UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO – UNIRIO

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS – PPGBIO
BIODIVERSIDADE NEOTROPICAL

Joana Paula da Silva Oliveira

Plataforma integrativa de bioprospecção orientada por dados: metabolômica, predição *in silico* de bioatividade e modelagem ecológica para valorização sustentável de baunilhas nativas da Mata Atlântica

Rio de Janeiro

Fevereiro de 2026

Joana Paula da Silva Oliveira

Plataforma integrativa de bioprospecção orientada por dados: metabolômica, predição *in silico* de bioatividade e modelagem ecológica para valorização sustentável de baunilhas nativas da Mata Atlântica

Tese apresentada ao Curso de Pós- Graduação em Ciências Biológicas (Biodiversidade Neotropical) do Instituto de Biociências da Universidade Federal do Estado do Rio de Janeiro, como requisito parcial para obtenção do título de Doutora em Ciências Biológicas.

Orientador(a): Profa. Dra. Andrea Furtado Macedo

Coorientadores:

Prof. Dr. Rodrigo Tardin

Prof. Dr. Murilo Lamim Bello

Prof. Dr. Carlos Araújo Lima

Rio de Janeiro

Fevereiro de 2026

Catalogação informatizada pelo(a) autor(a)

d48 da Silva Oliveira, Joana Paula
Plataforma integrativa de bioprospecção orientada por dados: metabolômica, predição in silico de bioatividade e modelagem ecológica para valorização sustentável de baunilhas nativas da Mata Atlântica / Joana Paula da Silva Oliveira. -- Rio de Janeiro : UNIRIO, 2026.

Orientadora: Andrea Furtado Macedo.
Coorientador: Rodrigo Tardin.
Coorientador: Murilo Lamim Bello.
Coorientador: Carlos Fernando Araújo-Lima.
Tese (Doutorado) - Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Ciências Biológicas, 2026.

1. Biodiversidade. 2. Bioeconomia. 3. Bioinformática. I. Macedo, Andrea Furtado, orient. II. Tardin, Rodrigo , coorient. III. Lamim Bello, Murilo, coorient. IV. Araújo-Lima, Carlos Fernando, coorient. V. Título.

JOANA PAULA DA SILVA OLIVEIRA

Plataforma integrativa de bioprospecção orientada por dados: metabolômica, predição *in silico* de bioatividade e modelagem ecológica para valorização sustentável de baunilhas nativas da Mata Atlântica

Trabalho de Conclusão de Curso de Doutorado, apresentado ao Programa de Pós-graduação em Ciências Biológicas da Universidade Federal do Estado do Rio de Janeiro, como requisito parcial para a obtenção do grau de Doutora em Ciências Biológicas. Área de concentração: Biodiversidade Neotropical.

Aprovado em: 25 / 02 / 2026

Banca examinadora:

Documento assinado digitalmente
RAQUEL DE ALMEIDA FERRANDO NEVES
Data: 05/05/2026 10:34:41-0300
Verifique em https://validar.iti.gov.br

Dra. Raquel de Almeida Ferrando Neves
Universidade Federal do Estado do Rio de Janeiro - UNIRIO

Documento assinado digitalmente
JOEL CAMPOS DE PAULA
Data: 11/05/2026 13:55:09-0300
Verifique em https://validar.iti.gov.br

Dr. Joel Campos de Paula
Universidade Federal do Estado do Rio de Janeiro - UNIRIO

Documento assinado digitalmente
GUILHERME AZEVEDO BARREIROS MARICATO
Data: 05/05/2026 19:59:33-0300
Verifique em https://validar.iti.gov.br

Dr. Guilherme Maricato
Universidade Federal do Rio de Janeiro □ UFRJ

Documento assinado digitalmente
ROBSON XAVIER FARIA
Data: 07/05/2026 11:57:12-0300
Verifique em https://validar.iti.gov.br

Dr. Robson Xavier Faria
Fundação Oswaldo Cruz - Fiocruz

Documento assinado digitalmente
RODRIGO HIPOLITO TARDIN OLIVEIRA
Data: 11/05/2026 11:50:00-0300
Verifique em https://validar.iti.gov.br

Dr. Rodrigo Tardin (Presidente da Banca)
Universidade Federal do Estado do Rio de Janeiro - UNIRIO

À minha mãe, Marisa da Silva Oliveira: quem me ensinou a sonhar e segue sonhando comigo, me recobra as forças, minha maior inspiração. *In memoriam*: Dona Maria Gomes da Silva, vó.

Agradecimentos

PhD - “Doctor of Philosophy” – Doutora em Filosofia de Ciências Biológicas (Biodiversidade Neotropical). Finalizo hoje um trabalho e o apresento com a finalidade de que seja avaliado meu mérito de receber esse título. Ao longo da execução desse trabalho, dediquei meu maior esforço para que não faltasse a essência desse título: a Filosofia. Quando defendi meu trabalho de conclusão de curso na graduação, apresentei um artigo que publicamos sobre as “respostas metabolômicas” de uma espécie de Lamiaceae a hormônios vegetais. Escrevendo aquele artigo tive, pela primeira vez, a oportunidade de perceber que um estudo dirigido por hipótese não necessariamente encontra suposições somente dentro do perímetro da esfera delineada pelas evidências. Percebi que a incorporação de outras disciplinas enriquece o estudo com propósito filosófico, para além das fronteiras de um nicho de pesquisa. E escrevi naquele artigo, a pedido de um revisor da revista em que publicaríamos, que, sob uma perspectiva ambiental, enfatizar o potencial biotecnológico de uma espécie ameaçada poderia motivar sua conservação e reintrodução. Eu senti, ao escrever aquilo, o peso da importância daquela ideia, na qual debrucei quando em contato com a obra de outros autores. Isso em meados de 2018. Sete anos depois, fizeram-me a seguinte pergunta em uma entrevista: “Como você resolveu ser cientista?”, a qual eu respondi tímida e intuitivamente: “A decisão de me tornar cientista foi já me formando bióloga, já me graduando.” Mal eu sabia que teria que refletir profundamente sobre essa minha resposta para conseguir finalizar a escrita de minha tese de doutorado. Assim, reencontrei a filosofia e o propósito humano dessa jornada que me permitiram finalizar e compartilhar esse trabalho e, quiçá, merecer tal título. Consequentemente, meus agradecimentos são para todos que, nessa jornada, me ajudaram a manter firme esse propósito e a não desistir. No âmbito acadêmico agradeço: ao meu orientador, Dr. Rodrigo Tardin, por me reanimar, pela confiança, parceria, presença e todos os ensinamentos, trocas inspiradoras e ideias para o futuro; meu co-orientador, Dr. Murilo Lamim Bello, por ter entrado nesse barco e não ter saído, apesar do mar revolto, por toda a inspiração - creio que construímos coisas boas, além de muitas ideias para o porvir; ao à época coordenador do curso de PPG Luciano Neves, por ter me escutado, apoiado e indicado um caminho possível (agradecimento que estendo a toda a Comissão de Assuntos Discentes do PPGBIO-UNIRIO e também à Samilly, do setor de Apoio Pedagógico PRAE-UNIRIO). Agradeço também ao Dr. Carlos Araújo Lima, por toda paciência, diálogo e compreensão e aos colegas de doutorado MSc Renatha Tavares e MSc Tibério Vale pela amizade e sorrisos nas horas em que muito precisei. Fora do âmbito

acadêmico e ainda no contexto de “manter firme esse propósito e a não desistir” agradeço: ao Daniel Pimentel (Emerge Biodiversidade), pelo exemplo, pela parceria, pela disposição no dia-a-dia em cada batalha e pela cumplicidade em um sonho que aos poucos se constrói; à Ornella Guzzo Vilardo, minha amiga e sócia em um sonho, pelos ensinamentos sobre como podemos impactar positivamente meio ambiente e sociedade ao realiza-lo, pela parceria, exemplo de determinação, pé no chão e olhar atento ao tempo certo das coisas. Ao Icarus Montessi pela linda parceria, exemplo de disciplina e criatividade; à Elisabeth Weigert Maier também pela linda parceria e por perfumar a vida quando ela ficou praticamente inolente; às Professoras Cláudia Furlan e Deborah Santos (IB-USP) pela confiança e parceria. Por último e mais importante, em minha família agradeço: à minha mãe, Marisa da Silva Oliveira, por esse amor e esforço ininterrupto pela manutenção de nossa estrutura, de nossas bases; minha companheira Jessica Porto, minha pedrinha preciosa, pelo amor, paciência e por seguirmos passinho-a-passinho; ao meu irmão, João Gabriel Oliveira, por esse amor e presença; ao meu tio materno, Marcos Pereira da Silva, por manter a infância nos agrados e desagrados do dia-a-dia; à minha tia materna Vilma Melo, por esse amor, cada abraço e cada cafezinho; à minha cunhada e meu sobrinho, Beatriz e Breno Scaramello, pela alegria, refúgio e sorrisos – um imenso e profundo agradecimento à minha família.

É característico da ciência que explicações completas sejam frequentemente apreendidas em sua essência pelo cientista perceptivo muito antes de qualquer prova possível.

John Desmond Bernal

Resumo

Esta tese investiga a integração entre metabolômica, análises *in silico* e modelagem ecológica como estratégia para a bioprospecção orientada por dados e a valorização sustentável da biodiversidade vegetal, com foco em espécies nativas brasileiras do gênero *Vanilla*. No Capítulo 1, é apresentada uma revisão metodológica de estudos metabolômicos de antocianinas baseados em LC–MS, sistematizando resultados comparativos sobre métodos de extração, abordagens analíticas direcionadas e não direcionadas e níveis de confiança em anotação molecular, evidenciando que a generalização por classes químicas se associa a maior robustez analítica em contextos de alta complexidade estrutural e limitada disponibilidade de padrões de referência. No Capítulo 2, uma abordagem metabolômica integrada foi aplicada a espécies nativas do gênero *Vanilla* para comparar seus perfis químicos; como principal resultado, *V. cribbiana* apresentou um conjunto distinto e recorrente de metabólitos especializados de interesse bioeconômico, evidenciando diferenciação química consistente em relação às demais espécies analisadas. No Capítulo 3, o estudo reutiliza dados de metabolômica não-alvo de folhas de *V. planifolia* para inferir bioatividade por meio de *docking* reverso integrado a modelos ADMET. O principal resultado mostra mudanças ontogenéticas nos alvos moleculares preditos, com priorização robusta de metabólitos como candidatos bioativos plausíveis. No Capítulo 4, é realizada a modelagem da distribuição de espécies, gerando mapas de adequabilidade ambiental e identificando áreas de convergência entre alta riqueza de *Vanilla* spp., potencial bioeconômico e lacunas de proteção ambiental, o que evidencia regiões estratégicas para conservação, coleta responsável e cultivo sustentável. Enfim, a Proposta de Divulgação Científica apresenta a plataforma digital ReSourceae (https://resourceae.com/), orientado à ciência aberta que reúne mapas de adequabilidade, *hotspots* de biodiversidade e indicadores de valor bioeconômico, transformando dados químicos, ecológicos e funcionais dispersos em um recurso informacional estruturado, rastreável e reutilizável. Conclui-se que a integração sistemática entre metabolômica, análises *in silico* e modelagem ecológica constitui uma abordagem robusta e generalizável para a bioprospecção orientada por dados, com implicações diretas para a conservação da biodiversidade, o fortalecimento de cadeias bioeconômicas sustentáveis e a inovação baseada em produtos naturais.

Palavras-chave: metabolômica; LC–MS; *Vanilla*; bioatividade; bioprospecção; sustentabilidade informacional.

Abstract

This thesis investigates the integration of metabolomics, *in silico* analyses, and ecological modeling as a strategy for data-driven bioprospecting and the sustainable valorization of plant biodiversity, with a focus on native Brazilian species of the genus *Vanilla*. Chapter 1 presents a methodological review of LC–MS-based metabolomics studies of anthocyanins, systematizing comparative results on extraction methods, targeted and untargeted analytical approaches, and confidence levels in molecular annotation, demonstrating that chemical-class generalization is associated with greater analytical robustness in contexts of high structural complexity and limited availability of reference standards. In Chapter 2, an integrated metabolomics approach was applied to native Vanilla species to compare their chemical profiles; as the main result, *V. cribbiana* exhibited a distinct and recurrent set of specialized metabolites of bioeconomic interest, evidencing consistent chemical differentiation relative to the other species analyzed. In Chapter 3, the study reuses untargeted metabolomics data from *V. planifolia* leaves to infer bioactivity through reverse *docking* integrated with ADMET models. The main result shows ontogenetic shifts in predicted molecular targets, with robust prioritization of metabolites as plausible bioactive candidates. Chapter 4 applies species distribution modeling, generating environmental suitability maps and identifying areas of convergence between high *Vanilla* spp. richness, bioeconomic potential, and gaps in environmental protection, thereby highlighting strategic regions for conservation, responsible harvesting, and sustainable cultivation. Finally, the Scientific Dissemination Proposal presents the digital platform ReSourceae (https://resourceae.com), an open-science-oriented web application that integrates suitability maps, biodiversity hotspots, and bioeconomic value indicators, transforming dispersed chemical, ecological, and functional data into a structured, traceable, and reusable informational resource. It is concluded that the systematic integration of metabolomics, *in silico* analyses, and ecological modeling constitutes a robust and generalizable approach to data-driven bioprospecting, with direct implications for biodiversity conservation, the strengthening of sustainable bioeconomic value chains, and innovation based on natural products.

Keywords: metabolomics; LC–MS; *Vanilla*; bioactivity; bioprospecting; informational sustainability.

Lista de Ilustrações

Introdução Geral. Figura 1. Pipeline integrativo de bioprospecção orientada pela sustentabilidade. 24

Capítulo 1. Figura 1. Fluxo de trabalho base geral com etapas usuais após a aquisição de dados metabolômicos, até a análise quimiométrica e anotações de metabólitos. 36

Capítulo 2. Figura 1. Análise diferencial baseada na mudança de dobra Log2 e nos valores p do teste t de moléculas anotadas em *V. cribbiana* e *V. planifolia*. 52

Capítulo 2. Figura 2. Análise direcionada para a estimativa de quantificação de moléculas flavorizantes características encontradas na baunilha comercial, anotadas na abordagem não direcionada deste estudo. 53

Capítulo 3. Figura 1. Representações bidimensionais dos dez ligantes de estudo (CIDs do PubChem: 5280459, 12005192, 12308704, 45359684, 13964506, 10073778, 5488152, 737375, 12241336, 70678861) desenhadas a partir de SMILES canônicos. 68

Capítulo 3. Figura 2. Mapa de calor de abundância dos ligantes selecionados através de estágios de desenvolvimento (1-5T20, 2T20, 4T20, 6T20, 8T20), mostrando a abundância de cada CID normalizada ao máximo específico do estágio. 69

Capítulo 3. Figura 3. Acoplamento molecular do CID 5280459 no bolso catalítico da 3',5'-fosfodiesterase cíclica específica de cGMP (PDE), usando a estrutura PDB 2H44. 70

Capítulo 3. Figura 4. Acoplamento molecular dos três complexos ligante-alvo com as pontuações mais altas identificados em classes de alvos biológicos distintas. 73

Capítulo 3. Figura 5. Modos de ligação previstos dos complexos ligante-alvo com as pontuações mais altas classificados pela Pontuação LigTMap. 75

Capítulo 3. Figura 6. Mapa de calor dos alvos proteicos mais fortemente previstos através de estágios de maturação com base nos valores mínimos de Pontuação de Acoplamento. 77

Capítulo 3. Figura 7. Mapa de calor das Pontuações LigTMap mais altas através de estágios de maturação, ilustrando o enriquecimento de alvos baseado em similaridade. 78

Capítulo 4. Figura 1. Curvas de resposta de preditores ambientais chave para cinco espécies de Vanilla: (A) *V. cribbiana*, (B) *V. chamissonis*, (C) *V. pompona*, (D) *V. phaeantha*, (E) *V. bahiana*. 101

Capítulo 4. Figura 2. Predição ponderada por TSS e adequabilidade de habitat ensemble para Vanilla spp. através da América Central e do Sul. 105

Capítulo 4. Figura 3. Mapas de riqueza potencial para espécies de Vanilla (V. cribbiana, V. chamissonis, V. pompona, V. bahiana e V. phaeantha), baseados em TSS ≥ 0.7 e adequabilidade do modelo ensemble de ≥ 80%. 107

Capítulo 5. Figura 2. Logomarca da plataforma online ReSourceae. 119

Lista de Tabelas

Capítulo 1. Tabela 1. Ferramentas e bancos de dados de metabólitos para a anotação de dados metabolômicos LC-MS. 37

Capítulo 2. Tabela 1. Características de desempenho analítico de padrões analíticos direcionados na análise de cromatografia líquida-espectrometria de massas (LC-MS). 50

Capítulo 2. Tabela 2. Anotações de moléculas de extratos metanólicos de *V. cribbiana* e *V. planifolia* da aquisição de dados de espectrometria de massas. 51

Capítulo 3. Tabela 1. Composição do painel de triagem virtual. 67

Capítulo 3. Tabela 2. Melhores acertos de acoplamento específicos do estágio. 70

Capítulo 3. Tabela 3. Melhores acertos de Pontuação LigTMap específicos do estágio. 71

Capítulo 3. Tabela 4. Melhores acertos de acoplamento por classe de alvo. 72

Capítulo 3. Tabela 5. Melhores acertos de Pontuação LigTMap por classe de alvo. 74

Capítulo 3. Tabela 6. Resumo das principais predições do ADMETlab 3.0 para metabólitos selecionados de *V. planifolia*. 80

Capítulo 4. Tabela 1. Métricas de desempenho para algoritmos individuais e o ensemble EMmean através de cinco espécies de *Vanilla*. 97

Capítulo 4. Tabela 2. Importância média (± DP) de permutação de preditores ambientais e climáticos através de quatro modelos de espécies de *Vanilla*. 102

Lista de Siglas e Abreviaturas

Capítulo 1. LC-MS. Cromatografia Líquida acoplada à Espectrometria de Massas. 26

Capítulo 1. HRMS. Espectrometria de Massas de Alta Resolução. 31

Capítulo 1. EAE. Extração Assistida por Enzima. 32

Capítulo 1. MAE. Extração Assistida por Microondas. 32

Capítulo 1. UAE. Extração Assistida por Ultrassom. 32

Capítulo 1. PLE. Extração com Líquido Pressurizado. 32

Capítulo 1. SFE. Extração com Fluido Supercrítico. 32

Capítulo 1. ME. Extração por Maceração. 33

Capítulo 1. SCDE. Extração com Dióxido de Carbono Supercrítico. 34

Capítulo 1. RP. Fase Reversa. 35

Capítulo 1. HILIC. Cromatografia Líquida de Interação Hidrofílica. 35

Capítulo 1. LTQ-Orbitrap. Armadilha Linear de Quadrupolo Orbitrap. 36

Capítulo 1. MRM. Monitoramento de Reação Múltipla. 36

Capítulo 1. MS/MS. Espectrometria de Massas em Tandem. 36

Capítulo 1. MSn. Espectrometria de Massas Multietapa. 36

Capítulo 1. KEGG. Enciclopédia de Genes e Genomas de Quioto. 37

Capítulo 1. PCA. Análise de Componentes Principais. 38

Capítulo 1. PLS-DA. Análise Discriminante de Mínimos Quadrados Parciais. 38

Capítulo 1. OPLS-DA. Análise Discriminante de Mínimos Quadrados Parciais Ortogonal. 38

