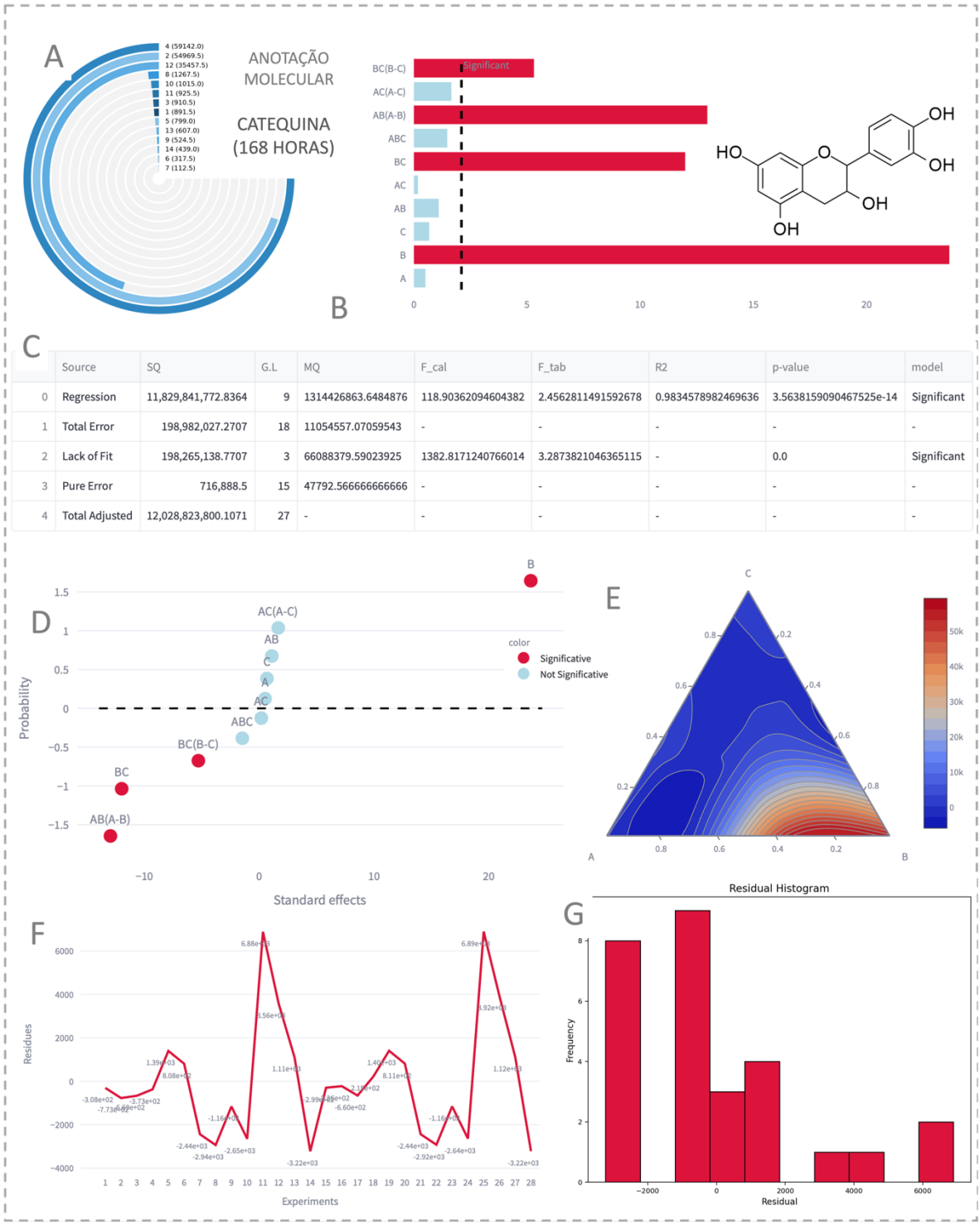
	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	2,169,296,732.4306	9	241032970.27006307	23.14318961460074	2.4562811491592678	0.9204555972946808	3.9848983890422573e-08	Significant
1	Total Error	187,467,394.8194	18	10414855.267745743	-	-	-	-	-
2	Lack of Fit	186,589,798.0694	3	62196599.35647446	1063.0725220291858	3.2873821046365115	-	0.0	Significant
3	Pure Error	877,596.75	15	58506.45	-	-	-	-	-
4	Total Adjusted	2,356,764,127.25	27	-	-	-	-	-	-

Figura Suplementar 38 - Análise estatística (Teobromina– 168 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.