Capítulo 1. QC. Controle de Qualidade. 49

Capítulo 2. ESI. Fonte de Ionização por Eletrospray. 49

Capítulo 2. ESI+. Modo de Ionização Positivo. 49

Capítulo 2. ESI-. Modo de Ionização Negativo. 49

Capítulo 2. LOD. Limite de Detecção. 49

Capítulo 2. LOQ. Limite de Quantificação. 49

Capítulo 2. SPME. Microextração em Fase Sólida. 49

Capítulo 2. GC-MS. Cromatografia Gasosa acoplada à Espectrometria de Massas. 49

Capítulo 3. ADMET. Absorção, Distribuição, Metabolismo, Excreção e Toxicidade. 57

Capítulo 3. PDB IDs. Identificadores do Banco de Dados de Proteínas. 64

Capítulo 3. MACCS. Chaves de Estrutura Molecular Codificadas em Computador. 64

Capítulo 3. EState. Estado Eletrotopológico. 65

Capítulo 3. PubChem CIDs. Identificadores de Compostos do PubChem. 68

Capítulo 3. QT. Prolongamento do Intervalo QT. 80

Capítulo 3. DILI. Lesão Hepática Induzida por Fármacos. 80

Capítulo 3. P-gp Sub. Substrato de P-glicoproteína. 80

Capítulo 3. HIA. Absorção Intestinal Humana. 80

Capítulo 3. PPB. Ligação a Proteínas Plasmáticas. 80

Capítulo 3. Ames. Teste de Mutagenicidade Ames. 80

Capítulo 3. Carc. Carcinogenicidade. 80

Capítulo 4. CWRs. Parentes Silvestres de Culturas. 75

Capítulo 4. SDM. Modelagem de Distribuição de Espécies. 94

Capítulo 4. NDVI. Índice de Vegetação por Diferença Normalizada. 95

Capítulo 4. VIF. Fator de Inflação da Variância. 96

Capítulo 4. PAs. Pseudo-ausências. 96

Capítulo 4. SDMs. Modelos de Distribuição de Espécies. 96

Capítulo 4. TSS. Estatística de Habilidade Verdadeira. 96

Capítulo 4. CNUC. Cadastro Nacional de Unidades de Conservação. 96

Capítulo 4. HLI. Índice de Carga de Calor. 102

Capítulo 4. WGS84. Sistema de Referência Geodésico Mundial 1984. 106

Capítulo 4. PA. Área Protegida. 106

Sumário

1. Introdução Geral 18
1.1. Metabolômica e bioprospecção 19
1.2. Sustentabilidade de dados e bioatividade 21
1.3. Modelagem ecológica aplicada 22
1.4. Pipeline integrativo de bioprospecção 24
1.5. Biodiversidade brasileira de baunilhas 24
1.6. Estrutura e objetivos da tese 25
Capítulo 1 - Estudos metabolômicos de antocianinas em frutas por meio de cromatografia líquida acoplada ao fluxo de trabalho de espectrometria de massas 28
Capítulo 2 - Perfil metabolômico integrado direcionado e não direcionado de espécies de baunilha da Mata Atlântica: Revelando o potencial bioeconômico da *Vanilla cribbiana* ... 44
Capítulo 3 - Triagem virtual baseada em reutilização de dados metabolômicos indica potenciais alvos moleculares dos metabólitos foliares de *Vanilla planifolia* 56
Capítulo 4 - Modelagem da distribuição de espécies revela *hotspots* de biodiversidade para baunilha selvagem (*Vanilla* spp.) de valor bioeconômico no Brasil 90
2. Proposta de Divulgação Científica - ReSourceae (https://resourceae.com/) 118
3. Conclusões Gerais 123
4. Referências Bibliográficas 124
5. Apêndice 131
5.2. Script do LigTMap (ligtmap.py) refatorado 131

1. Introdução Geral

A biodiversidade é amplamente reconhecida como um pilar central da inovação científica, tecnológica e econômica, por concentrar elevada diversidade de compostos químicos, funções biológicas e interações ecológicas com alto potencial de aplicação em saúde, agricultura, alimentação e indústria química (Newman e Cragg, 2020). Nesse contexto, a bioeconomia se consolida como uma estratégia global voltada ao uso sustentável de recursos biológicos renováveis, buscando integrar desenvolvimento econômico, conservação ambiental e bem-estar social por meio do conhecimento científico e tecnológico sobre a biodiversidade (Bugge et al., 2016; OECD, 2018). Contudo, a exploração histórica desses recursos ocorreu, em grande parte, dissociada de critérios ecológicos e sociais, resultando em uma tensão persistente entre conservação da biodiversidade, geração de valor econômico e inovação em cadeias produtivas baseadas em recursos naturais (Steffen et al., 2015; Díaz et al., 2019). Essa tensão é especialmente pronunciada em países megadiversos, onde a elevada riqueza biológica coexiste com degradação ambiental e subaproveitamento científico e tecnológico da biodiversidade (Mittermeier et al., 2011; Joly et al., 2018). Nesse cenário, torna-se evidente a necessidade de abordagens científicas integradas e multidisciplinares capazes de aprofundar o conhecimento sobre a biodiversidade e, simultaneamente, subsidiar estratégias inovadoras, sustentáveis e socialmente responsáveis de valorização dos recursos biológicos (Sutherland et al., 2018; Pascual et al., 2021).

Os sistemas biológicos são intrinsecamente complexos e organizados em múltiplos níveis de interação, do molecular ao ecológico, o que torna insuficiente sua compreensão por meio de abordagens analíticas isoladas (Kitano, 2002). Essa visão encontra respaldo na teoria da complexidade, que propõe a compreensão dos fenômenos vivos a partir da articulação entre níveis, relações e contextos, rejeitando a fragmentação disciplinar como estratégia exclusiva de produção de conhecimento (Morin, 2008). Nesse cenário, dados fragmentados ou restritos a um único nível de organização biológica apresentam baixo poder explicativo para inferir funções biológicas, interações ecológicas ou potenciais aplicações biotecnológicas de organismos e metabólitos naturais (Selvarajoo, 2021). Como resposta a essa limitação, observa-se uma tendência crescente à integração de dados experimentais de alto rendimento com ferramentas computacionais e modelos *in silico*, capazes de ampliar a interpretação biológica e funcional de grandes volumes de dados complexos (Selvarajoo, 2021). Essa convergência metodológica tem impulsionado o desenvolvimento de pipelines analíticos

integrados, especialmente na bioprospecção, nos quais dados químicos, biológicos e ambientais são combinados de forma sistemática para acelerar a descoberta de compostos bioativos, reduzir custos experimentais e aumentar a eficiência dos processos de triagem e validação (Harvey et al., 2015; Atanasov et al., 2021).

1.1. Metabolômica e bioprospecção

A metabolômica se consolidou como uma das principais ferramentas de alto rendimento para a caracterização química de sistemas biológicos, ao permitir a detecção simultânea de uma ampla variedade de metabólitos em diferentes matrizes biológicas por meio de abordagens analíticas baseadas, principalmente, em espectrometria de massas acoplada a técnicas cromatográficas (Nicholson et al., 1999; Dettmer et al., 2007). Diferentemente de estratégias analíticas direcionadas, a metabolômica não direcionada possibilita a geração de hipóteses a partir da observação global dos perfis metabólicos, favorecendo a descoberta de compostos inesperados, padrões químicos emergentes e assinaturas metabólicas associadas a condições biológicas, ambientais ou evolutivas específicas (Fiehn, 2002; Patti et al., 2012). Essa característica torna a metabolômica particularmente relevante para o estudo da biodiversidade vegetal, uma vez que plantas produzem uma diversidade extraordinária de metabólitos especializados, muitos dos quais estão diretamente relacionados a adaptações ecológicas, interações bióticas e potenciais aplicações biotecnológicas ainda pouco exploradas (Wink, 2003; Wolfender et al., 2019). No entanto, apesar de seu elevado poder descritivo, a metabolômica enfrenta desafios significativos relacionados à anotação molecular confiável, à reprodutibilidade entre plataformas analíticas e à interpretação biológica de grandes volumes de dados complexos, o que limita a conversão direta desses dados em conhecimento funcional ou aplicações práticas (Schrimpe-Rutledge et al., 2016; Blaženović et al., 2018). A anotação molecular confiável em abordagens não direcionadas permanece limitada pela cobertura incompleta das bibliotecas espectrais e pela ausência de validação por padrões de referência, exigindo que os resultados sejam interpretados como triagem estrutural sujeita a confirmação posterior (Schymanski et al., 2014). A reprodutibilidade entre plataformas é comprometida pela variabilidade nos métodos de extração, ionização e fragmentação, dificultando comparações entre estudos (Misra, 2021). Já a interpretação biológica de grandes conjuntos de metabólitos anotados esbarra na escassez de dados funcionais associados à maioria das moléculas detectadas, especialmente em espécies vegetais pouco estudadas (Wolfender et al., 2019). A presente tese aborda esses

desafios em diferentes frentes: o Capítulo 1 sistematiza estratégias analíticas e níveis de confiança na anotação de metabolômicos de antocianinas, contribuindo para a padronização metodológica; o Capítulo 3 mostra como dados não direcionados já publicados podem ser reinterpretados funcionalmente por meio de workflows *in silico* de *docking* reverso e modelagem ADMET, convertendo inventários químicos descritivos em hipóteses de bioatividade; e a plataforma ReSourceae, apresentada como proposta de divulgação científica, integra e torna acessíveis os dados químicos, ecológicos e funcionais gerados ao longo da tese, reduzindo barreiras à reutilização e interpretação do conhecimento produzido (Wilkinson et al., 2016). Assim, a metabolômica deve ser compreendida não como um fim em si mesma, mas como uma fonte estruturante de dados químicos que, quando integrados a outras abordagens experimentais e computacionais, potencializam a interpretação funcional da biodiversidade e ampliam o impacto científico e tecnológico da bioprospecção (Wolfender et al., 2009; Misra et al., 2019).

A anotação molecular confiável constitui um dos principais gargalos da metabolômica contemporânea, uma vez que a interpretação biológica e funcional dos perfis metabólicos depende diretamente da qualidade e do nível de confiança atribuídos às identificações químicas (Schrimpe-Rutledge et al., 2016; Blaženović et al., 2018). Nesse contexto, estratégias que operam em níveis hierárquicos mais amplos, como a generalização ao nível de classes químicas, têm sido amplamente reconhecidas como abordagens interpretativas robustas, especialmente em estudos de biodiversidade, nos quais a diversidade estrutural e a escassez de padrões de referência dificultam a anotação inequívoca de metabólitos individuais (Wink, 2003; Wolfender et al., 2019). A metabolômica apresenta elevado potencial para direcionar análises focadas em classes químicas específicas, permitindo não apenas a caracterização abrangente dessas classes, mas também a definição de parâmetros analíticos e experimentais mais adequados para sua detecção, anotação e interpretação funcional, como demonstrado em estudos centrados em polifenóis e outras classes de metabólitos especializados de plantas (Fiehn, 2002; Sumner et al., 2007). Nesse sentido, as antocianinas constituem um estudo de caso paradigmático, amplamente explorado na literatura como modelo para o desenvolvimento e a avaliação crítica de fluxos analíticos baseados em metabolômica, em razão de sua relevância biológica, diversidade estrutural e importância ecológica e funcional, conforme sistematizado em revisões metodológicas dedicadas a essa classe química (Castaneda-Ovando et al., 2009; Oliveira et al., 2022). Embora as antocianinas não representem uma classe molecular alvo central nos estudos

metabolômicos aplicados às espécies estudadas nesta tese, sua adoção como modelo conceitual permite ilustrar como classes químicas de elevado potencial bioativo podem ser acompanhadas ao longo de um pipeline integrado de avaliação, que abrange desde a bioprospecção química e a geração de dados metabolômicos até análises *in silico* de bioatividade e a contextualização ecológica das espécies produtoras (Harvey et al., 2015; Wolfender et al., 2019). Adicionalmente, revisões metodológicas desempenham papel fundamental na consolidação desse tipo de abordagem integrativa, ao orientar decisões relacionadas ao desenho experimental, à escolha de parâmetros analíticos e à definição de estratégias de *downstream analyses*, contribuindo para o uso mais eficiente, reprodutível e interpretável dos dados gerados em estudos metabolômicos de alta complexidade (Sumner et al., 2007; Schrimpe-Rutledge et al., 2016).

1.2. Sustentabilidade de dados e bioatividade

A expansão acelerada das tecnologias ômicas ao longo das últimas décadas resultou em uma produção sem precedentes de dados biológicos de alta complexidade e alto volume, impondo desafios substanciais relacionados ao armazenamento, à interoperabilidade, à interpretação e, sobretudo, ao reuso eficiente dessas informações (Stephens et al., 2015; Leonelli et al., 2015). Nesse contexto, a sustentabilidade informacional emerge como um conceito científico central, sustentado por iniciativas de ciência aberta e pelos princípios FAIR (*Findable, Accessible, Interoperable and Reusable*), que visam garantir que os dados gerados sejam não apenas preservados, mas também reutilizáveis, reprodutíveis e passíveis de integração em novos contextos analíticos (Wilkinson et al., 2016; Mons et al., 2017). A metabolômica ocupa posição estratégica nesse cenário, uma vez que seus dados, quando adequadamente anotados, padronizados e contextualizados, apresentam elevado potencial de reaproveitamento em estudos comparativos, meta-análises e aplicações *downstream*, incluindo avaliações funcionais, ecológicas e preditivas, ampliando significativamente o valor científico dos experimentos originais (Viant et al., 2017; Johnson et al., 2020). Entretanto, a ausência de estratégias integrativas e de plataformas analíticas capazes de conectar dados químicos, biológicos e ambientais de forma sistemática contribui para um desperdício informacional recorrente, no qual grandes volumes de dados permanecem subexplorados ou restritos a análises pontuais (Leonelli, 2019; Sansone et al., 2019). Assim, o desenvolvimento de plataformas conceituais e analíticas integrativas representa uma resposta necessária a esse desafio, ao permitir a articulação eficiente entre diferentes

camadas de dados e promover uma abordagem sustentável da geração, interpretação e reutilização de informações científicas no contexto da bioprospecção e da bioeconomia baseada na biodiversidade (Mons et al., 2017; Sansone et al., 2019).

A avaliação de bioatividade baseada exclusivamente em bioensaios tradicionais apresenta limitações significativas, incluindo alto custo, baixa escalabilidade, elevado consumo de recursos biológicos e químicos, além de restrições éticas e ambientais associadas ao uso extensivo de organismos modelo e ensaios experimentais repetitivos (Hughes et al., 2011; Atanasov et al., 2021). Nesse contexto, a integração de dados metabolômicos em estratégias de triagem racional tem se consolidado como uma abordagem eficiente para priorizar compostos e classes químicas com maior probabilidade de apresentar relevância biológica, reduzindo o espaço químico a ser explorado experimentalmente e aumentando a taxa de sucesso em etapas subsequentes de validação funcional (Wolfender et al., 2019; Johnson et al., 2020). Ferramentas *in silico*, como o *docking* molecular e os modelos farmacocinéticos baseados em fisiologia (PBPK), emergem como componentes complementares nesse processo, permitindo a predição de interações molécula-alvo, afinidade de ligação, propriedades ADMET e comportamento sistêmico de compostos bioativos a partir de estruturas químicas derivadas de dados metabolômicos. A aplicação integrada dessas abordagens computacionais possibilita a avaliação funcional preliminar de grandes conjuntos de metabólitos de forma rápida, reprodutível e escalável, contribuindo para a racionalização de *pipelines* de bioprospecção e para a diminuição da dependência de ensaios experimentais iniciais de alto custo (Ferreira et al., 2015; Shah et al., 2025). Dessa forma, a integração entre metabolômica e análises *in silico* de bioatividade representa uma estratégia alinhada aos princípios de sustentabilidade científica, ao reduzir tempo, custos e impacto ambiental associados à descoberta de compostos bioativos, sem comprometer a robustez analítica e a profundidade interpretativa dos estudos (Ferreira et al., 2015; Atanasov et al., 2021).

1.3. Modelagem ecológica aplicada

Embora a integração entre metabolômica e ferramentas *in silico* de bioatividade permita avançar substancialmente na interpretação funcional de compostos naturais e classes químicas de interesse, a valorização efetiva da biodiversidade exige que esses dados sejam contextualizados no espaço ecológico em que os organismos produtores estão inseridos, ampliando a análise do nível molecular para o territorial e socioambiental (Bezerra et al.,

2025). A distribuição geográfica das espécies é fortemente influenciada por fatores climáticos, edáficos e bióticos, sendo atualmente remodelada por mudanças climáticas aceleradas e por pressões antrópicas crescentes, o que impõe riscos simultâneos à conservação da biodiversidade e à sustentabilidade de cadeias produtivas baseadas em recursos naturais (Bellard et al., 2012; IPBES, 2019). Nesse contexto, a modelagem ecológica, em especial os modelos de distribuição de espécies, consolida-se como uma ferramenta preditiva essencial para estimar áreas potenciais de ocorrência, identificar regiões ambientalmente adequadas sob cenários atuais e futuros e subsidiar estratégias de manejo e uso sustentável (Elith & Leathwick, 2009; Zhao et al., 2025). Essas abordagens têm sido amplamente aplicadas tanto no planejamento de ações de conservação, por meio da identificação de *hotspots* de biodiversidade e áreas prioritárias para proteção legal, quanto na prospecção de áreas potenciais para cultivo sustentável de espécies de interesse econômico, reduzindo a pressão sobre populações naturais e alinhando produção e conservação (Myers et al., 2000; Fadda et al., 2026). Dessa forma, a incorporação da modelagem ecológica ao pipeline integrativo proposto nesta tese amplia o alcance das análises químicas e funcionais, conectando dados metabolômicos e predições de bioatividade a decisões territoriais e socioambientais, e deixando explícito que a bioprospecção orientada pela sustentabilidade não se encerra no laboratório, mas se projeta necessariamente sobre o espaço, o clima e a paisagem (Figura 1).