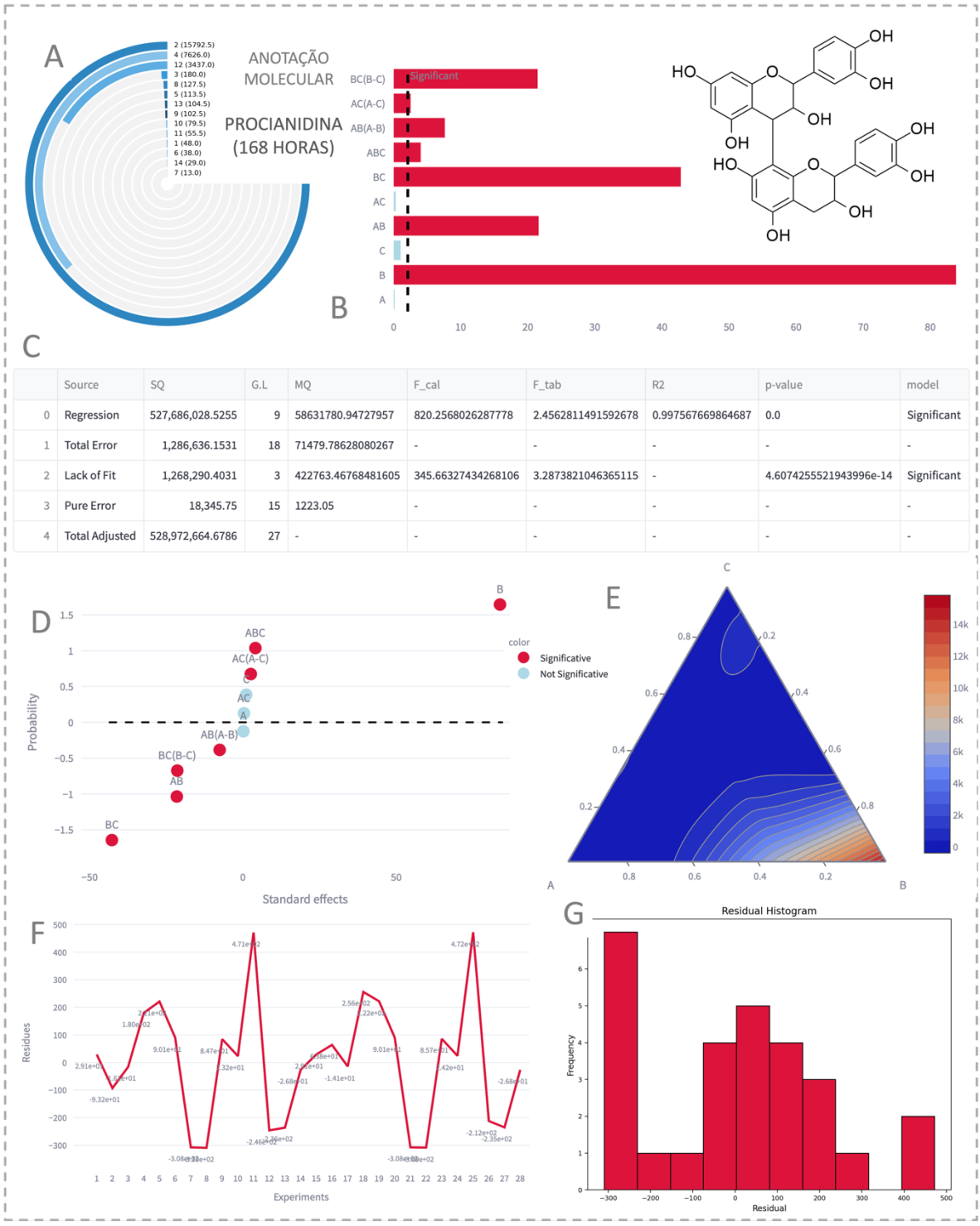
	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	4,152,348,715,281.8447	9	461372079475.76056	158.40184392048837	2.4562811491592678	0.9875313154068764	2.886579864025407e-15	Significant
1	Total Error	52,428,035,084.8209	18	2912668615.823384	-	-	-	-	-
2	Lack of Fit	51,053,038,309.3209	3	17017679436.440306	185.64784739497347	3.2873821046365115	-	4.428235556019899e-12	Significant
3	Pure Error	1,374,996,775.5	15	91666451.7	-	-	-	-	-
4	Total Adjusted	4,204,776,750,366.6655	27	-	-	-	-	-	-

Figura Suplementar 39 - Análise estatística (Catequina– 168 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.



	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	11,829,841,772.8364	9	1314426863.6484876	118.90362094604382	2.4562811491592678	0.9834578982469636	3.5638159090467525e-14	Significant
1	Total Error	198,982,027.2707	18	11054557.07059543	-	-	-	-	-
2	Lack of Fit	198,265,138.7707	3	66088379.59023925	1382.8171240766014	3.2873821046365115	-	0.0	Significant
3	Pure Error	716,888.5	15	47792.566666666666	-	-	-	-	-
4	Total Adjusted	12,028,823,800.1071	27	-	-	-	-	-	-

Figura Suplementar 40 - Análise estatística (Procianidina– 168 horas de fermentação). A – Distribuição da intensidade dos sinais nos ensaios. B – Gráfico de Pareto. C - ANOVA. D - Dispersão dos coeficientes padronizados. E - Mapa de contorno. F - Dispersão residual. G – Histograma residual.



	Source	SQ	G.L	MQ	F_cal	F_tab	R2	p-value	model
0	Regression	527,686,028.5255	9	58631780.94727957	820.2568026287778	2.4562811491592678	0.997567669864687	0.0	Significant
1	Total Error	1,286,636.1531	18	71479.78628080267	-	-	-	-	-
2	Lack of Fit	1,268,290.4031	3	422763.46768481605	345.66327434268106	3.2873821046365115	-	4.6074255521943996e-14	Significant
3	Pure Error	18,345.75	15	1223.05	-	-	-	-	-
4	Total Adjusted	528,972,664.6786	27	-	-	-	-	-	-

MATERIAL SUPLEMENTAR D – CAPÍTULO 5

Tabela Suplementar 5 - Planejamento de Misturas utilizando como resposta a quantidade de sinais moleculares em tempos diferentes de fermentação do cacau.

EXP	A	B	C	0 H	84 H	168 H
1	1,00	0,00	0,00	208	148	167
2	0,00	1,00	0,00	299	114	182
3	0,00	0,00	1,00	160	125	168
4	0,33	0,67	0,00	186	157	191
5	0,33	0,00	0,67	171	155	165
6	0,00	0,33	0,67	163	172	170
7	0,67	0,33	0,00	196	199	172
8	0,67	0,00	0,33	168	160	174
9	0,00	0,67	0,33	165	166	176
10	0,33	0,33	0,33	166	163	184
11	0,67	0,17	0,17	165	165	173
12	0,17	0,67	0,17	171	182	189
13	0,17	0,17	0,67	160	180	167
14	0,33	0,33	0,33	167	165	169
1	1,00	0,00	0,00	210	151	170
2	0,00	1,00	0,00	300	116	186
3	0,00	0,00	1,00	162	128	171
4	0,33	0,67	0,00	180	160	195
5	0,33	0,00	0,67	178	158	168
6	0,00	0,33	0,67	150	175	173
7	0,67	0,33	0,00	190	203	175
8	0,67	0,00	0,33	168	163	177
9	0,00	0,67	0,33	154	169	180
10	0,33	0,33	0,33	154	166	188
11	0,67	0,17	0,17	169	168	176
12	0,17	0,67	0,17	180	186	193
13	0,17	0,17	0,67	170	184	170
14	0,33	0,33	0,33	155	168	172

MATERIAL SUPLEMENTAR E – CAPÍTULO 6

Tabela Suplementar 6 - Decomposição residual referente ao complexo formado entre a proteína 1NC6 e ligante Procianidina.

RESÍDUO	VALOR PADRÃO 1NC6 kcal/mol	VALOR LIGANTE PROCIANIDINA kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
TRP:215	-2,22	-1,81	0,41	Comum	H bond	vdW
GLN:192	-1,88	-1,40	0,48	Comum	Pi-donor .H.B*	Pi-donor H.B
CYS:220	-1,49	-0,10	1,40	Comum	H bond	vdW
CYS:191	-1,35	-0,87	0,48	Comum	vdW	vdW
GLY:226	-0,96	----	----	Padrão	vdW	vdW
GLY:216	-0,95	-0,43	0,52	Comum	vdW	H bond
HIS:57	-0,64	-2,00	-1,36	Comum	vdW	Pi-pi-t-shaped
SER:195	-0,60	-0,42	0,19	Comum	vdW	H bond
SER:217	-0,55	----	----	Padrão	vdW	vdW
VAL:213	-0,54	-0,39	0,15	Comum	vdW	vdW
SER:214	-0,41	-1,07	-0,66	Comum	vdW	vdW
LEU:99	-0,40	-0,67	-0,27	Comum	vdW	Pi-alkyl
GLY:219	-0,36	-0,13	0,23	Comum	H bond	H bond
GLY:193	-0,29	-0,27	0,02	Comum	vdW	vdW
ALA:221: A	-0,22	----	----	Padrão	vdW	vdW
SER:190	-0,20	-0,53	-0,33	Comum	vdW	H bond
ASP:189	-0,08	----	----	Padrão	H bond	vdW
ASP:194	0,03	-0,17	-0,20	Comum	vdW	vdW
TYR:228	0,20	----	----	Padrão	vdW	vdW
SOMA	**-12,91**	**-10,25**				

Tabela Suplementar 7 - Decomposição residual referente ao complexo formado entre a proteína 4DD8 e ligante catequina.

RESÍDUO	VALOR PADRÃO 4DD8 kcal/mol	VALOR LIGANTE CATEQUINA kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
ILE:368	-1,72	-0,95	0,77	Comum	pi-donor H-Bond	vdW
GLY:366	-1,37	-0,30	1,07	Comum	vdW	vdW
VAL:301	-1,24	-1,75	-0,51	Comum	pi-alkyl	H Bond
THR:300	-0,88	-1,59	-0,71	Comum	H Bond	C-H Bond
HIS:344	-0,68	-0,19	0,49	Comum	H Bond	vdW
THR:299	-0,66	-0,76	-0,10	Comum	H Bond	Unf. Donor-donor
HIS:334	-0,61	-0,20	0,41	Comum	H Bond	Pi-Pi Stacked
SER:367	-0,55	-0,34	0,21	Comum	vdW	vdW
THR:331	-0,39	-0,94	-0,55	Comum	vdW	H Bond
PHE:303	-0,12			Padrão	pi-alkyl	vdW
HIS:321	-0,07			Padrão	vdW	vdW
GLY:302	-0,01	-0,48	-0,46	Comum	H Bond	vdW
ALA:365	0,00	0,05	0,05	Comum	vdW	vdW
HIS:338	0,08			Padrão	H-Bond/pi-alkyl	vdW
GLU:335	1,45	-3,56	-5,01	Comum	C-H Bond	vdW
ILE:363		0,01		Ligante	vdW	Pi-Alkyl
PRO:373		0,00		Ligante		vdW
SOMA	**-6,78**	**-11,01**				

Tabela Suplementar 8 - Decomposição residual referente ao complexo formado entre a proteína 4DD8 e ligante trealose.

RESÍDUO	VALOR PADRÃO 4DD8 kcal/mol	VALOR LIGANTE TREALOSE kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
ILE:368	-1,72	-1,50	0,22	Comum	pi-donor H-Bond	vdW
GLY:366	-1,37	-0,17	1,20	Comum	vdW	H-Bond/C-Hbond
VAL:301	-1,24	-1,02	0,22	Comum	pi-alkyl	vdW
THR:300	-0,88	-0,53	0,35	Comum	H Bond	C-H Bond
HIS:344	-0,68	-0,01	0,67	Comum	H Bond	vdW
THR:299	-0,66	-0,62	0,03	Comum	H Bond	C-H Bond
HIS:334	-0,61	-0,10	0,51	Comum	H Bond	Pi-Sigma
SER:367	-0,55	-0,14	0,41	Comum	vdW	C-H Bond
THR:331	-0,39	-0,55	-0,15	Comum	vdW	H Bond
PHE:303	-0,12	-0,01	0,11	Comum	pi-alkyl	vdW
HIS:321	-0,07	----	----	Padrão	vdW	vdW
GLY:302	-0,01	-0,01	0,01	Comum	H Bond	vdW
ALA:365	0,00	-0,17	-0,17	Comum	vdW	vdW
HIS:338	0,08	----	----	Padrão	H-Bond/pi-alkyl	vdW
GLU:335	1,45	0,21	-1,24	Comum	C-H Bond	H Bond
GLY:369	-----	0,04	----	Ligante	vdW	vdW
SOMA	**-6,78**	**-4,59**				

Tabela Suplementar 9 - Decomposição residual referente ao complexo formado entre a proteína 6VVU e ligante ácido ftálico.