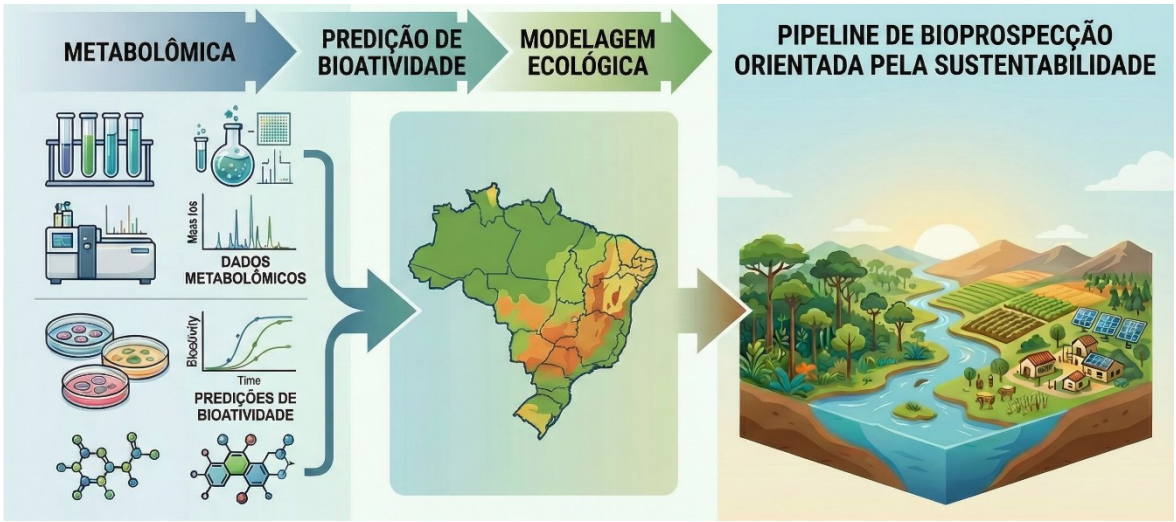


Figura 1. Pipeline integrativo de bioprospecção orientada pela sustentabilidade. O fluxo metodológico conecta dados experimentais de metabolômica e predições de bioatividade (esquerda) à modelagem de distribuição de espécies (centro). Essa integração permite projetar as análises químicas sobre o espaço

geográfico, fundamentando decisões territoriais e socioambientais (direita) que consideram o clima e a paisagem.

1.4. Pipeline integrativo de bioprospecção

A convergência entre metabolômica, análises *in silico* de bioatividade e modelagem ecológica estabelece um fluxo lógico e coerente que conecta a caracterização química da biodiversidade à sua interpretação funcional e, por fim, à sua contextualização socioambiental, configurando uma abordagem integrada de bioprospecção orientada por dados (Wolfender et al., 2019; Guisan et al., 2013). Nesse fluxo, a metabolômica atua como eixo estruturante ao gerar dados químicos de alto rendimento, que podem ser racionalmente priorizados por ferramentas computacionais de predição de bioatividade e posteriormente reinseridos no espaço geográfico e ecológico por meio de modelos de distribuição de espécies e análises territoriais (Johnson et al., 2020; Elith e Leathwick, 2009). A integração sistemática de dados químicos, biológicos e espaciais amplia significativamente o poder interpretativo dos estudos de biodiversidade, permitindo não apenas a identificação de compostos e funções potenciais, mas também a avaliação da viabilidade ecológica e territorial de estratégias de uso sustentável das espécies produtoras. Essa abordagem integrada fornece subsídios científicos robustos para o fortalecimento de uma bioeconomia baseada na biodiversidade, ao alinhar inovação biotecnológica, conservação ambiental e planejamento territorial em um mesmo arcabouço analítico (Bugge et al., 2016; OECD, 2018). Como resultado, a plataforma conceitual delineada nesta tese promove simultaneamente a sustentabilidade ambiental, ao reduzir pressões sobre ecossistemas naturais por meio de decisões informadas, e a sustentabilidade informacional, maximizar o valor, a reutilização e a integração de dados científicos ao longo de todo o pipeline de bioprospecção (Wilkinson et al., 2016; Mons et al., 2017).

1.5. Biodiversidade brasileira de baunilhas

A baunilha natural é um dos insumos de maior valor agregado no mercado de aromas e sabores, com cadeia produtiva global marcada por elevada demanda, vulnerabilidade a choques climáticos e instabilidade de oferta, em grande parte por depender de poucos polos produtores e de um número muito restrito de espécies cultivadas em escala comercial (Baker, 2018; de Oliveira et al., 2022; Hammami e Elliott, 2025). Essa combinação de concentração produtiva, volatilidade e pressões ambientais tem sido descrita como um cenário de “crise da baunilha”, no qual a diversificação de fontes botânicas e a ampliação do pool genético

disponível são estratégias relevantes para reduzir riscos, aumentar resiliência e sustentar inovação baseada em recursos naturais (de Oliveira et al., 2022; Oliveira et al., 2025). Nesse contexto, a biodiversidade brasileira do gênero *Vanilla* representa uma oportunidade estratégica, uma vez que o país abriga dezenas de espécies nativas (incluindo endêmicas) distribuídas em diferentes biomas, com lacunas de conhecimento ainda importantes quanto à distribuição, conservação e potencial de uso sustentável (Nadal et al., 2025; Pansarin et al., 2023). Evidências recentes apontam que espécies nativas brasileiras e seus extratos apresentam perfis químicos complexos e potencial de aplicação em cadeias de valor, incluindo resultados baseados em metabolômica de alta resolução e avaliação de segurança/toxicidade por abordagens in vitro e *in silico* (de Oliveira et al., 2024; Oliveira et al., 2025). Além disso, a caracterização química e o potencial sensorial de espécies brasileiras têm sido reportados como base para perspectivas de diferenciação e agregação de valor, reforçando a relevância de integrar bioprospecção química, avaliação funcional e planejamento socioambiental para desenvolvimento do setor (Silva et al., 2023; Oliveira et al., 2025).

1.6. Estrutura e objetivos da tese

Esta tese está estruturada no formato de capítulos correspondentes a artigos científicos, organizados de modo a refletir um eixo integrador baseado na bioprospecção orientada por dados e na sustentabilidade informacional e socioambiental, conforme recomendado para trabalhos que articulam abordagens multidisciplinares em biodiversidade e bioeconomia. O Capítulo 1 consiste em um artigo de revisão metodológica sobre o uso da cromatografia líquida acoplada à espectrometria de massas (LC–MS) em estudos metabolômicos de produtos naturais, elaborado a convite da revista *Plants* em reconhecimento à expertise prévia da autora e colaboradores na técnica analítica. Com foco em antocianinas de frutos como sistema-modelo, o artigo sistematiza estratégias de extração, métodos de ionização, abordagens direcionadas e não direcionadas, e protocolos de anotação com diferentes níveis de confiança. Sua inserção na tese é justificada pela transferibilidade direta do conhecimento técnico consolidado para outras matrizes vegetais complexas, como as espécies de *Vanilla* estudadas nos capítulos seguintes, os desafios analíticos associados à diversidade estrutural, isomeria e limitada disponibilidade de padrões de referência são compartilhados entre antocianinas e polifenóis da baunilha, tornando as conclusões metodológicas do capítulo diretamente aplicáveis às decisões experimentais e interpretativas adotadas nos Capítulos 2

e 3. O Capítulo 2 apresenta um estudo de metabolômica integrada aplicado a espécies nativas de *Vanilla* da biodiversidade brasileira, gerando dados químicos primários de alto rendimento e explorando seu potencial bioeconômico por meio de abordagens direcionadas e não direcionadas. O Capítulo 3, por sua vez, explora exclusivamente dados químicos disponíveis em bases públicas e literatura científica, adotando uma estratégia explícita de reuso de dados metabolômicos para análises *in silico* de bioatividade, incluindo *docking* molecular e propriedades farmacocinéticas e toxicológicas, demonstrando como informações previamente geradas podem ser reinterpretadas funcionalmente sem a necessidade de novos experimentos laboratoriais, em alinhamento com princípios de ciência aberta e sustentabilidade informacional. O Capítulo 4 amplia o escopo da tese ao incorporar a modelagem ecológica como ferramenta para mapear áreas potenciais de ocorrência, cultivo sustentável e conservação das espécies estudadas, conectando dados químicos e funcionais a dimensões espaciais, climáticas e territoriais. Em conjunto, os capítulos não devem ser interpretados como estudos independentes, mas como módulos complementares de uma proposta de sistema integrado de geração, reutilização, interpretação e aplicação de dados, no qual a metabolômica, as análises *in silico* e a modelagem ecológica se articulam para promover uma abordagem sustentável e orientada por dados da valorização da biodiversidade.

A hipótese central desta tese é que a integração sistemática de dados metabolômicos, análises *in silico* de bioatividade e modelagem ecológica constitui uma estratégia cientificamente robusta e sustentável para a bioprospecção e valorização da biodiversidade, ao permitir a transição do inventário químico para a interpretação funcional e territorial das espécies estudadas de forma orientada por dados e baseada em reuso informacional. Nesse sentido, o objetivo geral do trabalho é desenvolver e demonstrar um arcabouço integrativo capaz de conectar caracterização química de alto rendimento, predições computacionais de bioatividade e análises espaciais preditivas, produzindo conhecimento acionável para a bioeconomia baseada na biodiversidade e para estratégias de conservação e uso sustentável. Como desdobramento desse objetivo científico, a tese adota explicitamente uma perspectiva de divulgação científica e extensão, reconhecendo que a efetiva valorização da biodiversidade depende também da acessibilidade, transparência e comunicação dos dados gerados para públicos não especializados e tomadores de decisão (Leonelli, 2019; Sansone et al., 2019). Nesse contexto, é apresentada uma plataforma digital pública, na forma de um aplicativo *web*, destinada à disponibilização das informações geradas sobre potencial de

distribuição, adequabilidade ambiental e áreas prioritárias associadas às espécies estudadas, a partir dos mapas produzidos por modelagem ecológica, funcionando como uma ferramenta de apoio ao planejamento territorial, à conservação e a iniciativas de cultivo sustentável. Essa estratégia de extensão reforça o compromisso da tese com os princípios de ciência aberta e sustentabilidade informacional, ao promover o compartilhamento responsável de dados e resultados científicos e ampliar o impacto social do conhecimento produzido no âmbito acadêmico (Mons et al., 2017; Wilkinson et al., 2016).

Capítulo 1 - Estudos metabolômicos de antocianinas em frutas por meio de cromatografia líquida acoplada ao fluxo de trabalho de espectrometria de massas

O Capítulo 1 desta tese corresponde a um artigo de revisão elaborado a convite do periódico no qual foi publicado, *Plants* (ISSN: 2223-7747), refletindo a maturidade e a relevância científica do tema abordado no contexto da metabolômica aplicada à bioprospecção química (Oliveira et al., 2022). O artigo tem como objetivo central discutir criticamente o uso de abordagens metabolômicas baseadas em cromatografia líquida acoplada à espectrometria de massas para o estudo de antocianinas, uma classe de metabólitos especializados amplamente distribuída em plantas e reconhecida por sua diversidade estrutural, relevância biológica e interesse tecnológico. A revisão está estruturada em seções que abordam a diversidade química e biossíntese das antocianinas, os desafios analíticos associados à sua detecção e anotação, estratégias experimentais direcionadas e não direcionadas em metabolômica, bem como limitações relacionadas à reprodutibilidade, padronização e interpretação dos dados. Como principal conclusão, o artigo evidencia que a metabolômica, quando aplicada de forma criteriosa e apoiada por estratégias robustas de anotação e controle de qualidade, constitui uma ferramenta poderosa não apenas para o estudo aprofundado de antocianinas, mas também para a tomada de decisões experimentais em estudos de biodiversidade molecular vegetal. Embora as antocianinas não representem o foco químico central dos estudos empíricos desenvolvidos nos capítulos subsequentes desta tese, as discussões metodológicas e conceituais apresentadas neste capítulo são diretamente generalizáveis para outras classes químicas de metabólitos especializados, oferecendo um modelo analítico e interpretativo aplicável à metabolômica de plantas em sentido amplo. Dessa forma, o Capítulo 1 estabelece as bases conceituais e metodológicas que sustentam o uso da metabolômica como eixo estruturante do projeto de tese, orientando tanto a geração quanto o reaproveitamento de dados químicos em pipelines integrativos que conectam bioprospecção, bioatividade e sustentabilidade socioambiental.

Current Plant Biology 32 (2022) 100260

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

Metabolomic studies of anthocyanins in fruits by means of a liquid chromatography coupled to mass spectrometry workflow

Joana Paula da Silva Oliveira [1], Renatha Tavares de Oliveira [1], Amanda Lima Guedes [1], Matheus da Costa Oliveira, Andrea Furtado Macedo *

Integrated Laboratory of Plant Biology (LIBV), Institute of Biosciences, Federal University of the State of Rio de Janeiro – UNIRIO, Av. Pasteur, 458, Urca, Rio de Janeiro, RJ CEP 22290-240, Brazil

ARTICLE INFO

Keywords:
Extraction methods
Analytical methods
Chemometrics
LC-MS
Metabolites annotation

ABSTRACT

Among plant pigments, anthocyanins are one of the most important. However, there are great difficulties related to applying appropriate techniques able to identify molecules with similar structures (suchlike anthocyanins) and simultaneously estimate their variety and concentration within fruit matrices. Therefore, several methods have been evaluated to achieve the best workflows for anthocyanin analysis. For the first time, this review article presents the current findings on analytical procedures for anthocyanin investigation based on the robust liquid chromatography coupled to mass spectrometry (LC-MS) workflow. Anthocyanins extraction, LC-MS analytical instrumentation, and chemometric strategies that converge to the better performance of such workflow are herein presented. By means of this metabolomics platform, it is possible to obtain the most accurate identification of phenolic compounds, such as anthocyanins. This is a crucial step since these compounds are related to the sensory qualities of food, such as color, flavor, and aroma.

1. Introduction

Anthocyanins are natural water-soluble pigments, commonly found in the sap of plant cells of certain vegetables, fruits, and grains. These molecules belong to the group of phenolic compounds, which are considered secondary metabolites (Supplementary Figure 1, [27]). The basic molecular structure of anthocyanins is composed of two aromatic rings and one heterocyclic ring. The two benzene rings are connected by three carbon atoms to form a C6–C3–C6 skeleton, which is the anthocyanin motif (Supplementary Figure 1). Anthocyanins also occur in the form of aglycones, known as anthocyanidins. The most common anthocyanidins are cyanidin, delphinidin, pelargonidin, peonidin, malvidin, and petunidin (Supplementary Figure 2). However, anthocyanin aglycones are not common and are mostly found associated with glycosides in fruits. The main glycosides are composed of glucose, galactose, xylose, arabinose, rhamnose moieties or disaccharides, to which phenolic acids such as *p*-coumaric, caffeic, phenyl, and vanillic acids can be linked, to form anthocyanins through glycosidic bonds [112]. Anthocyanins contribute to the red, purple, and blue pigmentation in tissues of higher plants (Supplementary Figure 2) [72].

The color of these pigments and their chemical stability, bioavailability, biological activity, and diverse molecular structure depend upon their functional groups, including different numbers of substitutions, sites of hydroxylation on different rings, patterns of methylation, and

Abbreviations: APCI, Atmospheric pressure chemical ionization; ASE, Accelerated solvent extraction; BS, Blank samples; DES, Deep eutectic solvents; EAE, Enzyme-assisted extraction; ESI, Electrospray ionization; HILIC, Hydrophilic interaction liquid chromatography; LC, Liquid chromatography; MAE, Microwave-assisted extraction; ME, Maceration extraction; MRM, Multiple reaction monitoring; MS, Mass spectrometry; MS/MS, Tandem mass spectrometry; NADES, Natural Deep eutectic solvents; PEF, Pulsed electric fields; PLE, Pressurized liquid extraction; QC, Quality control; QqQ, Triple quadrupole; Q-TOF, Quadrupole time-of-flight; Q-TRAP, Triple-quadrupole ion trap; RP, Reversed phase; scCO2, Supercritical carbon dioxide; SCDE, Supercritical carbon dioxide extraction; SEM, Solvent extraction method; SFE, Supercritical fluid extraction; SWE, Subcritical water extraction; TFA, Trifluoroacetic acid; TOF, Time-of-flight; UAE, Ultrasound-assisted extraction; UV-Vis, Ultraviolet visible spectroscopy.

* Correspondence to: Integrated Laboratory of Plant Biology, Institute of Biosciences, Federal University of the State of Rio de Janeiro – UNIRIO, Av. Pasteur, 458, Urca, Rio de Janeiro 22290-240, Brazil.

E-mail addresses: joanapaulasoliveira@edu.unirio.br (J.P. da Silva Oliveira), renathat.oliveira@edu.unirio.br (R.T. de Oliveira), amandalguedes@edu.unirio.br (A.L. Guedes), matheusoliveira@edu.unirio.br (M. da Costa Oliveira), andrea.macedo@unirio.br (A.F. Macedo).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cpb.2022.100260
Received 23 April 2022; Received in revised form 30 September 2022; Accepted 2 October 2022
Available online 7 October 2022
2214-6628/© 2022 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

the number of glycosylation sites (Supplementary Figure 2) [140]. Intracellular environmental conditions where anthocyanins are found, such as pH and association with non-pigmented phenolics (like flavonols and flavones, considered co-pigments), can also change the final color. The resulting color of peripheral plant tissues, such as the epidermis, will depend on the diversity of anthocyanins present and the conditions mentioned above [41].

Anthocyanins have multiple functions, such as attracting insects to complete the pollination process and protecting plants from UV damage [112]. More than 635 types of anthocyanins have been identified and many have biological activities linked to human health [15,46,52,64,92, 132,141,144]. Anthocyanins are important molecules in the food sector and pharmaceutical industry, as natural pigments can replace artificial ones that may cause environmental damage and human health diseases [140].

The analysis of anthocyanins can become a great challenge. As representatives of polyphenols, anthocyanins have great diversity and present structural complexity, with highly polymerized structures, which can also form complexes with several other plant components [85,130]. In recent years, innovations in chromatographic performance in liquid chromatography (LC) column and mass spectrometry (MS) instrumentation have been fundamental to meeting the growing analytical demands for anthocyanins, in terms of resolving power, selectivity, and sensitivity [123,16,67].

1.1. Identification of anthocyanins in biological matrices

Anthocyanins can be easily identified by absorption spectroscopy, especially UV-Vis spectroscopy, which allows the characterization of their unsaturation and functional groups [123]. However, UV-Vis detectors fail to provide unambiguous identification for each peak, as the different metabolites in the separation process are often similar, and their response to radiation in this range of wavelengths is indistinguishable [123]. Anthocyanins exist as stereoisomers because of the presence of one or more chiral centers, which could render poor resolution and detection through UV-Vis. Thus, MS, especially LC-high-resolution MS (HRMS) has become a fundamental tool so far for the analysis of polymeric compounds (polyphenols) [123,125,23]. Therefore, this technology has advantages regarding chromatographic separation and accurate mass-based selectivity in resolving the complexity of natural matrices for accurate characterization [123]. Over time, LC-MS has become the preferred methodology for analyzing phenolic compounds as it has some advantages: i) it can separate quite complex mixtures of low and high molecular weight compounds, as well as different polarities and acid-base properties in various matrices, and ii) derivatization is not necessary [28].

1.2. Metabolomics for the identification of anthocyanins

Metabolomics is one of the platforms of the post-genomics era that constitutes the omics sciences. One of the objectives of this platform is to identify or annotate specific metabolites, investigate their abundance, and/or perform structural characterization of different molecules under different conditions [20]. This technology, like other omics technologies, can contribute to the understanding of functional compounds with health benefits and the productivity and nutritional status of crop improvement [96].

Metabolomics technologies allow a high-throughput analysis of different types of samples with quantitative and qualitative approaches by means of a) the identification of metabolites using reference standards (targeted); and b) the tentative identification of the compounds (untargeted). The latter approach is feasible, as it is based on the presence of precursor and fragment ions with high mass accuracy (mass error <5 ppm) [8,123,6,76]. Hyphenation of LC with MS and various ion sources is one of the most complete techniques for structural elucidation and investigation of phenolic components [109]. The advantage of the mass spectrometry technique is the ability to indicate the exact molecular formula of the substance without the need for large amounts of samples, as the technique can work with trace-level quantities [109]. Thus, this review article is focused on the liquid chromatography coupled to mass spectrometry (LC-MS) technique.

A summary of the latest research related to analytical tools based on LC-MS for the metabolomic study of anthocyanins in fruits is hereby presented. Although fruits are the main source of functional foods, as they are rich in several antioxidants such as anthocyanins, metabolomic studies are restricted to some fruits [96]. Consequently, our first aim is to describe common sample preparation methods adequate to metabolomics workflows. We also focus on recent methodologies employed for data acquisition in LC-MS (instruments and techniques). Subsequently, spectra processing will be detailed in an ensuing section of this article. In addition, recent advances in bioinformatics and chemometric tools for the interpretation of metabolomics data are discussed to provide a current perspective on the state of the art in anthocyanin identification and annotation.

2. Sample collection, extraction, and preparation for metabolomics analysis

Sample preparation is an important step that must be carefully defined and planned from the beginning of the experiment. Since it is related to other steps of the metabolomic analysis, it should, therefore, be carefully integrated with other analytical methods employed in the study [73]. During the sample preparation, it is also important to consider the physicochemical features of anthocyanin molecules, aiming to reduce the degradation of compounds and improve evaluation by optimizing experimental conditions [167]. Different steps for sample preparation will be further discussed in this section.

2.1. Sample collection

The collection of samples must be conducted quickly because of the rapid metabolic changes intrinsic to organisms. Hence, it must be followed by a method that completely stops cellular metabolism known as quenching [94]. An example is the flash-freezing of fruits with liquid nitrogen immediately after removal from the plant [73]. The fruit maturation stage is an important factor to consider during collection, as many fruits develop higher anthocyanin content after ripening [56]. In addition, positioning on the plant and solar radiation exposure can interfere with the levels of metabolites found in each fruit [117]. Collection parameters may vary according to the objectives of the study.

2.2. Sample drying

After collection, samples need to be dried. The drying stage aims at reducing the water levels of the plant material, thus inhibiting the proclivity of enzymatic and microbial activities. This process avoids the decomposition and/or degradation of the metabolites present in the samples, and they can be conserved for significantly longer periods [94, 73]. Among several drying methods, freeze drying is the most applied to vegetable samples [73]. In this technique, the sample is frozen and submitted to high vacuum conditions along with low temperatures, causing sublimation of water content [101]. Freeze drying performs better than heat-based drying methods in which the target compounds are susceptible to degradation under high temperatures, as is the case of anthocyanins [94,136]. Çoklar and Akbulut [24] performed three different drying strategies on black grapes and observed higher anthocyanin levels in freeze-dried units. Moreover, the authors reported a considerable decrease in anthocyanin concentration in oven-dried and sun-dried grapes. After the drying step, samples should be stored at low temperatures to avoid water reabsorption [94].

2.3. Anthocyanins extraction

Extraction is an important step in sample preparation. It enables the isolation of analytes or groups of molecules while removing interfering compounds [94]. As previously mentioned, the properties of anthocyanins are directly related to their structures. These structures preservation throughout extraction depends on several factors such as pH, temperature, and solvent. It is necessary to stabilize the flavylium cation, which acts as an acid and makes anthocyanins highly reactive. Accordingly, it is necessary to consider these factors to avoid a decrease in the quality and activity of the resulting extract [144].

Anthocyanins are commonly found inside cellular organelles, and prior procedures should be performed to favor their extraction, such as particle size reduction. Reducing particle size increases the contact area of the material and facilitates the entry of the solvent into the cells and organelles, allowing the diffusion of targeted compounds to the solvent [136]. Despite their reproducibility, classical extraction techniques can have negative outcomes: low yields; degradation of metabolites due to high temperatures; large sample demand; consumption of large volumes of mostly toxic and expensive organic solvents; production of large volumes of waste; difficulties with automation; lack of the possibility of making a direct connection between samples preparation and the analysis stages; need of additional operations on the sample (e.g. filtering the extract, its concentration, etc.); and finally the processes are time-consuming and laborious [144]. The most modern extraction methods are focused not only on reducing the consumption of organic solvents but also on preventing the degradation of sample components, as well as eliminating additional pre- and post-sample preparation steps, such as concentration level and homogenization [17]. New extraction techniques have been developed to bypass these hindrances by applying more sustainable, efficient, faster, and safer methodologies [144,14,37]. They are Enzyme-Assisted Extraction (EAE), Microwave-Assisted Extraction (MAE), Ultrasound-Assisted Extraction (UAE), Pressurized Liquid Extraction (PLE), Supercritical Fluid Extraction (SFE), and Supercritical carbon dioxide extraction (SCDE), among others [37]. However, despite recent developments, several studies claim that conventional extraction methods can have better extraction performance than modern ones. López et al. [83] observed that extraction with ethanol, for 5 min, at 90ºC, is more efficient than UAE for extraction of anthocyanin from *Arbutus unedo* fruits. Conventional and modern approaches commonly used for extracting anthocyanins from fruits are briefly discussed in Sections 2.4 to 2.11 and illustrated by some examples of current studies found in the literature (Supplementary Table 1).

2.4. Solvent Extraction Method (SEM)

Solvent extraction methods (SEM) are widely used for anthocyanins extraction and they are based on the principle of "like dissolves like" [144,37]. Anthocyanins have a polar character, so they are usually extracted with the use of polar organic solvents, such as water, methanol, ethanol, and acetone [114]. Since anthocyanins are more stable under acidic conditions (pH ~3) and tend to degrade under neutral and alkaline conditions (pH above 7), pH is an important aspect to consider in extraction [136]. Therefore, it is common to combine acidic solutions with solvents to decrease the pH of the medium, preserving molecular stability, mainly of non-acylated anthocyanins, and improving extraction quality [144]. However, the addition of strong acids can lead to the hydrolysis of anthocyanin glycosidic bonds, producing anthocyanidins and sugar [136,37]. Thus, weak acids such as formic, citric, and acetic acids may be more suitable, and the amount employed is usually less than 7 % of the solvent/acid mixture [10,144]. The water-solvent ratio and the different combinations of solvents and acids can result in the extraction of different yields, which also vary depending on the type of sample and its compounds [10,136]. In general, a solvent is considered suitable when it maximizes the yield of extracted compounds while still preserving their chemical nature [4].

Conventional methods of solvent extraction of anthocyanins include mainly maceration, infusion, digestion (also known as heat-assisted extraction (HAE)), decoction, percolation and filtration, and Soxhlet extraction. Maceration is the procedure of soaking a powdered sample in the proper solvent, under specific conditions. This method requires subsequent separation of solvent and solid particles [144,2,4]. The infusion consists of a maceration process performed with water. The digestion method involves maceration under mild heating conditions. The decoction extraction is based on an infusion technique carried out with boiling water [144]. The percolation technique involves the use of a percolator device in which the solvent moves continuously through the sample powder, from top to bottom, and a filtration procedure is applied subsequently [4]. In the Soxhlet extractor, the condensed solvent reaches the sample after the system is heated, and then returns to the distillation flask, carrying the extracted compounds. This procedure is repeated cyclically until extraction is complete [84].

In summary, SEM has some advantages: convenient operation, simple equipment, and ease of implementation in the extraction of anthocyanins. However, despite being widely used in the industry, the application of SEM presents a few significant disadvantages: it is time-consuming, has low efficiency, and has high applied temperature and solvent consumption [142,144].

2.5. Deep Eutectic Solvents (DES)

DES has been used nowadays as an alternative to organic solvents. They are produced by combining naturally occurring components that are capable of hydrogen bond interactions as hydrogen bond acceptors or donors. Natural Deep Eutectic Solvents (NADES), a type of DES, are mixtures of primary metabolites such as sugars, organic acids, alcohols, amino acids, and amines, united by strong intermolecular interactions, particularly hydrogen bonds, together with other interactions such as van der Waals and electrostatic forces [150]. The use of NADES composed of combinations of quaternary ammonium salt-alcohol, quaternary ammonium salt-carboxylic acid, and carboxylic acid-sugar to extract anthocyanins from plant matrices has been registered. The most widely used, both in binary and ternary combinations, are citric acid, malic acid, lactic acid, glucose, fructose, xylose, glycerol, and choline chloride [150]. Low cost and ecological impact are among the advantages of DES [102]. Jeong et al. [58] reported that the most effective extraction was performed using the eutectic solvent formed by citric acid as hydrogen bond acceptor, and maltose as hydrogen bond donor, whereby 63.3 mg/g of anthocyanins were extracted. On the other hand, Bubalo et al. [11] found that the most effective combination to extract anthocyanins was choline chloride and oxalic acid.

2.6. Enzyme Assisted Extraction (EAE)

Since anthocyanins are often found inside plant cell organelles, it is extremely important to access the interior of these compartments during extraction [136]. EAE is based on inducing the disruption of plant cell walls by enzymatic activity, in order to release target compounds stored inside the cells [144]. It is a simple method with advantages such as swiftness, efficiency, use of lower amounts of solvent, low instrumental costs, low energy consumption, and it is an eco-friendly method [87,47]. The main types of enzymes used are cellulases, pectinases, hemicellulases, and proteases [87]. Pectinases, widely used in the extraction of compounds from plant matrices, assist in the extraction process by degrading the pectin present in cell walls and structures [47,74].

EAE can be combined with other extraction techniques, such as Ultrasound-Assisted Extraction (UAE) in order to obtain higher yields in the extraction of anthocyanins [174]. Enzyme extraction protocols have been developed and optimized for anthocyanins extraction from different types of fruit, determining the ideal conditions to achieve the highest yields [47,74]. For instance, according to González et al. [47], the optimal conditions for blackcurrant EAE are 10 min of enzyme

extraction time; enzyme amount of 91.0 enzyme units per gram of sample using pectinase from *Aspergillus niger*; 10 % ethanol in water (pH 4.1); at 30 °C. In the mulberry fruit EAE, the optimal conditions were a 1:5 w/v material/water ratio with an enzyme concentration of 3.5 % pectinase Pectinex Ultra SP-L (v/w) and 1.5 % citric acid (w/w), 113 min of reaction time, at 50 °C [74].

2.7. Microwave-Assisted Extraction (MAE)

MAE uses microwaves that promote heating and vaporization of the water contained in the cell matrix, increasing the pressure on the cell walls, and finally inducing their rupture. Additionally, microwaves also raise the temperature of the solvent, which can reduce the viscosity of the extract and promote the solubility of the target components [168]. Thus, the entry of the solvent into the cell structures and the diffusion of the compounds become possible [144]. MAE is an efficient technique for extracting compounds, with advantages such as swiftness, high yields of anthocyanins, good reproducibility, and low consumption of solvents. It is considered eco-friendly compared to conventional SEM [106,139,142, 151]. However, the higher diffusion rate promoted by this method also increases the extraction of non-targeted compounds [136]. In MAE, it is important to consider the variables: i) solvent/water ratio; ii) solvent composition and volume; iii) temperature; iv) extraction time; v) microwave power and sample mass (W/g); vi) and pH value [106,163,39, 45].

Fernandez-Aulis et al. [39] compared different MAE parameters in a corn cob and corn husk. The authors observed higher yields of anthocyanins in a solvent system of ethanol:water:lactic acid (80:19:1), in the extraction time of 60 s, using a microwave at 600 W. The yield obtained by applying MAE was analogous to the yield obtained by applying two subsequent Maceration Extraction (ME), indicating that the MAE uses less time to obtain a higher yield. Also, this MAE workflow was considered efficient in the breakdown of cell walls [39]. Moreover, the anthocyanin profile was the same in the different extraction methods assessed, comprised mainly of glycosylated and acylated anthocyanins, the latter with malonic acid [39]. Vilas-Boas et al. [151] demonstrated that the MAE technique provided higher yields in the recovery of anthocyanins in Portuguese sweet cherries (Saco variety) compared to the UAE and conventional extraction techniques. This result may be related to the homogeneous heating present in the MAE technique [151].

2.8. Ultrasound-Assisted Extraction (UAE)

This approach is based on the use of ultrasound frequencies (from 20 to 2000 kHz) that induce the cavitation phenomenon and the production of bubbles that gradually increase in size and implode, leading to cell wall rupture due to increasing pressure and energy of the system [144, 4]. UAE can improve the yields obtained for anthocyanins extraction to a certain extent when compared to traditional SEM. The method demands careful implementation because the mechanical effects can destroy the structure of anthocyanins [142]. Important optimal conditions to be determined are solvent composition, temperature, extraction time, solid-liquid ratio, ultrasound power, and the number of extractions [118,21,39,45,82]. Nonetheless, some advantages are evident: it is a simple and eco-friendly method, with good reproducibility, lower solvent and energy consumption, and greater automation in comparison to other methods [37,4].

Chen et al. [21] reported that the best conditions for the extraction of anthocyanins from purple corn were 40ºC, ultrasonication for 90 min set at 400 W, and a solid-liquid ratio of 1:8. Through tandem LC-MS analysis, six types of anthocyanins were identified, with a predominance of cyanidin-3-*O*-glucoside and cyanidin-3-*O*-(6-malonyglucoside). The study by Fernandez-Aulis et al. [39] also compared UAE, MAE, and ME. It was observed that the most efficient method was the UAE under the following conditions: ethanol/water/lactic acid (80:19:1), 1:10 solvent to solid ratio, at 25ºC, ultrasound power of 100 W, 20 min, performing 2 subsequent extractions for obtaining anthocyanins from corn cob and corn husk. Lončarić et al. [82] observed that the extraction of anthocyanins from blueberry pomace using the UAE resulted in high yields under the following conditions: using methanol-based solvent, at 80ºC, 35 kHz, 15 min. Subsequently, 7 anthocyanins were identified in blueberry extracts and peonidin-3-glucosidase was the most abundant. UAE combined with EAE was performed by Zhang et al. [174] for the yield of anthocyanins from mulberry wine residues. The optimal conditions were determined as 52ºC, ultrasonication power of 315 W, 0.22 % enzyme, and 94 min. Furthermore, through UPLC-MS analysis, two main anthocyanins were identified in mulberry wine residue extracts: cyanidin-3-*O*-glucoside and cyanidin-3-*O*-rutinoside. The high dosage of enzymes promoted enzymolysis, accelerating the dissolution of anthocyanins [174].

2.9. Pressurized Liquid Extraction (PLE)

Pressurized Liquid Extraction (PLE) is known as Accelerated Solvent Extraction (ASE), or, when the solvent used is water, it is also known as Subcritical Water Extraction (SWE). This method uses high pressure to keep the solvent in a liquid state at high temperatures, enhancing the solubility of the compounds and facilitating solvent extraction [144,4]. Extracts obtained by means of PLE are clean with minimal background noise during analysis [4]. In addition, PLE requires lower extraction time and solvent amounts. [144,4]. Despite these advantages, the high temperatures applied in this technique can induce the degradation of anthocyanins. Therefore, temperature is a parameter that must be carefully established and optimized [136].

In a study carried out with wild blackberry fruits, the PLE showed a 29 % increase in anthocyanin yield when compared to the UAE, demonstrating that PLE can be efficient under the conditions used (50 % methanol acidified with 1 % acetic acid; 50ºC; 100 bar; 5 repetitions) [103]. Another study comparing extraction methods showed that PLE is more efficient and adequate for the extraction of anthocyanins from colored grains than MAE. The optimal conditions of PLE for extraction of anthocyanins from colored grains were 50 °C, 2500 psi, 10 min using 5 cycles, and 100 % flush for black rice, blue wheat, and purple corn [1]. The optimal conditions of PLE for extracting anthocyanins from jambolan fruit were 90 °C, ethanol 80 % v/v (trifluoroacetic acid (TFA), 0.1 %), 2 extraction cycles, obtaining the content of 10–45 mg cyanidin-3-*O*-glucoside per gram of dry fruit [121].

Machado et al. [88] evaluated the yields of phenolic compounds in blackberry residues obtained by PLE and by conventional methods (Soxhlet and maceration). They observed that PLE, despite being more effective in terms of yielding phenolic content, did not render the highest values of anthocyanins. In this case, the extraction using the Soxhlet technique and ethanol as solvent afforded the highest yield of monomeric anthocyanins [88]. The higher phenolic content and lower yields of anthocyanins obtained with this high-temperature approach could be explained by the thermal sensitivity of anthocyanins [103, 136].

SWE uses water below its critical points of temperature and pressure (100–374.15 °C and 22.1 MPa) [22]. Under these conditions, the dielectric constant and viscosity of water reduce when temperature increases, whereas the diffusion coefficient increases at high temperatures. This allows the extraction of compounds with different polarities by modifying temperature and pressure settings [22]. The SWE is an inexpensive, non-toxic and eco-friendly technique that can be optimized through a careful selection of variables such as temperature, pressure, repetitions, extraction time, and solvent pH [144,69]. Kang et al. [69] investigated the optimal conditions for the extraction of anthocyanins from blueberry (*Vaccinium corymbosum*) and chokeberry (*Aronia melanocarpa*) by means of SWE. Thirteen anthocyanins were identified from blueberry extracts and the optimal conditions to obtain higher contents of anthocyanins from these were 130 °C, for 3 min, using water and 1 % citric acid as solvent. For chokeberries, the optimal conditions were

190 °C, for 1 min, with the use of water and 1 % citric acid as solvent. Eight anthocyanins were identified from these fruits [69]. The authors also reported that SWE provided higher concentrations of anthocyanins than pressing and hot water classical extraction techniques. Wang et al. [155] reported that the extraction efficiency of anthocyanin from raspberry through SWE was significantly higher than that from hot water or methanol extractions.

2.10. Supercritical Fluid Extraction (SFE)

Supercritical Fluid Extraction (SFE) is a technique that uses supercritical fluids, i.e., fluids at temperature and pressure above their critical points in order to extract compounds of interest from plant matrices [19,100]. These supercritical conditions promote low viscosity and high diffusivity of fluids and contribute to faster mass transfer and extraction rates [100,135]. The SFE presents several advantages: it is an eco-friendly technique, it uses solvents that are generally recognized as safe, it has reduced extraction times, and provides higher yields of targeted compounds [135]. In addition, the extraction process occurs in light- and atmospheric oxygen-controlled environments, preventing anthocyanins oxidation [136].

Carbon dioxide (CO_2) is commonly used as a solvent for SFE. In this case, the technique is known as Supercritical Carbon Dioxide Extraction (SCDE). SCDE has been mainly applied in food matrices because of the following characteristics: i) it is non-toxic, and considered safe for the environment and human health [100,135]; ii) it is non-flammable [19]; iii) it is reusable, which reduces energy costs and consumption [71]; iv) its critical temperature is moderate (31.3 °C), avoiding degradation of thermolabile compounds [19,135]; v) and native enzymes that induce anthocyanins degradation are inhibited by applying supercritical carbon dioxide ($scCO_2$) and pressure [136]. The $scCO_2$ has advantages over other types of traditional solvents: strong solubility, high mass transfer coefficient, non-toxic, high-cost performance, and easy access [113]. As the SCDE technique uses low temperature, the obtainment of anthocyanins that are thermolabile is favored [142]. Despite the numerous benefits of this method, its application on a large scale and in the industrial sector, such as the food industry, is unlikely due to the high cost of equipment and large technology investment. In addition, the extraction parameters (CO_2 pressure and velocity) are difficult to control.