RESÍDUO	VALOR PADRÃO 6VVU kcal/mol	VALOR LIGANTE ÁC. FTÁLICO kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
GLN:192	-3,98	-0,39	3,60	Comum	H Bond*	H Bond*
TRP:215	-2,63	-0,65	1,98	Comum	vdW	Amide-Pi Stack.
SER:195	-1,78	-0,10	1,68	Comum	H Bond	vdW
GLY:216	-1,62	-0,95	0,67	Comum	H Bond	C-H Bond
GLY:219	-1,51	-0,03	1,49	Comum	vdW	vdW
CYS:191	-1,21	-0,89	0,32	Comum	C-H Bond	Amida-Pi Stack
GLY:193	-1,14	-----	-----	Padrão	vdW	vdW
VAL:213	-0,87	-1,25	-0,38	Comum	vdW	vdW
SER:214	-0,85	-1,04	-0,19	Comum	Unf. D-Donor[1]	Amide-Pi Stack
CYS:220	-0,79	-0,69	0,10	Comum	vdW	C-H Bond
SER:190	-0,49	-1,37	-0,87	Comum	vdW	H Bond*
GLU:217	-0,34	-----	-----	Padrão	vdW	vdW
ASP:194	-0,30	-----	-----	Padrão	vdW	vdW
GLY:226	-0,22	-0,65	-0,43	Comum	vdW	C-H Bond*
HIS:57	-0,16	0,02	0,18	Comum	H-Bond/pi-alkyl	vdW
ILE:227	-0,08	-1,16	-1,08	Comum	H Bond	vdW
ALA:97	-0,07	-----	-----	Padrão	vdW	vdW
CYS:42	-0,06	-----	-----	Padrão	vdW	vdW
GLN:98	-0,06	-----	-----	Padrão	vdW	vdW
ALA:221	-0,04	-0,06	-0,03	Comum	vdW	vdW
TYR:228	0,09	-1,31	-1,40	Comum	vdW	vdW
ASP:189	0,34	-0,92	-1,27	Comum	vdW	H Bond*
SOMA	**-17,78**	**-11,44**				

1 - Unf. D-Donor1 = Doador-doador desfavorável

* realiza ligações de hidrogênio com duas regiões do ligante

Tabela Suplementar 10 - Decomposição residual referente ao complexo formado entre a proteína 6VVU e ligante adenina.

RESÍDUO	VALOR PADRÃO 6VVU kcal/mol	VALOR LIGANTE ADENINA kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
GLN:192	-3,98	-0,46	3,53	Comum	H Bond*	H Bond*
TRP:215	-2,63	-0,40	2,23	Comum	vdW	Amide-Pi Stacked
SER:195	-1,78	-0,06	1,71	Comum	H Bond	C-H Bond
GLY:216	-1,62	-0,25	1,37	Comum	H Bond	vdW
GLY:219	-1,51	-0,17	1,34	Comum	vdW	vdW
CYS:191	-1,21	-0,40	0,81	Comum	C-H Bond	Amide-Pi Stacked
GLY:193	-1,14	-----	-----	Padrão	vdW	H Bond
VAL:213	-0,87	-0,10	0,76	Comum	vdW	Pi-Alkyl
SER:214	-0,85	-0,10	0,75	Comum	Unf. D-Donor	C-H/Am-Pi Stac.
CYS:220	-0,79	-0,17	0,62	Comum	vdW	vdW
SER:190	-0,49	-0,35	0,14	Comum	vdW	C-H Bond
GLU:217	-0,34	-----	-----	Padrão	vdW	vdW
ASP:194	-0,30	-----	-----	Padrão	vdW	vdW
GLY:226	-0,22	-0,03	0,18	Comum	vdW	vdW
HIS:57	-0,16	-0,03	0,13	Comum	H-B/pi-alkyl	vdW
ILE:227	-0,08	-0,04	0,05	Comum	H Bond	vdW
ALA:97	-0,07	-----	-----	Padrão	vdW	vdW
CYS:42	-0,06	-----	-----	Padrão	vdW	vdW
GLN:98	-0,06	-----	-----	Padrão	vdW	vdW
ALA:221	-0,04	-0,02	0,02	Comum	vdW	vdW
TYR:228	0,09	-0,02	-0,11	Comum	vdW	vdW
ASP:189	0,34	-0,02	-0,37	Comum	vdW	vdW
SOMA	**-17,78**	**-2,63**				

Tabela Suplementar 11 - Decomposição residual referente ao complexo formado entre a proteína 6VVU e ligante anidrido ftálico.