$scCO_2$ is a non-polar solvent, hence, the extraction of polar compounds such as anthocyanins is limited. Therefore, to extract anthocyanins and other polar compounds a polar co-solvent must be added in order to increase the extraction efficiency [37,100,173]. Ethanol, methanol, and water are commonly used as co-solvents, generally in small amounts, with ratios depending on the plant material [100,136,144].

To increase extraction efficiency, it is important to consider some parameters, such as temperature, pressure, type of co-solvent and its volume, extraction time, and flow rate of CO_2, among other factors [144]. Paes et al. [104] investigated the optimal conditions for anthocyanins recovery from blueberry (*Vaccinium myrtillus* L.) residues through SFE with $scCO_2$, and observed that higher yields were obtained with 5 % water and 5 % ethanol as co-solvents, at 40 °C, 20 MPa and CO_2 flow rate of 1.4×10^{-4} kg/s. The authors pointed out the importance of water as a co-solvent to improve anthocyanins extraction since they are highly soluble in water and the interaction of water and CO_2 would possibly contribute to an acidic environment [104]. A total of sixteen anthocyanins were identified from blueberry residue extracts and higher concentrations of anthocyanins were obtained by means of SFE when compared to PLE [104].

The SFE is often combined with other techniques, such as pressurized liquid extraction or ultrasound, in order to increase efficiency and anthocyanins yield. Garcia-Mendoza et al. [42] demonstrated that performing PLE followed by SFE using an acidified solution of ethanol and water (50 % v/v, pH 2.0) as co-solvent is an effective strategy for obtaining anthocyanins from juçara residues (*Euterpe edulis* Mart.). Higher yields were obtained at moderate temperatures (40 °C), and four anthocyanins from juçara residues were identified. In a study with blackberry (*Rubus* sp.) residues, ultrasound application during SFE and the addition of ethanol and water as co-solvents enhanced the process of anthocyanins extraction, minimizing extraction time and increasing yields of target compounds [115]. There were four major anthocyanins identified from blackberry bagasse and the highest content of anthocyanin was achieved with SFE combined with ultrasound having water as co-solvent at 60 °C, 15 MPa, and 200 W. This method provided higher anthocyanin content in comparison with Soxhlet and maceration methods [115].

2.11. Pulsed Electric Fields (PEF)

This is a non-thermal process based on the application of external electric fields, improving the permeabilization of the cell membranes. Consequently, it enhances the release of liquid and targeted compounds from the cells. Electroporation of cell membranes is commonly achieved at electric fields lower than 10 kV/cm and specific energies lower than 10 kJ/kg [111]. However, different parameters have been reported. Zhou et al. [176] optimized a PEF extraction process in blueberry, and the most efficient condition was at 20 kV/cm to extract 5.6 mg/g. Gagneten et al. [40] reported that, under optimal conditions, an increment of 6 % of total monomeric anthocyanins was extracted from blackcurrants using PEF.

2.12. Steps after extraction

After extraction, if storage is required, the extract should be kept in the dark at a low temperature. Finally, one step in between sample extraction and LC-MS data acquisition is sample preparation. Some sample types are essential to an investigative metabolomics workflow. For more adequate analysis with LC-MS, it is important to use blank samples (BS), both for mobile phases and for extraction solutions (used to subtract data unrelated to the samples), and quality control samples (QC) [36]. QC can preferably be a pool of the samples of interest, but also synthetic mixtures or standard compounds. They are used to assess the stability of the LC-MS system and the quality of the data during acquisition and post-acquisition handling. After LC-MS acquisition of the investigated samples, QC and BS integrate data pre-processing steps. The use of this strategy was observed in some of the studies mentioned in the present review article [123,134].

3. Data acquisition through LC-MS in fruit metabolomics

As a result of the complementary advantages of chromatographic separation and accurate mass-based selectivity, the use of LC-MS as an analytical tool has become increasingly popular over the past decade and has provided considerable insight into the identification and quantification of polyphenols components in complex biological samples, such as fruit extracts [123]. Regarding the study of anthocyanins in fruits, several recent studies use LC-MS-based methods for a better understanding of the color formation, metabolic changes during fruit development, postharvest quality control, and the overall metabolic profile present in the extract. Hence, the goal of this section is to discuss recent data acquisition methodologies used in LC-MS applied to the screening and discovery of anthocyanins in fruits.

3.1. LC types and columns

Separation is a crucial aspect of metabolomic measurements. As plant extracts are remarkably complex, it is sometimes necessary to purify the samples to carry out the identification of the substances. The metabolites are separated in LC as they elute from the device at different times, thus reducing the complexity of biological samples prior to MS detection [9]. Typically, phenolic compounds separation by LC is

performed in the reversed-phase (RP) mode using columns generally filled with particles of silica bonded with alkyl chains (C8 or C18), in which the compounds are eluted according to their polarity [85]. In the surveyed studies of anthocyanin in fruits, C18 is the most used and column particle size mainly ranges from 1.7 μm to 3.0 μm [126,166,172, 25,38,60]. For instance, Yang et al. [166] used a CSH C18 column for an untargeted metabolomics study aiming to identify and quantify metabolic changes in sweet cherry development and ripening. Results show an increase in flavonoids, including the anthocyanin cyanidin [166]. The use of hydrophilic interaction as a separation method for LC (HILIC) has been developed as an alternative for the targeted analysis of diverse anthocyanins in blueberries and grape skins, among other sources [156]. The authors evaluated a range of different stationary phases and reported that the BEH Amide phase showed the best performance in terms of both retention and selectivity. HILIC was found to provide an alternative separation mechanism to RP-LC, with clear separation according to the degree of glycosylation and acylation, allowing for tentative compound identification [156]. HILIC was designed to retain polar analytes and metabolites that are too polar to be retained by RP, which is often used in the separation of nonpolar to slightly polar molecules [12].

3.2. Mobile phases

The analytes are carried through the LC column by a liquid mobile phase (solvents) and eluted from it when they have a higher degree of chemical affinity with the mobile phase than the stationary phase [9]. Ideal chromatographic conditions should use mobile phases that help promote ionization of as many analytes as possible, retaining and separating the complex mixture within samples [164]. In general, the separation of phenolic compounds is carried out by gradient elution using binary systems that include an aqueous component (solvent A) and a less polar organic solvent (solvent B) such as methanol or acetonitrile [85]. Most studies reviewed in this section show the same results [25,41, 60,126,162,166,172]. Alternatively, Singh et al. [137] performed a comparative study applying gradient and isocratic elution methods for the separation and identification of mixtures comprising up to 10 anthocyanins in jamun, blackberries, and other sources. The isocratic method failed to resolve all the peaks when the number of anthocyanins was high, whereas the gradient elution method successfully resolved all the peaks independent of the number of anthocyanins [137]. Moreover, as the composition of the solvent influences the gas-phase acid-base processes required for electrospray ionization, it is also important to consider the addition of acids in the mobile phase in order to control pH (and therefore the charge of the molecules) [29,32]. In contrast to the desired anthocyanins structure stability in the extraction step, highly acidic mobile phases are applied to guarantee the predominance of the flavylium cationic species (AH+) and increase chromatographic performances [156]. The reviewed studies mostly used formic acid and acetic acid in one or both solvent phases, and their concentrations varied from 0.04 % to 5 % (v/v) (Supplementary Table 2). Deineka et al. [32] reported that the substitution of formic acid by phosphoric acid for the acidification of an acetonitrile-water mobile phase leads to a substantial increase in anthocyanins retention, although both systems have presented similar properties. TFA was also studied by Willemse et al. [156] as an acidic mobile phase component for HILIC-LC (0.4 % v/v) and the peak shapes of all samples improved greatly in comparison to the formic acid phase (10 % v/v). This could be largely attributed to the lower polarity of the solvent containing TFA [156]. As for the mobile phase solvents reported in the surveyed literature, water, methanol, and acetonitrile were used in different proportions (Supplementary Table 2). Acetonitrile is typically preferred over methanol due to the lower resulting pressure in the column [89]. Nonetheless, the use of the latter did not cause problems related to high pressure, as this was not addressed in studies that implemented methanol in the mobile phase [25,41]. Molecules eluted from the LC are ionized in MS ion sources, which are further discussed below.

3.3. MS ionization sources

One of the most important steps in MS-based metabolite measurements is ionization. The level of ionization of a metabolite determines its proclivity to detection and quantification [49]. Soft ionization techniques have been developed for nonvolatile or thermodynamically unstable natural compounds [123]. By applying this method, intact molecular ions and/or a few fragments are generated in MS1 [25]. Electrospray ionization (ESI), matrix-assisted laser desorption ionization (MALDI), and atmospheric pressure chemical ionization (APCI) hyphenated to mass analyzers have been reported for the characterization of anthocyanins [123,25]. Among all the LC-MS hyphenated techniques available, the application of ESI is the most common for the analysis of anthocyanins in fruit samples (Supplementary Table 2). An important feature of ESI is that the samples must be introduced in the form of a solution, allowing the coupling to many separation techniques, such as ultra-performance liquid chromatography (UPLC). For example, both targeted and untargeted metabolomics of *Myrtus communis* L. berries were investigated through LC-ESI-MS and it was possible to identify ten anthocyanins glycosides, among other flavonoids [126]. Similarly, using an ESI-MS target metabolomics approach, another study detected 11 anthocyanin-related metabolites with significant differences among the developmental stages of fruit peel [60]. On the other hand, Commisso et al. [25] compared ESI and APCI performances for both targeted and untargeted metabolomic analysis of Corvina berries. The authors reported that APCI was particularly suitable for the ionization of strongly polar metabolites with lower retention time, such as sugars and organic acids, whereas ESI seemed to ionize molecules regardless of their polarity. The latter, though, was more efficient in ionizing moderately and weakly polar metabolites such as flavanols, flavones, and both glycosylated and acylated anthocyanins. However, ESI was more susceptible to matrix effects than APCI [25]. As reported before, moderately polar metabolites, especially strongly polar metabolites, are subject to extreme matrix effects when using an ESI source [145].

Anthocyanins, unlike other flavonoids or polyphenols, exist in cationic forms. Thus, positive ionization is extensively used in ESI analyses. Moreover, the MS spectra acquired in the negative mode have also proved to be valuable tools for the differentiation of anthocyanins from non-anthocyanin polyphenols, such as flavonol glycosides [138]. In recent years, both ionization modes have been commonly reported for metabolomic analyses of anthocyanins in fruit samples (Supplementary Table 2).

3.4. Mass analyzers and tandem mass spectrometry

The ions produced in the MS source are separated according to their mass-to-charge ratio (m/z) in the analyzers. Therefore, they provide an accurate measurement of the masses of ionized biomolecules [9]. The ability to distinguish ions of very similar m/z values is determined by mass resolution. High-resolution instruments provide mass resolution above 10,000 (m/Dm, FWHM), thus excluding quadrupole mass filter, triple quadrupole (QqQ), and quadrupole ion trap (Q-TRAP) mass analyzers [159]. By enabling precise mass measurements with sub-ppm errors, high mass resolution analyzers improve the sensitivity and information content of global approaches. High-resolution mass spectrometry includes time-of-flight (TOF), Fourier transform ion cyclotron resonance (FT-ICR), and Orbitrap-based instruments [65]. This review article discusses both high- and low-resolution LC-MS approaches applied toward the annotation of anthocyanins in fruits (Supplementary Table 2).

There are also several options including single (MS) or tandem (MS/MS) mass analyzers to choose from, each with different sensitivity and resolution performance. Hyphenation of ESI with ion trap (IT), QqQ, TOF, quadrupole time-of-flight (Q-TOF), and Orbitrap are frequently reported for the analysis of polyphenols, particularly anthocyanins [85, 102]. According to Gowda and Djukovic [49], QqQ and Q-TRAP are the

most common analyzers in targeted metabolic studies due to their high sensitivity and selectivity, whereas Q-TOF and linear trap quadrupole orbitrap (LTQ-Orbitrap) are more suitable for an untargeted approach because of their higher mass-resolving power [49].

Multiple reaction monitoring (MRM) refers to an MS/MS mode able to select, identify and quantify predefined ions for analysis when coupled to QqQ [62]. MRM quantification involves the inspection of individual chemical species in terms of their parent and product ions, the latter generated through collision gas application in a mass analyzer, as expected reactions to be monitored [78]. This method can significantly reduce analysis time and still provide excellent sensitivity, especially regarding phenolic compounds [62]. In this respect, MRM combined with Q-TRAP has been reported in numerous studies that integrated transcriptomic and metabolomic analyses in fruits, considering color formation and color change during developmental stages, whereby anthocyanins were identified [133,172,41,59,60,81]. Alternatively, Sarais et al. [126] also made use of multistage mass spectrometry (MS^n) for both ionization modes to get information on the structural features and the glycosylated forms of phenolic compounds. MS^n is highly suitable for the structural characterization of glycoconjugates in the case of phenolic compounds because different modifications can potentially produce a complex combination of isomeric and isobaric compounds [109,66].

4. Bioinformatic and chemometric tools for the interpretation of metabolomics data

Metabolomics data processing and interpretation rapidly evolved along with the improvement and diversification of LC-MS analytical techniques. Nonetheless, some analysis workflows are preferably applied to specific case studies, and they are considered stable, unbiased, and reproducible [48]. Flavonoids have well-established structures distinguishing their subclasses and, therefore, the elucidation of anthocyanins through mainly LC-MS/MS approaches is based on the annotation of their fragments and precursors *m/z* and/or retention time patterns [131,170]. The workflows generally adopted for the annotation of anthocyanins are consistent and show rather relevant results. Even though chemometric and statistical analysis parameters are widely disclosed, thorough chromatographic and mass spectrometry raw data pre-processing, processing, and molecules annotation parameters are scarce among reviewed studies. Still, a rich compilation of bioinformatic and chemometric strategies and their intricacies aimed at the annotation of anthocyanins in fruits are hereby presented. A general base workflow is shown in Fig. 1.

4.1. Data pre-processing

LC-MS experiments result in a high throughput complex raw data output. The next step is a daunting filtering and meticulous organizational procedure, often referred to as data pre-processing, in which patterns from detected data signals are evaluated, and relevant features are distinguished from the noise [107]. Instrumental conditions can vary, which affects overall performance and results. Hence, strategies to overcome such proclivity are indispensable. As files respective to each sample must be aligned, it is a priority in the pre-processing stage to define significant similarities and differences among these data matrices by considering acquisition and sample unevenness. Dedicated algorithms can resolve peak detection and alignment and perform retention time corrections [107]. These algorithms are usually integrated into raw LC-MS data interpretation software, of which some were used to pre-process data for the annotation of anthocyanins in fruits (Supplementary Table 3).

Among the few studies that provided detailed pre-processing parameters, attention was given to the steps of noise removal, blank subtraction, and constraint parameters for retention times and *m/z* alignments (see References in Supplementary Table 3). Tohge et al. [146] applied a strategy for noise removal based on the internal compounds calibration, while Allwood et al. [5] set the signal-to-noise ratio (S/N) parameter to 6. Renai et al. [116] set a retention time tolerance for data alignment of 0.05 min, while Calumpang et al. [13] defined it as 20 s (0.33 min). Parameter adjustments for each of the pre-processing steps occurred on a case-by-case basis. Such settings require thorough exploratory analysis, capacity for data pattern recognition, and are dependent on each software demand. Nonetheless, pre-processing aiming at precursory data filtering, cleaning, and reduction was reported in most of the literature regarding fruits anthocyanins metabolomics through LC-MS.

4.2. Data Normalization, scaling, and transformation

LC-MS metabolomics data have a 3D (three-dimensional) structure, consisting of *m/z*, retention time, and signal intensity. After pre-processing, software usually returns noise-cleared 2D and/or 1D data tables to be implemented in the subsequent chemometric analysis and metabolites annotation. The objective of performing data normalization is to evaluate each row from the returned data tables by a mathematical factor [70]. This process allows the reduction or removal of fluctuation so that samples can be cross-analyzed quantitatively or semi-quantitatively. Normalized data fit as close as possible to a Gaussian distribution. Further adjustments are performed through data scaling and transformations, which are usually based on the statistical analysis or algorithm to be applied [90,70].

Normalization based on internal chemical standards was one of the strategies observed within studies reporting the presence of anthocyanins in fruits through LC-MS metabolomics [30,31,57,154]. Internal standards cited within these studies are reserpine (in positive ionization mode), morin (negative ionization mode), and ampicillin (no ionization mode disclosed). The known value of internal standards concentration functions as the basis for the normalization of overall data. Yet, this procedure is not recommended for untargeted studies [70]. Sample weight normalization was also observed in some studies [149,171,26]. Degu et al. [31] and Wang et al. [154] performed both internal standard and sample weight normalization. Some studies specified that the sample weight used was dry or fresh, whereas others simply referred to it as weight. Other normalization methods applied can be outlined as data-based, using average, median, log-based intensities, and total ion count (TIC) values as factors [93,25,77]. Moreover, less applied normalization methods within workflows aimed at anthocyanins annotation in fruits via LC-MS metabolomics were the extreme difference method [97]; R scale normalization (a built-in R software normalization function within the base package, which uses as a calculation basis the mean and standard deviation of the data, [161]); and range method

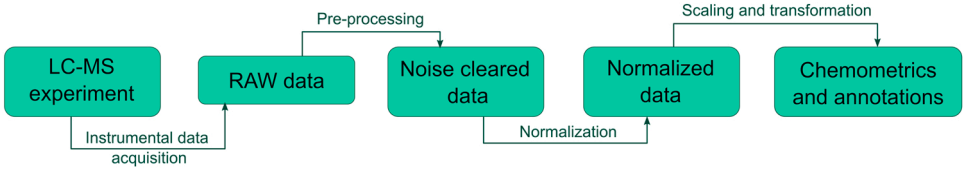


Fig. 1. General base workflow with usual steps after metabolomics data acquisition, until chemometric analysis and metabolite annotations. The steps were adapted from van den Berg et al. [34] and Karaman [70].

normalization [170,60]. Observed scaling strategies were Pareto and autoscaling, and most applied data transformation strategy was based on log values [7,147]. Autoscaling applies the calculated mean and standard deviation of the features, which are thereby mean-centered and attributed equal weights. This strategy allows for the equal contribution of low- and high-abundance metabolites within multivariate models [70]. In its turn, the Pareto scaling follows a similar workflow as the autoscaling, though the mean-centered values are divided by the square root of the standard deviation. This strategy tends to highlight the features with higher abundances [70].

4.3. Metabolites databases for the annotation of anthocyanins

A myriad of databases has been cited in studies with the successful annotation of anthocyanins from LC-MS metabolomics data (Table 1). Most studies applied a combination of several of these resources [95, 143,35]. None of them were databases completely dedicated to the compilation of anthocyanins and/or fruits. However, in-house-built databases were often cited [175,5]. Several studies have applied annotation by means of comparing chemical standards, and it is also the basis for quantification with MRM, which will be discussed in the next section [61]. Moreover, another annotation strategy was the combination of MS/MS evaluation and UV-Vis spectroscopy results, although the latter has not been discussed in the present review article [99,123].

4.4. Anthocyanins annotation and quantification strategies

According to Schrimpe-Rutledge et al. [127], annotations based on pure chemical standards are classified as level 1. Those annotations based on tentative processes that apply *in-silico* chemical structure fragmentation (theoretical *m/z*) as the basis for comparison with MS/MS experimental data are classified as level 2. Nour et al. [99] report a minute strategy applied in the method of anthocyanin annotation from blackcurrant fruits. The authors annotated the anthocyanins delphinidin 3-glucoside, delphinidin 3-rutinoside, cyanidin 3-glucoside, cyanidin 3-rutinoside, petunidin 3-rutinoside, pelargonidin 3-rutinoside, peonidin 3-rutinoside, petunidin 3-(6-coumaroyl)-glucoside and cyanidin 3-(6-coumaroyl)-glucoside. They listed the obtained *m/z* of precursors and fragments, associated with the sugar unit losses for the prediction of the metabolites within positive ionization mode analysis. The annotation of delphinidin 3-glucoside was established as the positive molecular ion with precursor with *m/z* 465 and a fragment at *m/z* 303 (the aglycon delphinidin), which was formed through the loss of one glucosyl unit with *m/z* 162. Another example was the annotation of cyanidin-3-glucoside: a precursor with *m/z* 449 and a fragment ion with *m/z* 287 (indicating again the loss of a glucosyl unit with *m/z* 162). These inferences were grounded by other experiments reported in the literature and they also applied UV-Vis spectra with respective external standards [99]. In general, most studies that applied level 2 annotation strategies have done so by checking isotopic patterns and comparing *m/z* values, retention times, and fragmentation patterns (so peculiar in the case of flavonoids) to the metabolites present in the databases (Table 1). As mentioned in Section 3.3, these studies often only considered the ESI positive ionization mode (M+), and respective adducts (e.g. [M+H]+, [M+NH_4]+, [M+Na]+, and [M+K]+) [86,110, 63]. Nonetheless, ESI negative ionization mode and respective adducts were also considered for the annotation of anthocyanins [146,148].

By applying an LC-Q-TOF-MS approach, Scorrano et al. [128] profiled the anthocyanins in *M. communis* berries, annotating the following molecules: delphinidin 3-*O*-glucoside, petunidin 3-*O*-glucoside, and malvidin 3-*O*-glucoside in higher concentrations (> 24.3 %); and delphinidin-pentose, delphinidin-pentose, cyanidin 3-*O*-glucoside, petunidin-pentose, petunidin-pentose, and peonidin 3-*O*-glucoside in lower concentrations (< 6.3 %). Their annotation workflow was based on the comparison of metabolites spectra from samples, including their fragmentation, with spectral data acquired from literature. Also, a

Table 1
Tools and metabolites databases for the annotation of metabolomics LC-MS data. The references column refers to literature with the record of the respective software used in the workflow of LC-MS metabolomics of fruits aiming at the annotation of anthocyanins.

Tools and databases, license type	Main functionalities	References
Bruker MetaboBASE® (https://store.bruker.com/products/bruker-metabobase-personal-library-3–0), proprietary	Spectra data from more than 100 K from multiple classes and sources. 700 K+ *in silico*-generated spectra are available in the database.	[116]
CFM-ID ([152], https://cfmid.wishartlab.com), free	Online *in-silico* application for accurately predicting ESI-MS/MS spectrum for compound structures.	[143]
Chemspider ([105], http://www.chemspider.com), free	114 M+ chemical structure database, including chemical properties, and spectral data.	[44,95, 154]
FooDB (https://foodb.ca), free	Comprehensive database on food constituents, chemistry, and biology. Includes structural chemical information.	[143]
GNPS ([98], https://gnps.ucsd.edu), free	Open-access online platform for metabolites dereplication, through a community-wide database, including molecular networking, experimental MS/MS data, and *in-silico* annotation tools.	[116]
HMDB ([157], https://hmdb.ca), free	Small molecules found in the human body with detailed information.	[175,79, 61]
KNApSAcK ([3], http://www.knapsackfamily.com/knapsack_core/top.php), free	Tools for analyzing mass spectra data sets, with queries of metabolites names, organisms, molecular weight, molecular formula, and more.	[35,134, 172]
KomicMarket ([55], http://webs2.kazusa.or.jp/komicmarket/index.php), free	High-resolution LC-MS sample-centric database.	[13,146]
Kyoto Encyclopedia of Genes and Genomes (KEGG) ([68], http://www.kegg.jp), free	Online database with information on small molecules relevant to biological systems (formula, mass, mol files), including molecular interaction and reaction networks information.	[169,158, 160]
Manchester Metabolomics Database (http://dbkgroup.org/MMD), free	Information about 42 K+ endogenous and exogenous chemical species.	[5]
MassBank ([54], https://massbank.eu/MassBank), free	Mass spectra repository of small molecules (< 3000 Da), including a variety of electrospray ionization (ESI)-MS^n, authentic, volatile, natural, and synthetic compounds data.	[129,35, 134,172]
MetFrag ([120], https://ipb-halle.github.io/MetFrag), free	*In-silico* online predictor of molecular structure fragmentation focused on the annotation of tandem mass spectrometry data.	[95]
MetFusion [43]* **	Combination of annotation results from several resources, including MetFrag and MassBank. Discontinued, but functionalities were integrated into the MetFrag project.	[95]
METLIN ([51], http://metlin.scripps.edu), free	Annotation processing tool and database with fragment similarity search function, including reference	[35,77,41]

(*continued on next page*)

Table 1 (*continued*)

Tools and databases, license type	Main functionalities	References
	standards and labeled stable isotope analogs databases.	
MetWare (https://www.metware.cn), proprietary	Custom and dedicated in-house metabolites databases.	[134,61,41]
MoTo DB ([91], http://appliedbioinformatics.wur.nl/)* ,* *	LC-MS-based metabolomics database for tomato (*Solanum lycopersicum*).	[146,35,172]
MZCloud (https://www.mzcloud.org), free	Online application for the annotation of small molecules using tandem mass spectrometry.	[143]
Phenol-Explorer ([119], http://phenol-explorer.eu), free	Comprehensive database on polyphenol, with information on small molecules content in foods (formula, mass, sdf files).	[124,122]
PlantCyc/Plant Metabolic Network (PMN) ([53], https://plantcyc.org), free	Online database with information about small molecules relevant to plant biology (formula, mass, mol files), including pathway databases.	[80,124,5]
PubChem ([75], https://pubchem.ncbi.nlm.nih.gov), free	Chemical information about 111 M+ compounds, 2D and 3D molecular structure, and available spectral information.	[95,110,146];
SciFinder® (https://www.cas.org/solutions/cas-scifinder-discovery-platform/cas-scifinder), proprietary	Curated chemical database.	[167]

*The website was unavailable/offline by the time this review was written.
* *Licensing information was not found.
* **Discontinued.

cyanidin-3-glucoside chloride standard was used for the linear calibration curve determination for anthocyanins quantification [128]. Shen et al. [129] assessed the anthocyanins present in *Fragaria nilgerrensis* and *F. pentaphylla*, applying the MassBank, KNAPSAcK, and HMDB databases and an MRM quantification approach. They annotated a list of 28 anthocyanins, including cyanidin, cyanidin 3-rutinoside, pelargonidin 3-glucoside, pelargonidin 3-glucoside, pelargonidin 3,5-di-*O*-glucoside, pelargonidin 3-*O*-malonylhexoside, pelargonidin *O*-acetylhexoside, which were differentially expressed in the two species, indicating chemotypes in terms of anthocyanins expression [129].

Regarding the quantification of annotated anthocyanins in fruits from LC-MS-derived data, the MRM method was extensively applied [97,134,172,35,41]. Most studies hereby surveyed applied a QqQ and calibration curves obtained from chemical standards to implement this method. Shi et al. [131] introduced standard solutions of 93 anthocyanins to the MRM analysis in several different concentrations as the basis to obtain analytes *m/z*, intensity, and quantitative signals to be monitored from samples of mango (*Mangifera indica*). The authors successfully annotated cyanidin-3-*O*-galactoside (anthocyanin), procyanidin B1 (proanthocyanidin), procyanidin B3 (proanthocyanidin), and procyanidin C1 (proanthocyanidin). Calibration curves from the chemical standards were considered for the calculations, and QqQ was used in their experiment [131]. These are examples of how the annotation workflows described so far are feasible strategies for the analysis of anthocyanins in fruits interpreted from LC-MS data.

4.5. Statistical and chemometric analyses of anthocyanins LC-MS metabolomics data

One straightforward designation of chemometrics is "the chemical discipline that uses mathematical, statistical, and other methods employing formal logic, to design or select optimal measurement procedures and experiments, and to provide maximum relevant chemical information by analyzing chemical data." [33,108]. Practically, within the scope of this review, chemometrics has been applied within studies to highlight significant differential expression (signal intensities) of anthocyanins annotated from LC-MS data acquired from samples or sample groups derived from fruits. Most chemometric strategies applied in such studies were principal component analysis (PCA); projection to latent structures (PLS); orthogonal projection to latent structures (OPLS); the discriminant analyses (DA) referent to PLS and OPLS; hierarchical cluster analysis (HCA); K-means clustering; Random Forest algorithm and variable importance on the projection (VIP). Other main statistical methods used to assign and/or denote significant differences between sample groups or metabolite intensities were one-way analysis of variance (ANOVA); Tukey's test; Duncan's test; Dunnett's test; Student's t-test; fold change and Pearson correlations [18,153,154,158,165,170,175].

PCA was the most widely used chemometric approach to identify the variation of signal intensities of anthocyanins within fruits acquired through LC-MS analysis. As a non-supervised technique, no premises or presuppositions were previously imposed on the method in terms of samples or group distinction. To sum up, PCA returns three matrices, one refers to the scores (usually in reference to the samples, with dimensions $n \times p$), the second refers to the loadings (variables in reference to metabolites, with dimensions $m \times p$) and the third refers to the residual variance (with dimensions $n \times m$). Scores and loadings analyzed together can provide useful insights, given the clusters formed by scores and the way the loadings variables fall within a PCA scatter plot. Both axes of the PCA scatter plots describe the variation explained by each of the principal components plotted [108]. Samples grouped in a cluster suggest similarities of their components, while variables clustered suggest correlations between them, in terms of feature intensities in the case of LC-MS derived data [108].

The study of anthocyanins found in Cabernet Sauvignon grape berries based on regulated deficit irrigation utilized PCA to assess the group-specific accumulation of the significantly changed metabolites [165]. The authors show the scores plot in which sample groups, representing the different amounts of irrigation received by the grapes, are distinguished from one another. Based on that, further analysis shows by means of Venn diagram, Volcano plots (based on log2 fold change), and paired clusters heatmaps that such distinctions were entailed by the differential presence of specific anthocyanins and anthocyanin-like molecules. The comparison of the groups CK (watered with 100 % evapotranspiration) and RDI2 (watered with 50 % evapotranspiration) showed that the anthocyanin keracyanin (a synonym for cyanidin 3-*O*-rutinoside) had a higher concentration in CK sample groups, while malvidin 3-*O*-galactoside, petunidin 3-*O*-glucoside, and other anthocyanins had higher concentrations in RDI2 sample groups, thereby characterizing the differences between these two groups. The same chemometric analysis workflow was performed between other paired groups [165]. Moreover, through the integration of PCA, HCA, fold changes, and VIP, Fei et al. [38] were able to distinguish most flavonoids differentially accumulated in red and green prickly ash fruits (*Zanthoxylum bungeanum* and *Z. armatum*). PCA clearly explained the variation of 45.4 % (PC1) between samples of both species in relation to the detected metabolites. The authors reported a significantly higher accumulation of anthocyanins in red prickly ash fruits, attributing its red color in the mature stage to the presence of anthocyanins of such color tone. Anthocyanins with higher content were cyanidin-3-*O*-galactoside and cyanidin-3-*O*-glucoside [38].

In the analysis of kiwifruit/berry, a PCA (an unsupervised method, as previously mentioned) was applied to explore the variation of the flavonoid profiles among cultivars and to assess the integrity and reproducibility of the biological samples regarding quality control samples. To recognize which flavonoid metabolites denote the metabolic variation between cultivars, a partial least squares discriminant analysis (PLS-DA) was applied, following the PCA [170]. PLS-DA is a supervised method applied in chemometrics, largely implemented in the

field of metabolomics, to perform the selection of discriminative variables among groups. Through PLS-DA, data groups are a priori established and fed into the algorithm to statistically optimize their separation [50]. Thus, Yu et al. [170] verified through PLS-DA the high predictability and strong goodness of fit of the generated models of the tested groups comparisons. The supervised analysis, hence, provided a robust statistical basis for the characterization of important discriminative features. Such characterization was held through fold changes and VIP analysis. They were able to discern 39 differential metabolites, from which 6 were anthocyanins responsible for the metabolic variation between cultivars. The anthocyanins cyanidin 3-*O*-glucoside, cyanidin 3-*O*-rutinoside, and pelargonidin 3-*O*-beta-*D*-glucoside were annotated and distinguishably accumulated (higher fold change values) in one of the cultivars (red-fleshed kiwi). The result led to the confirmation of the predominance of cyanidins in kiwifruit with red colors [170]. Wu et al. [158], along with PCA and PLS-DA, applied OPLS-DA to assess differentially expressed metabolites. OPLS-DA is a modification of PLS-DA and, for its "orthogonal" character, it can provide improved interpretation of the distinguishably expressed features among groups [108]. Subsequently, the authors applied ANOVA and Duncan's test to appraise significant differences in the accumulation of the anthocyanin cyanidin-3-sambubioside in *Rubus* fruits (raspberries cultivars) of different colors [158].

In summary, to perform robust chemometric analysis of anthocyanins in fruits from LC-MS derived data, most studies adopted comprehensive workflows that pervaded from unsupervised (PCA) to supervised (PLS-DA, OPLS-DA) approaches, together with other statistical methods (fold changes, VIP, ANOVA, Duncan's test and others). An exploratory perspective is given by unsupervised methods, applying the goodness of fit and predictability that can be provided by supervised methods and, at last, the characterization of the observed discriminations by means of other statistical approaches.

5. Conclusions

Anthocyanins are very relevant molecules since they are related to the sensory qualities of food, such as color, flavor, and aroma. Therefore, they have a great impact on the worldwide economy. In addition, anthocyanins have great therapeutic potential and industrial applications. Applications range from wine production and adulteration analysis, natural dyes production, and even in the cultivation of ornamental plants, in which the type of anthocyanin can define the color of the plant, influencing the final price and acceptance of the products.

However, a thorough compilation of the available knowledge on LC-MS-based metabolomics tools used to evaluate anthocyanins in fruits had not yet been gathered. Little has been revealed about the recent methodology used to analyze the content of anthocyanins in various sources. Metabolomics tools such as UPLC-MS/MS followed by chemometric and bioinformatic analyses, preceded by careful and efficient extraction, are powerful implements by which the variability of anthocyanin may be investigated. Rapid developments of cutting-edge technologies both regarding instrumentation for acquisition and interpretation of LC-MS-based data offer suitable and feasible workflows for the characterization of anthocyanins. This advance can be applied to further understanding natural products carrying anthocyanins, food composition, and agricultural residues. As it has been shown in this review article, literature within the scope of anthocyanins present in fruits is remarkably rich. Most compelling contributions encompass the metabolomics-based studies of fruit color expression related to the presence and accumulation of anthocyanins, in which abundant experimental and results references are given. From sample collection, through LC-MS instrumentation, to chemometrics and bioinformatics, a comprehensive overview of the theme was offered in this article. The authors hope that the assembly and attentive description of these studies will contribute to and facilitate upcoming research in this field.

CRediT authorship contribution statement

Joana Paula da Silva Oliveira: Investigation, Formal analysis, Writing – review & editing,Writing – original draft, Visualization. **Renatha Tavares de Oliveira:** Investigation, Writing – review & editing. **Amanda Lima Guedes:** Investigation, Writing – review & editing. **Matheus da Costa Oliveira:** Investigation, Visualization. **Andrea Furtado Macedo:** Conceptualization, Methodology, Investigation, Supervision, Writing – review & editing, Resources, Project administration, Funding acquisition. All authors critically revised the manuscript and gave their final approval.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the Coordination of Superior Level Staff Improvement (CAPES) (J.P.S.O is CAPES 001 scholar at The Graduate Program in Biological Sciences (Neotropical Biodiversity) - PPGBIO, no. 88887.685735/2022–00), (R.T.O. is CAPES scholar at the Graduate Program in Biological Sciences (Neotropical Biodiversity) PPGBIO, no. 88887.685726/2022-00); Foundation for Research of Rio de Janeiro State (FAPERJ) (M.C.O. - no. e-26/204.104/2021 - scholarship no. 2019.01223.0) and Federal University of the State of Rio de Janeiro (UNIRIO) (A.L.G - scholarship no. 20162110054), which granted fellowships and financial support (Edital INOVA-UNIRIO no. 01/2021; Edital PPQ-UNIRIO no. 03/2022).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.cpb.2022.100260.

References

[1] El-Sayed M. Abdel-Aal, Humayoun Akhtar, Iwona Rabalski, Michael Bryan, Accelerated, microwave-assisted, and conventional solvent extraction methods affect anthocyanin composition from colored grains, J. Food Sci. 79 (2) (2014) C138–C146, https://doi.org/10.1111/1750-3841.12346.
[2] Maria Assunta Acquavia, et al., Analytical methods for extraction and identification of primary and secondary metabolites of apple (Malus domestica) fruits: a review, Separations 8.7 (2021) 91, https://doi.org/10.3390/separations8070091.
[3] Farit Mochamad Afendi, Taketo Okada, Mami Yamazaki, Aki Hirai-Morita, Yukiko Nakamura, Kensuke Nakamura, Shun Ikeda, et al., KNApSAcK family databases: integrated metabolite–plant species databases for multifaceted plant research, Plant Cell Physiol. 53 (2) (2012), https://doi.org/10.1093/pcp/pcr165.
[4] Alfredo Aires, Phenolics in foods: Extraction, analysis and measurements, Phenolic Compd. (2017) 61–88, https://doi.org/10.5772/66889.
[5] J.William Allwood, Tomasz L. Woznicki, Yun Xu, Alexandre Foito, Kjersti Aaby, Julie Sungurtas, Sabine Freitag, et al., Application of HPLC–PDA–MS metabolite profiling to investigate the effect of growth temperature and day length on blackcurrant fruit, Metabolomics 15 (1) (2019) 1–17, https://doi.org/10.1007/s11306-018-1462-5.
[6] Gerardo Alvarez-Rivera, Diego Ballesteros-Vivas, Fabian Parada-Alfonso, Elena Ibañez, Alejandro Cifuentes, Recent applications of high resolution mass spectrometry for the characterization of plant natural products, TrAC Trends Anal. Chem. 112 (2019) 87–101, https://doi.org/10.1016/j.trac.2019.01.002.
[7] Carmen Arlotta, Giuseppe D. Puglia, Claudia Genovese, Valeria Toscano, Rumyana Karlova, Jules Beekwilder, Ric C.H. De Vos, Salvatore A. Raccuia, MYB5-like and bHLH influence flavonoid composition in pomegranate, Plant Sci. 298 (2020), 110563, https://doi.org/10.1016/j.plantsci.2020.110563.
[8] Cemil Aydoğan, Recent advances and applications in LC-HRMS for food and plant natural products: a critical review, Anal. Bioanal. Chem. 412 (9) (2020) 1973–1991, https://doi.org/10.1007/s00216-019-02328-6.