RESÍDUO	VALOR PADRÃO 6VVU kcal/mol	VALOR LIGANTE AN. FTÁLICO kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
GLN:192	-3,98	-0,74	3,24	Comum	H Bond*	vdW
TRP:215	-2,63	-0,67	1,96	Comum	vdW	vdW
SER:195	-1,78	-0,30	1,48	Comum	H Bond	vdW
GLY:216	-1,62	-0,51	1,11	Comum	H Bond	vdW
GLY:219	-1,51	-0,15	1,37	Comum	vdW	vdW
CYS:191	-1,21	-0,89	0,33	Comum	C-H Bond	Amida-Pi Stac
GLY:193	-1,14	-----	-----	Padrão	vdW	H Bond
VAL:213	-0,87	-0,46	0,41	Comum	vdW	Pi-Alkyl
SER:214	-0,85	-0,34	0,51	Comum	Unf. D-D*	H Bond
CYS:220	-0,79	-0,21	0,58	Comum	vdW	vdW
SER:190	-0,49	-0,96	-0,46	Comum	vdW	vdW
GLU:217	-0,34	-----	-----	Padrão	vdW	vdW
ASP:194	-0,30	-----	-----	Padrão	vdW	vdW
GLY:226	-0,22	0,04	0,26	Comum	vdW	C-H Bond
HIS:57	-0,16	-0,16	-0,01	Comum	H-B/pi-alkyl	vdW
ILE:227	-0,08	-0,05	0,03	Comum	H Bond	vdW
ALA:97	-0,07	-----	-----	Padrão	vdW	vdW
CYS:42	-0,06	-----	-----	Padrão	vdW	vdW
GLN:98	-0,06	-----	-----	Padrão	vdW	vdW
ALA:221	-0,04	-0,04	0,00	Comum	vdW	vdW
TYR:228	0,09	-0,31	-0,40	Comum	vdW	vdW
ASP:189	0,34	0,07	-0,28	Comum	vdW	vdW
SOMA	-17,78	-5,66				

*Doador-Doador desfavorável

Tabela Suplementar 12 - Decomposição residual referente ao complexo formado entre a proteína 6VVU e ligante catequina.

RESÍDUO	VALOR PADRÃO 6VVU kcal/mol	VALOR LIGANTE CATEQUINA kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
GLN:192	-3,98	-1,26	2,72	Comum	H Bond*	vdW
TRP:215	-2,63	-2,27	0,36	Comum	vdW	H Bond
SER:195	-1,78	-0,86	0,92	Comum	H Bond	H Bond
GLY:216	-1,62	-1,35	0,27	Comum	H Bond	H Bond
GLY:219	-1,51	-0,21	1,30	Comum	vdW	vdW
CYS:191	-1,21	-2,23	-1,01	Comum	C-H Bond	vdW
GLY:193	-1,14	-0,57	0,57	Comum	vdW	Unf. D-D
VAL:213	-0,87	-1,07	-0,21	Comum	vdW	Alkyl
SER:214	-0,85	-1,06	-0,21	Comum	Unf. D-D	Unf. D-D
CYS:220	-0,79	-1,27	-0,48	Comum	vdW	Pi-Sulfur
SER:190	-0,49	-1,24	-0,75	Comum	vdW	H Bond*
GLU:217	-0,34	-----	-----	Padrão	vdW	vdW
ASP:194	-0,30	-0,67	-0,37	Comum	vdW	Unf. D-D
GLY:226	-0,22	-0,08	0,14	Comum	vdW	vdW
HIS:57	-0,16	-0,01	0,15	Comum	H-Bond/pi-alkyl	vdW
ILE:227	-0,08	-0,74	-0,66	Comum	H Bond	H Bond
ALA:97	-0,07	-0,07	0,01	Comum	vdW	vdW
CYS:42	-0,06	-----	-----	Padrão	vdW	vdW
GLN:98	-0,06	-0,17	-0,11	Comum	vdW	vdW
ALA:221	-0,04	-0,43	-0,40	Comum	vdW	vdW
TYR:228	0,09	-0,64	-0,74	Comum	vdW	vdW
ASP:189	0,34	-0,85	-1,19	Comum	vdW	H Bond
SOMA	-17,78	-17,06				

Tabela Suplementar 13 - Decomposição residual referente ao complexo formado entre a proteína 6VVU e ligante Indol-3-Acetamida.

RESÍDUO	VALOR PADRÃO 6VVU kcal/mol	VALOR LIGANTE INDOL-3-AC. kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
GLN:192	-3,98	-1,62	2,36	Comum	H Bond*	vdW
TRP:215	-2,63	-1,78	0,85	Comum	vdW	Pi-Sigma
SER:195	-1,78	-0,23	1,54	Comum	H Bond	H Bond
GLY:216	-1,62	-1,11	0,51	Comum	H Bond	vdW
GLY:219	-1,51	-0,26	1,25	Comum	vdW	vdW
CYS:191	-1,21	-1,55	-0,33	Comum	C-H Bond	vdW
GLY:193	-1,14	-----	-----	Padrão	vdW	vdW
VAL:213	-0,87	-0,83	0,04	Comum	vdW	vdW
SER:214	-0,85	-0,67	0,19	Comum	Unf. D-D	vdW
CYS:220	-0,79	-0,62	0,18	Comum	vdW	vdW
SER:190	-0,49	-0,83	-0,34	Comum	vdW	H Bond
GLU:217	-0,34	-----	-----	Padrão	vdW	vdW
ASP:194	-0,30	-0,25	0,05	Comum	vdW	vdW
GLY:226	-0,22	-0,83	-0,61	Comum	vdW	C-H BOND
HIS:57	-0,16	-0,07	0,08	Comum	H-B/pi-alkyl	vdW
ILE:227	-0,08	-0,80	-0,72	Comum	H Bond	vdW
ALA:97	-0,07	-----	-----	Padrão	vdW	vdW
CYS:42	-0,06	-----	-----	Padrão	vdW	vdW
GLN:98	-0,06	-----	-----	Padrão	vdW	vdW
ALA:221	-0,04	-----	-----	Padrão	vdW	vdW
TYR:228	0,09	-----	-----	Padrão	vdW	vdW
ASP:189	0,34	0,51	0,17	Comum	vdW	H Bond
SOMA	**-17,78**	**-10,92**				

Tabela Suplementar 14 - Decomposição residual referente ao complexo formado entre a proteína 6VVU e ligante Teobromina.