[9] Edward E.K. Baidoo, Veronica Teixeira Benites, Mass spectrometry-based microbial metabolomics: Techniques, analysis, and applications. Microbial metabolomics, Humana Press,, New York, NY, 2019, pp. 11–69, https://doi.org/10.1007/978-1-4939-8757-3_2.
[10] Jeremy S. Barnes, Hien P. Nguyen, Sijia Shen, Kevin A. Schug, General method for extraction of blueberry anthocyanins and identification using high performance liquid chromatography–electrospray ionization-ion trap-time of flight-mass spectrometry, J. Chromatogr. A 1216 (23) (2009) 4728–4735, https://doi.org/10.1016/j.chroma.2009.04.032.
[11] Marina Cvjetko Bubalo, Natka Ćurko, Marina Tomašević, Karin Kovačević Ganić, Ivana Radojčić Redovniković, Green extraction of grape skin phenolics by using deep eutectic solvents, Food Chemistry 200 (2016) 159–166, https://doi.org/10.1016/j.foodchem.2016.01.040.
[12] Bogusław Buszewski, Sylwia Noga, Hydrophilic interaction liquid chromatography (HILIC)—a powerful separation technique, Anal. Bioanal. Chem. 402 (1) (2012) 231–247, https://doi.org/10.1007/S00216-011-5308-5.
[13] Carla Lenore F. Calumpang, Tomoki Saigo, Mutsumi Watanabe, Takayuki Tohge, Cross-species comparison of fruit-metabolomics to elucidate metabolic regulation of fruit polyphenolics among solanaceous crops, Metabolites 10 (5) (2020) 209, https://doi.org/10.3390/metabo10050209.
[14] Lei Cao, Yena Park, Sanggil Lee, Dae-Ok Kim, Extraction, identification, and health benefits of anthocyanins in blackcurrants (Ribes nigrum L.), Appl. Sci. 11 (4) (2021) 1863, https://doi.org/10.3390/app11041863.
[15] Francesca Cappellini, Alessandra Marinelli, Marta Toccaceli, Chiara Tonelli, Katia Petroni, Anthocyanins: from mechanisms of regulation in plants to health benefits in foods, Front. Plant Sci. 12 (2021), https://doi.org/10.3389/fpls.2021.748049.
[16] Anna Laura Capriotti, Chiara Cavaliere, Patrizia Foglia, Susy Piovesana, Salvatore Ventura, Chromatographic methods coupled to mass spectrometry detection for the determination of phenolic acids in plants and fruits, J. Liq. Chromatogr. Relat. Technol. 38 (3) (2015) 353–370, https://doi.org/10.1080/10826076.2014.941263.
[17] Anna Laura Capriotti, Chiara Cavaliere, Patrizia Foglia, Roberto Samperi, Serena Stampachiacchiere, Salvatore Ventura, Aldo Laganà, Recent advances and developments in matrix solid-phase dispersion, TrAC Trends in Analytical Chemistry 71 (2015) 186–193, https://doi.org/10.1016/j.trac.2015.03.012.
[18] Elisabete Carvalho, Pietro Franceschi, Antje Feller, Luisa Palmieri, Ron Wehrens, Stefan Martens, A targeted metabolomics approach to understand differences in flavonoid biosynthesis in red and yellow raspberries, Plant Physiol. Biochem. 72 (2013) 79–86, https://doi.org/10.1016/j.plaphy.2013.04.001.
[19] Roberto Castro-Muñoz, Elsa Díaz-Montes, Emilia Gontarek-Castro, Grzegorz Boczkaj, Charis M. Galanakis, A comprehensive review on current and emerging technologies toward the valorization of bio-based wastes and by products from foods", Compr. Rev. Food Sci. Food Saf. 21 (1) (2022) 46–105, https://doi.org/10.1111/1541-4337.12894.
[20] Úrsula Catalán, Laura Barrubés, Rosa Maria Valls, Rosa Solà, Laura Rubió, In vitro metabolomic approaches to investigating the potential biological effects of phenolic compounds: An update, Genom., Proteom. Bioinforma. 15 (4) (2017) 236–245, https://doi.org/10.1016/j.gpb.2016.12.007.
[21] Liyuan Chen, Min Yang, Haijin Mou, Qing Kong, Ultrasound-assisted extraction and characterization of anthocyanins from purple corn bran, J. Food Process. Preserv. 42 (1) (2018), e13377, https://doi.org/10.1111/jfpp.13377.
[22] Yan Cheng, Fumin Xue, Shuai Yu, Shichao Du, Yu Yang, Subcritical water extraction of natural products, Molecules 26 (13) (2021) 4004, https://doi.org/10.3390/molecules26134004.
[23] V.éronique Cheynier, Phenolic compounds: from plants to foods, Phytochem. Rev. 11 (2) (2012) 153–177, https://doi.org/10.1007/s11101-012-9242-8.
[24] H. Çoklar, M. Akbulut, Effect of sun, oven and freeze-drying on anthocyanins, phenolic compounds and antioxidant activity of black grape (Ekşikara)(Vitis vinifera L.), South Afr. J. Enol. Vitic. 38 (2) (2017) 264–272, https://doi.org/10.21548/38-2-2127.
[25] Mauro Commisso, Andrea Anesi, Silvia Dal Santo, Flavia Guzzo, Performance comparison of electrospray ionization and atmospheric pressure chemical ionization in untargeted and targeted liquid chromatography/mass spectrometry based metabolomics analysis of grapeberry metabolites, Rapid Commun. Mass Spectrom. 31 (3) (2017) 292–300, https://doi.org/10.1002/rcm.7789.
[26] Artur Conde, Diana Pimentel, Andreia Neves, Lia-T.ânia Dinis, Sara Bernardo, Carlos M. Correia, Hernâni Gerós, José Moutinho-Pereira, Kaolin foliar application has a stimulatory effect on phenylpropanoid and flavonoid pathways in grape berries, Front. Plant Sci. 7 (2016) 1150, https://doi.org/10.3389/fpls.2016.01150.
[27] C.ândice L.K. Copetti, Fernando Diefenthaeler, Fernanda Hansen, Francilene G. K. Vieira, Patricia F. Di Pietro, Fruit-Derived Anthocyanins: Effects on Cycling-Induced Responses and Cycling Performance, Antioxidants 11 (2) (2022) 387, https://doi.org/10.3390/antiox11020387.
[28] Montserrat Cortina-Puig, Hector Gallart-Ayala, S.ílvia Lacorte, Liquid chromatography coupled to electrochemical detection and mass spectrometry for the determination of phenolic compounds in food and beverages, Curr. Anal. Chem. 8 (4) (2012) 436–455, https://doi.org/10.2174/157341112803216681.
[29] Marina Creydt, Markus Fischer, Plant Metabolomics: Maximizing Metabolome Coverage by Optimizing Mobile Phase Additives for Nontargeted Mass Spectrometry in Positive and Negative Electrospray Ionization Mode, Anal. Chem. 89 (19) (2017) 10474–10486, https://doi.org/10.1021/acs.analchem.7b02592.
[30] Asfaw Degu, Biruk Ayenew, Grant R. Cramer, Aaron Fait, Polyphenolic responses of grapevine berries to light, temperature, oxidative stress, abscisic acid and jasmonic acid show specific developmental-dependent degrees of metabolic resilience to perturbation, Food Chem. 212 (2016) 828–836, https://doi.org/10.1016/j.foodchem.2016.05.164.
[31] Asfaw Degu, Uri Hochberg, Noga Sikron, Luca Venturini, Genny Buson, Ryan Ghan, Inbar Plaschkes, et al., Metabolite and transcript profiling of berry skin during fruit development elucidates differential regulation between Cabernet Sauvignon and Shiraz cultivars at branching points in the polyphenol pathway, BMC Plant Biol. 14 (1) (2014) 1–20, https://doi.org/10.1186/s12870-014-0188-4.
[32] V.I. Deineka, L.A. Deineka, I.I. Saenko, Regularities of Anthocyanins Retention in RP HPLC for "Water–Acetonitrile–Phosphoric Acid" Mobile Phases, J. Anal. Methods Chem. (2015)) (2015), https://doi.org/10.1155/2015/732918.
[33] S.N. Deming, Y. Michotte, Desiré Luc Massart, L. Kaufman, B.G.M. Vandeginste, Chemometrics: A textbook, Elsevier,, 1988.
[34] Robert A. van den Berg, Huub C.J. Hoefsloot, Johan A. Westerhuis, Age K. Smilde, Mariët J. van der Werf, Centering, scaling, and transformations: improving the biological information content of metabolomics data, BMC Genom. 7 (1) (2006) 1–15, https://doi.org/10.1186/1471-2164-7-142.
[35] Wenkai Duan, Wanlu Shao, Weida Lin, Ling Yuan, Qiuwei Lu, Luxi Chen, Lyuben Zagorchev, Junmin Li, Integrated metabolomics and transcriptomics reveal the differences in fruit quality of the red and white Fragaria pentaphylla morphs, Food Biosci. 40 (2021), 100896, https://doi.org/10.1016/j.fbio.2021.100896.
[36] Anne M. Evans, Claire O'Donovan, Mary Playdon, Chris Beecher, Richard D. Beger, John A. Bowden, David Broadhurst, et al., Dissemination and analysis of the quality assurance (QA) and quality control (QC) practices of LC–MS based untargeted metabolomics practitioners, Metabolomics 16 (10) (2020) 1–16, https://doi.org/10.1007/s11306-020-01728-5.
[37] Saqib Farooq, Manzoor Ahmad Shah, Mohammed Wasim Siddiqui, B.N. Dar, Shabir Ahmad Mir, Asgar Ali, Recent trends in extraction techniques of anthocyanins from plant materials, J. Food Meas. Charact. 14 (6) (2020) 3508–3519, https://doi.org/10.1007/s11694-020-00598-8.
[38] Xitong Fei, Yuan Wei, Yichen Qi, Yingli Luo, Haichao Hu, Anzhi Wei, Integrated LC-MS/MS and Transcriptome Sequencing Analysis Reveals the Mechanism of Color Formation During Prickly Ash Fruit Ripening, Front. Nutr. 9 (2022), https://doi.org/10.3389/fnut.2022.847823.
[39] Fernanda Fernandez-Aulis, Liliana Hernandez-Vazquez, Guillermo Aguilar-Osorio, Daniel Arrieta-Baez, Arturo Navarro-Ocana, Extraction and identification of anthocyanins in corn cob and corn husk from Cacahuacintle maize, J. Food Sci. 84 (5) (2019) 954–962, https://doi.org/10.1111/1750-3841.14589.
[40] Maite Gagneten, Graciela Leiva, Daniela Salvatori, Carolina Schebor, Nahuel Olaiz, Optimization of Pulsed Electric Field Treatment for the Extraction of Bioactive Compounds from Blackcurrant, 2019, Food Bioprocess Technol. 12 (2019) 1102–1109, https://doi.org/10.1007/s11947-019-02283-1.
[41] Sichen Gan, Gang Zheng, Shoukuo Zhu, Jieyu Qian, Lijun Liang, Integrative Analysis of Metabolome and Transcriptome Reveals the Mechanism of Color Formation in Liriope spicata Fruit, Metabolites 12 (2) (2022) 144, https://doi.org/10.3390/metabo12020144.
[42] Maria del Pilar Garcia-Mendoza, Faber Ariel Espinosa-Pardo, Andressa Mara Baseggio, Gerardo Fernández Barbero, M.ário Roberto Maróstica Junior, Mauricio Ariel Rostagno, Julian Martínez, Extraction of phenolic compounds and anthocyanins from juçara (*Euterpe edulis* Mart.) residues using pressurized liquids and supercritical fluids, J. Supercrit. Fluids 119 (2017) 9–16, https://doi.org/10.1016/j.supflu.2016.08.014.
[43] Michael Gerlich, Steffen Neumann, MetFusion: integration of compound identification strategies.", J. Mass Spectrom. 48 (3) (2013) 291–298, https://doi.org/10.1002/jms.3123.
[44] Ryan Ghan, Steven C. Van Sluyter, Uri Hochberg, Asfaw Degu, Daniel W. Hopper, Richard L. Tillet, Karen A. Schlauch, Paul A. Haynes, Aaron Fait, Grant R. Cramer, Five omic technologies are concordant in differentiating the biochemical characteristics of the berries of five grapevine (*Vitis vinifera* L.) cultivars, BMC Genom. 16 (1) (2015) 1–26, https://doi.org/10.1186/s12864-015-2115-y.
[45] Alireza Ghassempour, Rouhollah Heydari, Zahra Talebpour, Ali Reza Fakhari, Ali Rassouli, Noel Davies, Hassan Y. Aboul-Enein, Study of new extraction methods for separation of anthocyanins from red grape skins: analysis by HPLC and LC-MS/MS, J. Liq. Chromatogr. Relat. Technologies® 31 (17) (2008) 2686–2703, https://doi.org/10.1080/10826070802353247.
[46] Ana C. Gonçalves, Ana R. Nunes, Amílcar Falcão, Gilberto Alves, Luís R. Silva, Dietary effects of anthocyanins in human health: A comprehensive review, Pharmaceuticals 14 (7) (2021) 690, https://doi.org/10.3390/ph14070690.
[47] María José Aliaño González, Ceferino Carrera, Gerardo F. Barbero, Miguel Palma, A comparison study between ultrasound–assisted and enzyme–assisted extraction of anthocyanins from blackcurrant (Ribes nigrum L.), Food Chem.: X 13 (2022), 100192, https://doi.org/10.1016/j.fochx.2021.100192.
[48] Eva Gorrochategui, Joaquim Jaumot, S.ílvia Lacorte, Romà Tauler, Data analysis strategies for targeted and untargeted LC-MS metabolomic studies: Overview and workflow, TrAC Trends Anal. Chem. 82 (2016) 425–442, https://doi.org/10.1016/j.trac.2016.07.004.
[49] G.A. Gowda, Danijel Djukovic, Overview of mass spectrometry-based metabolomics: opportunities and challenges, Mass Spectrom. Metab. (2014) 3–12, https://doi.org/10.1007/978-1-4939-1258-2.
[50] Piotr S. Gromski, Howbeer Muhamadali, David I. Ellis, Yun Xu, Elon Correa, Michael L. Turner, Royston Goodacre, A tutorial review: Metabolomics and partial least squares-discriminant analysis–a marriage of convenience or a shotgun wedding, Anal. Chim. Acta 879 (2015) 10–23, https://doi.org/10.1016/j.aca.2015.02.012.

[51] Carlos Guijas, J.Rafael Montenegro-Burke, Xavier Domingo-Almenara, Amelia Palermo, Benedikt Warth, Gerrit Hermann, Gunda Koellensperger, et al., METLIN: a technology platform for identifying knowns and unknowns, Anal. Chem. 90 (5) (2018) 3156–3164, https://doi.org/10.1021/acs.analchem.7b04424.

[52] Rachel Hair, Junichi R. Sakaki, Ock K. Chun, Anthocyanins, microbiome and health benefits in aging, Molecules 26 (3) (2021) 537, https://doi.org/10.3390/molecules26030537.

[53] Charles Hawkins, Daniel Ginzburg, Kangmei Zhao, William Dwyer, Bo Xue, Angela Xu, Selena Rice, et al., Plant Metabolic Network 15: A resource of genome-wide metabolism databases for 126 plants and algae, J. Integr. Plant Biol. 63 (11) (2021) 1888–1905, https://doi.org/10.1111/jipb.13163.

[54] Hisayuki Horai, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, et al., MassBank: a public repository for sharing mass spectral data for life sciences, J. Mass Spectrom. 45 (7) (2010) 703–714, https://doi.org/10.1002/jms.1777.

[55] Yoko Iijima, Yukiko Nakamura, Yoshiyuki Ogata, Ken'ichi Tanaka, Nozomu Sakurai, Kunihiro Suda, Tatsuya Suzuki, et al., Metabolite annotations based on the integration of mass spectral information, Plant J. 54 (5) (2008) 949–962, https://doi.org/10.1111/j.1365-313X.2008.03434.x.

[56] Laura Jaakola, New insights into the regulation of anthocyanin biosynthesis in fruits, Trends Plant Sci. 18 (9) (2013) 477–483, https://doi.org/10.1016/j.tplants.2013.06.003.

[57] Dorota A. Jarret, Jenny Morris, Danny W. Cullen, Sandra L. Gordon, Susan R. Verrall, Linda Milne, Pete E. Hedley, J.William Allwood, Rex M. Brennan, Robert D. Hancock, A transcript and metabolite atlas of blackcurrant fruit development highlights hormonal regulation and reveals the role of key transcription factors, Front. Plant Sci. (2018) 1235, https://doi.org/10.3389/fpls.2018.01235.

[58] K.M. Jeong, J. Zhao, Y. Jin, et al., Highly efficient extraction of anthocyanins from grape skin using deep eutectic solvents as green and tunable media, Archives of Pharmacal Research 38 (2015) 2143–2152, https://doi.org/10.1007/s12272-015-0678-4.

[59] Xiaolong Ji, Jing Ren, Shaoyu Lang, Di Wang, Liwei Zhu, Xingshun Song, Differential Regulation of Anthocyanins in Cerasus humilis Fruit Color Revealed by Combined Transcriptome and Metabolome Analysis, Forests 11 (10) (2020) 1065, https://doi.org/10.3390/f11101065.

[60] Xiaolong Ji, Jing Ren, Yixin Zhang, Shaoyu Lang, Di Wang, Xingshun Song, Integrated Analysis of the Metabolome and Transcriptome on Anthocyanin Biosynthesis in Four Developmental Stages of Cerasus humilis Peel Coloration, Int. J. Mol. Sci. 22 (21) (2021) 11880, https://doi.org/10.3390/ijms222111880.

[61] Nonghui Jiang, Huili Zhu, Wei Liu, Chao Fan, Feng Jin, Xu Xiang, Metabolite differences of polyphenols in different Litchi Cultivars (Litchi chinensis Sonn.) based on extensive targeted metabonomics, Molecules 26 (4) (2021) 1181, https://doi.org/10.3390/molecules26041181.

[62] Yang Jiao, Donglin Chen, Mingtao Fan, Siew Young Quek, UPLC-QqQ-MS/MS-based phenolic quantification and antioxidant activity assessment for thinned young kiwifruits, Food Chem. 281 (2019) 97–105, https://doi.org/10.1016/j.foodchem.2018.12.062.

[63] Young Ho Jin, Hyun Ju Shin, Hye Jung Ahn, Bong Keun Jang, Na Hyun Park, Jongki Hong, Identification of anthocyanins in berry species using LC-ESI-MS combined with high-speed counter-current chromatography, Bull. Korean Chem. Soc. 37 (10) (2016) 1618–1624, https://doi.org/10.1002/bkcs.10915.