RESÍDUO	VALOR PADRÃO 6VVU kcal/mol	VALOR LIGANTE TEOBROMINA kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
GLN:192	-3,98	-1,28	2,70	Comum	H Bond*	vdW
TRP:215	-2,63	-0,30	2,33	Comum	vdW	Pi-Sigma
SER:195	-1,78	-0,09	1,69	Comum	H Bond	H Bond
GLY:216	-1,62	-0,68	0,94	Comum	H Bond	vdW
GLY:219	-1,51	-0,30	1,22	Comum	vdW	C-H Bond
CYS:191	-1,21	-0,18	1,03	Comum	C-H Bond	H Bond
GLY:193	-1,14	-0,07	1,07	Comum	vdW	vdW
VAL:213	-0,87	-0,18	0,68	Comum	vdW	Alkyl
SER:214	-0,85	-0,11	0,75	Comum	Unf. D-D	vdW
CYS:220	-0,79	-0,24	0,55	Comum	vdW	vdW
SER:190	-0,49	-0,05	0,44	Comum	vdW	C-H Bond
GLU:217	-0,34	-----	-----	Padrão	vdW	vdW
ASP:194	-0,30	-0,16	0,14	Comum	vdW	vdW
GLY:226	-0,22	-----	-----	Padrão	vdW	vdW
HIS:57	-0,16	-0,05	0,11	Comum	H-B/pi-alkyl	H Bond*
ILE:227	-0,08	-----	-----	Padrão	H Bond	vdW
ALA:97	-0,07	-----	-----	Padrão	vdW	vdW
CYS:42	-0,06	-----	-----	Padrão	vdW	vdW
GLN:98	-0,06	-----	-----	Padrão	vdW	vdW
ALA:221	-0,04	-----	-----	Padrão	vdW	vdW
TYR:228	0,09	-----	-----	Padrão	vdW	vdW
ASP:189	0,34	-----	-----	Padrão	vdW	vdW
SOMA	-17,78	-3,68				

Tabela Suplementar 15 - Decomposição residual referente ao complexo formado entre a proteína 6VVU e ligante Trealose.

RESÍDUO	VALOR PADRÃO 6VVU kcal/mol	VALOR LIGANTE TREALOSE kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
GLN:192	-3,98	-2,28	1,70	Comum	H Bond*	H Bond*
TRP:215	-2,63	-1,27	1,36	Comum	vdW	H-B/C-H Bond
SER:195	-1,78	-1,77	0,01	Comum	H Bond	vdW
GLY:216	-1,62	-1,08	0,54	Comum	H Bond	H Bond
GLY:219	-1,51	-1,63	-0,12	Comum	vdW	H Bond
CYS:191	-1,21	-2,18	-0,97	Comum	C-H Bond	vdW
GLY:193	-1,14	-----	-----	Padrão	vdW	vdW
VAL:213	-0,87	-0,69	0,18	Comum	vdW	vdW
SER:214	-0,85	-0,22	0,63	Comum	Unf. D-D	vdW
CYS:220	-0,79	-1,22	-0,43	Comum	vdW	vdW
SER:190	-0,49	-3,61	-3,12	Comum	vdW	Unf. D-D
GLU:217	-0,34	-0,58	-0,24	Comum	vdW	vdW
ASP:194	-0,30	-0,39	-0,09	Comum	vdW	vdW
GLY:226	-0,22	-0,24	-0,02	Comum	vdW	vdW
HIS:57	-0,16	0,18	0,34	Comum	H-B/pi-alkyl	vdW
ILE:227	-0,08	-0,23	-0,14	Comum	H Bond	vdW
ALA:97	-0,07	-0,04	0,04	Comum	vdW	vdW
CYS:42	-0,06	-----	-----	Padrão	vdW	vdW
GLN:98	-0,06	-0,06	0,00	Comum	vdW	vdW
ALA:221	-0,04	-0,08	-0,04	Comum	vdW	vdW
TYR:228	0,09	-----	-----	Padrão	vdW	vdW
ASP:189	0,34	1,24	0,90	Comum	vdW	H Bond
SOMA	-17,78	-16,15				

Tabela Suplementar 16 - Decomposição residual referente ao complexo formado entre a proteína 7P2G e ligante Catequina.

RESÍDUO	VALOR PADRÃO 7P2G kcal/mol	VALOR LIGANTE CATEQUINA kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
MET:165	-1,53	-1,98	-0,45	Comum	Pi-sigma	Pi-alkyl
MET:49	-1,53	-1,16	0,36	Comum	Pi-alkyl	Pi-alkyl
HIS:41	-1,18	-1,04	0,14	Comum	Pi-Pi-T-shap.	Pi-Pi-t-shap.
CYS:145	-1,06	-1,26	-0,20	Comum	VdW	Pi-Alkyl
GLN:189	-0,63	-0,89	-0,27	Comum	VdW	VdW
ARG:188	-0,46	-0,26	0,20	Padrão	VdW	VdW
ASN:142	-0,42	-0,55	-0,13	Comum	VdW	H Bond
ASP:187	-0,41	-1,05	-0,64	Padrão	VdW	VdW
HIS:164	-0,10	-0,21	-0,11	Comum	VdW	VdW
SER:144	-0,07	-0,61	-0,54	Comum	VdW	VdW
HIS:163	-0,04	-0,03	0,01	Comum	Conv. H bond	Unfav. D-D
TYR:54	-0,02	0,00	0,03	Padrão	VdW	VdW
LEU:141	-0,02	-0,06	-0,04	Comum	VdW	VdW
PHE:140	0,01	-0,01	-0,02	Comum	VdW	VdW
GLU:166	0,21	-0,43	-0,64	Comum	VdW	H Bond
HIS:172	-----	-0,54	----	Comum	-----	vdW
GLY:143	-----	-10,09	----	Ligante	-----	-----
SOMA	-7,25	-10,09	-2,30			

Tabela Suplementar 17 - Decomposição residual referente ao complexo formado entre a proteína 7P2G e ligante Trealose.

RESÍDUO	VALOR PADRÃO 7P2G kcal/mol	VALOR LIGANTE TREALOSE kcal/mol	DIF. kcal/mol	TIPO	INTERAÇÃO PADRÃO	INTERAÇÃO LIGANTE
MET:165	-1,53	-1,30	0,23	Comum	Pi-sigma	vdW
MET:49	-1,53	-1,04	0,48	Comum	Pi-alkyl	vdW
HIS:41	-1,18	-1,28	-0,09	Comum	Pi-Pi-t-shaped	vdW
CYS:145	-1,06	-0,58	0,48	Comum	VdW	H Bond*
GLN:189	-0,63	-1,20	-0,57	Comum	VdW	H Bond
ARG:188	-0,46	-----	-----	Padrão	VdW	vdW
ASN:142	-0,42	-0,33	0,09	Comum	VdW	H Bond
ASP:187	-0,41	-----	-----	Padrão	VdW	vdW
HIS:164	-0,10	-0,05	0,06	Comum	VdW	H Bond*
SER:144	-0,07	-0,01	0,06	Comum	VdW	vdW
HIS:163	-0,04	0,01	0,06	Comum	Conv. H bond	H Bond
TYR:54	-0,02	-----	-----	Padrão	VdW	vdW
LEU:141	-0,02	0,00	0,02	Comum	VdW	vdW
PHE:140	0,01	0,01	0,00	Comum	VdW	vdW
GLU:166	0,21	-0,27	-0,48	Comum	VdW	H Bond*
HIS:172	-----	0,01	-----	Ligante	-----	vdW
GLY:143	-----	-0,11	-----	Ligante	-----	vdW
SOMA	**-7,25**	**-6,12**				

Tabela Suplementar 18 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 4DD8 e Catequina.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
4DD8 vs Catequina	Densidade	0,92	Não rejeita H0	Sem diferença significativa
4DD8 vs Catequina	Temperatura	0,08	Não rejeita H0	Sem diferença significativa
4DD8 vs Catequina	Pressão	0,31	Não rejeita H0	Sem diferença significativa
4DD8 vs Catequina	Potencial	0,32	Não rejeita H0	Sem diferença significativa
4DD8 vs Catequina	SASA	0,75	Não rejeita H0	Sem diferença significativa
4DD8 vs Catequina	Raio Giro	0,00	Rejeita H0	Há diferença significativa
4DD8 vs Catequina	RMSF	0,67	Não rejeita H0	Sem diferença significativa
4DD8 vs Catequina	RMSD	0,00	Rejeita H0	Há diferença significativa

Tabela Suplementar 19 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 4DD8 e Trealose.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
4DD8 vc Trealose	Densidade	0,79	Não rejeita H0	Sem diferença significativa
4DD8 vc Trealose	Temperatura	0,21	Não rejeita H0	Sem diferença significativa
4DD8 vc Trealose	Pressão	0,24	Não rejeita H0	Sem diferença significativa
4DD8 vc Trealose	Potencial	0,18	Não rejeita H0	Sem diferença significativa
4DD8 vc Trealose	SASA	0,77	Não rejeita H0	Sem diferença significativa
4DD8 vc Trealose	Raio Giro	0,00	Rejeita H0	Há diferença significativa
4DD8 vc Trealose	RMSF	0,07	Não rejeita H0	Sem diferença significativa
4DD8 vc Trealose	RMSD	0,00	Rejeita H0	Há diferença significativa

Tabela Suplementar 20 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 1NC6 e Procianidina.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
1NC6 vs Procianidina	Densidade	0,95	Falha ao rejeitar H0	Sem diferença significativa
1NC6 vs Procianidina	Temperatura	0,36	Falha ao rejeitar H0	Sem diferença significativa
1NC6 vs Procianidina	Pressão	0,32	Falha ao rejeitar H0	Sem diferença significativa
1NC6 vs Procianidina	Potencial	0,09	Falha ao rejeitar H0	Sem diferença significativa
1NC6 vs Procianidina	SASA	0,65	Falha ao rejeitar H0	Sem diferença significativa
1NC6 vs Procianidina	Raio Giro	0,00	Rejeitar H0	Há diferença significativa
1NC6 vs Procianidina	RMSF	0,00	Rejeitar H0	Há diferença significativa
1NC6 vs Procianidina	RMSD	0,00	Rejeitar H0	Há diferença significativa
1NC6 vs Procianidina	Energia Livre	0,00	Rejeitar H0	Há diferença significativa

Tabela Suplementar 21 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 6VVU e ácido ftálico.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
6VVU vs Ác. Ftálico	Densidade	0,61	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Ác. Ftálico	Temperatura	0,87	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Ác. Ftálico	Pressão	0,33	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Ác. Ftálico	Potencial	0,38	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Ác. Ftálico	SASA	0,74	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Ác. Ftálico	Raio Giro	0,00	Rejeitar H0	Há diferença significativa
6VVU vs Ác. Ftálico	RMSF	0,10	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Ác. Ftálico	RMSD	0,00	Rejeitar H0	Há diferença significativa
6VVU vs Ác. Ftálico	Energia Livre	0,00	Rejeitar H0	Há diferença significativa

Tabela Suplementar 22 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 6VVU e adenina.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
6VVU vc Adenina	Densidade	0,89	Falha ao rejeitar H0	Sem diferença significativa
6VVU vc Adenina	Temperatura	0,52	Falha ao rejeitar H0	Sem diferença significativa
6VVU vc Adenina	Pressão	0,45	Falha ao rejeitar H0	Sem diferença significativa
6VVU vc Adenina	Potencial	0,50	Falha ao rejeitar H0	Sem diferença significativa
6VVU vc Adenina	SASA	0,81	Falha ao rejeitar H0	Sem diferença significativa
6VVU vc Adenina	Raio Giro	0,00	Rejeitar H0	Há diferença significativa
6VVU vc Adenina	RMSF	0,15	Falha ao rejeitar H0	Sem diferença significativa
6VVU vc Adenina	RMSD	$1,08x10^{-14}$	Rejeitar H0	Há diferença significativa

Tabela Suplementar 23 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 6VVU e anidrido ftálico.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
6VVU vs Anidrido Ftálico	Densidade	0,45	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Anidrido Ftálico	Temperatura	0,89	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Anidrido Ftálico	Pressão	0,74	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Anidrido Ftálico	Potencial	0,52	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Anidrido Ftálico	SASA	1,00	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Anidrido Ftálico	Raio Giro	0,00	Rejeitar H0	Há diferença significativa
6VVU vs Anidrido Ftálico	RMSF	0,00	Rejeitar H0	Há diferença significativa
6VVU vs Anidrido Ftálico	RMSD	0,00	Rejeitar H0	Há diferença significativa

Tabela Suplementar 24 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 6VVU e catequina.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
6VVU vs Catechin	Densidade	0,70	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Catechin	Temperatura	0,81	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Catechin	Pressão	0,65	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Catechin	Potencial	0,69	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Catechin	SASA	0,75	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Catechin	Raio Giro	0,00	Rejeitar H0	Há diferença significativa
6VVU vs Catechin	RMSF	0,00	Rejeitar H0	Há diferença significativa
6VVU vs Catechin	RMSD	0,00	Rejeitar H0	Há diferença significativa

Tabela Suplementar 25 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 6VVU e Indol-3-Acetamida.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
6VVU vs Indol-3-Ac.	Densidade	0,57	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Indol-3-Ac.	Temperatura	0,33	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Indol-3-Ac.	Pressão	0,47	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Indol-3-Ac.	Potencial	0,73	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Indol-3-Ac.	SASA	0,61	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Indol-3-Ac.	Raio Giro	0,00	Rejeitar H0	Há diferença significativa
6VVU vs Indol-3-Ac.	RMSF	0,00	Rejeitar H0	Há diferença significativa
6VVU vs Indol-3-Ac.	RMSD	0,00	Rejeitar H0	Há diferença significativa

Tabela Suplementar 26 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 6VVU e teobromina.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
6VVU vs Teobromina	Densidade	0,81	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Teobromina	Temperatura	0,61	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Teobromina	Pressão	0,18	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Teobromina	Potencial	0,31	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Teobromina	SASA	0,79	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Teobromina	Raio Giro	0,00	Rejeitar H0	Há diferença significativa
6VVU vs Teobromina	RMSF	$8,93x10^{-7}$	Rejeitar H0	Há diferença significativa
6VVU vs Teobromina	RMSD	0,00	Rejeitar H0	Há diferença significativa

Tabela Suplementar 27 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 6VVU e trealose.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
6VVU vs Trealose	Densidade	0,70	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Trealose	Temperatura	0,73	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Trealose	Pressão	0,61	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Trealose	Potencial	0,32	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Trealose	SASA	0,88	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Trealose	Raio Giro	0,00	Rejeitar H0	Há diferença significativa
6VVU vs Trealose	RMSF	0,54	Falha ao rejeitar H0	Sem diferença significativa
6VVU vs Trealose	RMSD	0,00	Rejeitar H0	Há diferença significativa

Tabela Suplementar 28 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 7P2G e trealose.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
7P2G vs Trealose	Densidade	0,93	Falha ao rejeitar H0	Sem diferença significativa
7P2G vs Trealose	Temperatura	0,31	Falha ao rejeitar H0	Sem diferença significativa
7P2G vs Trealose	Pressão	0,75	Falha ao rejeitar H0	Sem diferença significativa
7P2G vs Trealose	Potencial	0,27	Falha ao rejeitar H0	Sem diferença significativa
7P2G vs Trealose	SASA	0,88	Falha ao rejeitar H0	Sem diferença significativa
7P2G vs Trealose	Raio Giro	0,00	Rejeitar H0	Há diferença significativa
7P2G vs Trealose	RMSF	0,01	Rejeitar H0	Há diferença significativa
7P2G vs Trealose	RMSD	0,00	Rejeitar H0	Há diferença significativa

Tabela Suplementar 29 - Métricas obtidas pela técnica de dinâmica molecular para o complexo formado entre a proteína 7P2G e catequina.

COMPONENTES	MÉTRICA	P-VALUE $\alpha = 0,05$	RESULTADOS HIPÓTESES	CONCLUSÃO
7P2G vs Catequina	Densidade	0,81	Falha ao rejeitar H0	Sem diferença significativa
7P2G vs Catequina	Temperatura	0,24	Falha ao rejeitar H0	Sem diferença significativa
7P2G vs Catequina	Pressão	0,47	Falha ao rejeitar H0	Sem diferença significativa
7P2G vs Catequina	Potencial	0,00	Rejeitar H0	Há diferença significativa
7P2G vs Catequina	SASA	0,83	Falha ao rejeitar H0	Sem diferença significativa
7P2G vs Catequina	Raio Giro	0,00	Rejeitar H0	Há diferença significativa
7P2G vs Catequina	RMSF	0,83	Falha ao rejeitar H0	Sem diferença significativa
7P2G vs Catequina	RMSD	0,00	Rejeitar H0	Há diferença significativa