[64] Johanna Jokioja, Baoru Yang, Kaisa M. Linderborg, Acylated anthocyanins: A review on their bioavailability and effects on postprandial carbohydrate metabolism and inflammation, Compr. Rev. Food Sci. Food Saf. 20 (6) (2021) 5570–5615, https://doi.org/10.1111/1541-4337.12836.

[65] Christophe Junot, Benoit Colsch, High resolution mass spectrometry based techniques at the crossroads of metabolic pathways, Mass Spectrom. Rev. 33 (2014) 471–500, https://doi.org/10.1002/mas.21401.

[66] Piotr Kachlicki, Anna Piasecka, Maciej Stobiecki, Łukasz Marczak, Structural characterization of flavonoid glycoconjugates and their derivatives with mass spectrometric techniques, Molecules 21 (11) (2016) 1494, https://doi.org/10.3390/molecules21111494.

[67] Kathithileni M. Kalili, Andre de Villiers, Recent developments in the HPLC separation of phenolic compounds, J. Sep. Sci. 34 (8) (2011) 854–876, https://doi.org/10.1002/jssc.201000811.

[68] Minoru Kanehisa, KEGG bioinformatics resource for plant genomics and metabolomics. Plant bioinformatics, Humana Press,, New York, NY, 2016, pp. 55–70, https://doi.org/10.1007/978-1-4939-3167-5_3.

[69] Hye-Ji Kang, Min-Jung Ko, Myong-Soo Chung, Anthocyanin structure and pH dependent extraction characteristics from blueberries (*Vaccinium corymbosum*) and chokeberries (*Aronia melanocarpa*) in Subcritical Water State, Foods 10 (3) (2021) 527, https://doi.org/10.3390/foods10030527.

[70] Ibrahim Karaman, Preprocessing and pretreatment of metabolomics data for statistical analysis, Metab.: Fundam. Clin. Appl. (2017) 145–161, https://doi.org/10.1007/978-3-319-47656-8_6.

[71] Kooi-Yeong Khaw, Marie-Odile Parat, Paul Nicholas Shaw, James Robert Falconer, Solvent supercritical fluid technologies to extract bioactive compounds from natural sources: A review, Molecules 22 (7) (2017) 1186, https://doi.org/10.3390/molecules22071186.

[72] Emil Khusnutdinov, Anna Sukhareva, Maria Panfilova, Elena Mikhaylova, Anthocyanin biosynthesis genes as model genes for genome editing in plants, Int. J. Mol. Sci. 22 (16) (2021) 8752, https://doi.org/10.3390/ijms22168752.

[73] Hye Kyong Kim, Rob Verpoorte, Sample preparation for plant metabolomics, Phytochem. Anal.: Int. J. Plant Chem. Biochem. Tech. 21 (1) (2010) 4–13, https://doi.org/10.1002/pca.1188.

[74] Mina Kim, Dong-Geon Nam, Jeong-Sook Choe, Kyung-A. Hwang, Ae-Jin Choi, Optimization of pectinase-assisted extraction condition of mulberry (Morus alba L.) fruit using response surface methodology and its effect on anthocyanin synthesis pathway-related metabolites, J. Food Sci. 86 (9) (2021) 3926–3938, https://doi.org/10.1111/1750-3841.15853.

[75] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, et al., PubChem in 2021: new data content and improved web interfaces, Nucleic Acids Res. 49 (D1) (2021) D1388–D1395, https://doi.org/10.1093/nar/gkaa971.

[76] Tobias Kind, Hiroshi Tsugawa, Tomas Cajka, Yan Ma, Zijuan Lai, Sajjan S. Mehta, Gert Wohlgemuth, et al., Identification of small molecules using accurate mass MS/MS search, Mass Spectrom. Rev. 37 (4) (2018) 513–532, https://doi.org/10.1002/mas.21535.

[77] Kodsara Ramachandra Kiran, Puthanvila Surendrababu Swathy, Bobby Paul, Kabekkodu Shama Prasada, Mattu Radhakrishna Rao, Manjunath B. Joshi, Padmalatha S. Rai, Kapaettu Satyamoorthy, Annamalai Muthusamy, Untargeted metabolomics and DNA barcoding for discrimination of Phyllanthus species, J. Ethnopharmacol. 273 (2021), 113928, https://doi.org/10.1016/j.jep.2021.113928.

[78] Neil R. Kitteringham, Rosalind E. Jenkins, Catherine S. Lane, Victoria L. Elliott, B. Kevin Park, Multiple reaction monitoring for quantitative biomarker analysis in proteomics and metabolomics, J. Chromatogr. B 877 (13) (2009) 1229–1239, https://doi.org/10.1016/j.jchromb.2008.11.013.

[79] Tingting Li, Yunfang Fan, Huan Qin, Guoli Dai, Guoxiu Li, Yanlong Li, Jingjin Wang, et al., Transcriptome and flavonoids metabolomic analysis identifies regulatory networks and hub genes in black and white fruits of Lycium ruthenicum Murray, Front. Plant Sci. (2020) 1256, https://doi.org/10.3389/fpls.2020.01256.

[80] Yukuo Li, Jinbao Fang, Xiujuan Qi, Miaomiao Lin, Yunpeng Zhong, Leiming Sun, Wen Cui, Combined analysis of the fruit metabolome and transcriptome reveals candidate genes involved in flavonoid biosynthesis in Actinidia arguta, Int. J. Mol. Sci. 19 (5) (2018) 1471, https://doi.org/10.3390/ijms19051471.

[81] Yuhua Liu, Junheng Lv, Zhoubin Liu, Jing Wang, Bozhi Yang, Wenchao Chen, Lijun Ou, Xiongze Dai, Zhuqing Zhang, Xuexiao Zou, Integrative analysis of metabolome and transcriptome reveals the mechanism of color formation in pepper fruit (Capsicum annuum L.), Food Chem. 306 (2020), 125629, https://doi.org/10.1016/j.foodchem.2019.125629.

[82] Ante Lončarić, Maria Celeiro, Antun Jozinović, Josip Jelinić, Tihomir Kovač, Stela Jokić, Jurislav Babić, Tihomir Moslavac, Sandra Zavadlav, Marta Lores, Green extraction methods for extraction of polyphenolic compounds from blueberry pomace, Foods 9 (11) (2020) 1521, https://doi.org/10.3390/foods9111521.

[83] Cecilia Jiménez López, Cristina Caleja, M.A. Prieto, Maria Filomena Barreiro, Lillian Barros, Isabel, C.F.R. Ferreira, Optimization and Comparison of Heat and Ultrasound Assisted Extraction Techniques to Obtain Anthocyanin Compounds from Arbutus unedo L. Fruits, Food Chem. 264 (2018) 81–91, https://doi.org/10.1016/j.foodchem.2018.04.103.

[84] M.A. López-Bascón, M.D. Luque De Castro, Soxhlet extraction. Liquid-phase extraction, Elsevier,, 2020, pp. 327–354, https://doi.org/10.1016/B978-0-12-816911-7.00011-6.

[85] Olalla López-Fernández, Rubén Domínguez, Mirian Pateiro, Paulo E.S. Munekata, Gabriele Rocchetti, José M. Lorenzo, Determination of polyphenols using liquid chromatography–tandem mass spectrometry technique (LC–MS/MS): A review, Antioxidants 9 (6) (2020) 479, https://doi.org/10.3390/antiox9060479.

[86] Suwen Lu, Jiayang Wang, Yaxian Zhuge, Mengwei Zhang, Chang Liu, Haifeng Jia, Jinggui Fang, Integrative Analyses of Metabolomes and Transcriptomes Provide Insights into Flavonoid Variation in Grape Berries, J. Agric. Food Chem. 69 (41) (2021) 12354–12367, https://doi.org/10.1021/acs.jafc.1c02703.

[87] Agnieszka Łubek-Nguyen, Wojciech Ziemichód, Marta Olech, Application of Enzyme-Assisted Extraction for the Recovery of Natural Bioactive Compounds for Nutraceutical and Pharmaceutical Applications, Appl. Sci. 12 (7) (2022) 3232, https://doi.org/10.3390/app12073232.

[88] Ana Paula Da. Fonseca Machado, José Luis Pasquel-Reátegui, Gerardo Fernández Barbero, Julian Martínez, Pressurized liquid extraction of bioactive compounds from blackberry (Rubus fruticosus L.) residues: a comparison with conventional methods, Food Res. Int. 77 (2015) 675–683, https://doi.org/10.1016/j.foodres.2014.12.042.

[89] Motilva Maria-José, Aida Serra, Alba Macià, Analysis of food polyphenols by ultra high-performance liquid chromatography coupled to mass spectrometry: An overview, J. Chromatogr. A 1292 (2013) 66–82, https://doi.org/10.1016/j.chroma.2013.01.012.

[90] Biswapriya B. Misra, Data normalization strategies in metabolomics: Current challenges, approaches, and tools, Eur. J. Mass Spectrom. 26 (3) (2020) 165–174, https://doi.org/10.1177/1469066720918446.

[91] Sofia Moco, Raoul J. Bino, Oscar Vorst, Harrie A. Verhoeven, Joost de Groot, Teris A. van Beek, Jacques Vervoort, C.H.Ric De Vos, A liquid chromatography-mass spectrometry-based metabolome database for tomato, Plant Physiol. 141 (4) (2006) 1205–1218, https://doi.org/10.1104/pp.106.078428.

[92] Hamdoon A. Mohammed, Riaz A. Khan, Anthocyanins: Traditional Uses, Structural and Functional Variations, Approaches to Increase Yields and Products' Quality, Hepatoprotection, Liver Longevity, and Commercial Products, Int. J. Mol. Sci. 23 (4) (2022) 2149, https://doi.org/10.3390/ijms23042149.

[93] Domenico Montesano, Gabriele Rocchetti, Lina Cossignani, Luigi Lucini, Maria Stella Simonetti, Frances Blasia, Italian Lycium barbarum L. berry: Chemical characterization and nutraceutical value, 1934578X1801300913, Nat. Prod. Commun. 13 (9) (2018), https://doi.org/10.1177/1934578X1801300913.
[94] Mian Yahya Mushtaq, Young Hae Choi, Robert Verpoorte, Erica G. Wilson, Extraction for metabolomics: access to the metabolome, Phytochem. Anal. 25 (4) (2014) 291–306, https://doi.org/10.1002/pca.2505.
[95] Luca Narduzzi, Jan Stanstrup, Fulvio Mattivi, Comparing wild American grapes with Vitis vinifera: a metabolomics study of grape composition, J. Agric. Food Chem. 63 (30) (2015) 6823–6834, https://doi.org/10.1021/acs.jafc.5b01999.
[96] Spurthi N. Nayak, B. Aravind, Sachin S. Malavalli, B.S. Sukanth, R. Poornima, Pushpa Bharati, Kathleen Hefferon, Chittaranjan Kole, Naveen Puppala, Omics Technologies to Enhance Plant Based Functional Foods: An Overview, 742095-742095, Front. Genet. 12 (2021), https://doi.org/10.3389/fgene.2021.742095.
[97] Hui Nie, Huangcan Chen, Guanli Li, Kezhen Su, Mubo Song, Zhenhua Duan, Xiaochun Li, et al., Comparison of flavonoids and phenylpropanoids compounds in Chinese water chestnut processed with different methods, Food Chem. 335 (2021), 127662, https://doi.org/10.1016/j.foodchem.2020.127662.
[98] Louis-F.élix Nothias, Daniel Petras, Robin Schmid, Kai Dührkop, Johannes Rainer, Abinesh Sarvepalli, Ivan Protsyuk, et al., Feature-based molecular networking in the GNPS analysis environment, Nat. Methods 17 (9) (2020) 905–908, https://doi.org/10.1038/s41592-020-0933-6.
[99] Violeta Nour, Franci Stampar, Robert Veberic, Jerneja Jakopic, Anthocyanins profile, total phenolics and antioxidant activity of black currant ethanolic extracts as influenced by genotype and ethanol concentration, Food Chem. 141 (2) (2013) 961–966, https://doi.org/10.1016/j.foodchem.2013.03.105.
[100] Ana N. Nunes, Alexandra Borges, Ana A. Matias, Maria Rosário Bronze, Joana Oliveira, Alternative Extraction and Downstream Purification Processes for Anthocyanins, Molecules 27 (2) (2022) 368, https://doi.org/10.3390/molecules27020368.
[101] Akira Oikawa, Takao Otsuka, Yusuke Jikumaru, Shinjiro Yamaguchi, Fumio Matsuda, Ryo Nakabayashi, Tadashi Takashina, Kanji Isuzugawa, Kazuki Saito, Katsuhiro Shiratake, Effects of freeze-drying of samples on metabolite levels in metabolome analyses, J. Sep. Sci. 34 (24) (2011) 3561–3567, https://doi.org/10.1002/jssc.201100466.
[102] Paulina Ongkowijoyo, Diego A. Luna-Vital, Elvira Gonzalez de Mejia, Extraction techniques and analysis of anthocyanins from food sources by mass spectrometry: An update, Food Chem. 250 (2018) 113–126, https://doi.org/10.1016/j.foodchem.2018.01.055.
[103] Jan Oszmiański, Paulina Nowicka, Mirosława Teleszko, Aneta Wojdyło, Tomasz Cebulak, Krzysztof Oklejewicz, Analysis of phenolic compounds and antioxidant activity in wild blackberry fruits, Int. J. Mol. Sci. 16 (7) (2015) 14540–14553, https://doi.org/10.3390/ijms160714540.
[104] Juliana Paes, Raquel Dotta, Gerardo F. Barbero, Julian Martínez, Extraction of phenolic compounds and anthocyanins from blueberry (*Vaccinium myrtillus* L.) residues using supercritical CO_2 and pressurized liquids, J. Supercrit. Fluids 95 (2014) 8–16, https://doi.org/10.1016/j.supflu.2014.07.025.
[105] Harry E. Pence, Antony Williams, ChemSpider: Online Chem. Inf. Resour. (2010) 1123–1124, https://doi.org/10.1021/ed100697w.
[106] Ana V.González de Peredo, Mercedes Vázquez-Espinosa, Estrella Espada-Bellido, Ana Jiménez-Cantizano, Marta Ferreiro-González, Antonio Amores-Arrocha, Miguel Palma, Carmelo G. Barroso, Gerardo F. Barbero, Development of new analytical microwave-assisted extraction methods for bioactive compounds from myrtle (Myrtus communis L.), Molecules 23 (11) (2018) 2992, https://doi.org/10.3390/molecules23112992.
[107] Miriam Pérez-Cova, Carmen Bedia, Dwight R. Stoll, Romà Tauler, Joaquim Jaumot, MSroi: A pre-processing tool for mass spectrometry-based studies, Chemom. Intell. Lab. Syst. 215 (2021), 104333, https://doi.org/10.1016/j.chemolab.2021.104333.
[108] Rui Climaco Pinto, Chemometrics methods and strategies in metabolomics. Metabolomics: From Fundamentals to Clinical Applications, Springer,, Cham, 2017, pp. 163–190, https://doi.org/10.1007/978-3-319-47656-8_7.
[109] Susy Piovesana, Chiara Cavaliere, Andrea Cerrato, Carmela Maria Montone, Aldo Laganà, Anna Laura Capriotti, Developments and pitfalls in the characterization of phenolic compounds in food: From targeted analysis to metabolomics-based approaches, TrAC Trends Anal. Chem. 133 (2020), 116083, https://doi.org/10.1016/J.TRAC.2020.116083.
[110] Irina Porras-Mija, Rosana Chirinos, Diego Garcia-Rios, Ana Aguilar-Galvez, Cinthya Huaman-Alvino, Romina Pedreschi, David Campos, Physico-chemical characterization, metabolomic profile and in vitro antioxidant, antihypertensive, antiobesity and antidiabetic properties of Andean elderberry (Sambucus nigra subsp. peruviana), J. Berry Res. 10 (2) (2020) 193–208, https://doi.org/10.3233/JBR-190439.
[111] Eduardo Puértolas, Oliver Cregenzán, Elisa Luengo, Ignacio Álvarez, Javier Raso, Pulsed-electric-field-assisted extraction of anthocyanins from purple-fleshed potato, Food Chemistry 136 (3–4) (2013) 1330–1336, https://doi.org/10.1016/j.foodchem.2012.09.080.
[112] Qianqian Qi, Meijun Chu, Xiuting Yu, Yanning Xie, Yali Li, Yongmei Du, Xinmin Liu, Zhongfeng Zhang, John Shi, Ning Yan, Anthocyanins and Proanthocyanidins: Chemical Structures, Food Sources, Bioactivities, and Product Development, Food Rev. Int. (2022) 1–29, https://doi.org/10.1080/87559129.2022.2029479.
[113] G.W. Qin, H. Han, W. Ding, Y.L. Wang, Optimization of extracting technology of anthocyanins from blueberry pomace by supercritical carbon dioxide, Appl. Chem. Ind. 48 (2019) 109–112.
[114] Muangrat Rattana, Paul T. Williams, Phanat Saengcharoenrat, Subcritical Solvent Extraction of Total Anthocyanins from Dried Purple Waxy Corn: Influence of Process Conditions, J. Food Process. Preserv. 41 (6) (2017), e13252, https://doi.org/10.1111/jfpp.13252.
[115] José Luis Pasquel Reátegui, Ana Paula da Fonseca Machado, Gerardo F. Barbero, Camila A. Rezende, Julian Martínez, Extraction of antioxidant compounds from blackberry (*Rubus* sp.) bagasse using supercritical CO_2 assisted by ultrasound, J. Supercrit. Fluids 94 (2014) 223–233, https://doi.org/10.1016/j.supflu.2014.07.019.
[116] L. Renai, C.V.A. Scordo, U. Chiuminatto, M. Ulaszewska, E. Giordani, W. A. Petrucci, F. Tozzi, S. Nin, M. Del Bubba, Liquid chromatographic quadrupole time-of-flight mass spectrometric untargeted profiling of (poly) phenolic compounds in Rubus idaeus L. and Rubus occidentalis L. fruits and their comparative evaluation, Antioxidants 10 (5) (2021) 704, https://doi.org/10.3390/antiox10050704.
[117] Noam Reshef, Aaron Fait, Nurit Agam, Grape berry position affects the diurnal dynamics of its metabolic profile, Plant, Cell Environ. 42 (6) (2019) 1897–1912, https://doi.org/10.1111/pce.13522.
[118] Sueli Rodrigues, Fabiano A.N. Fernandes, Edy Sousa de Brito, Adriana Dutra Sousa, Narendra Narain, Ultrasound extraction of phenolics and anthocyanins from jabuticaba peel, Ind. Crops Prod. 69 (2015) 400–407, https://doi.org/10.1016/j.indcrop.2015.02.059.
[119] Joseph A. Rothwell, Jara Perez-Jimenez, Vanessa Neveu, Alexander Medina-Remon, Nouha M'hiri, Paula García-Lobato, Claudine Manach, et al., Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content, Database 2013 (2013), https://doi.org/10.1093/database/bat070.
[120] Christoph Ruttkies, Emma L. Schymanski, Sebastian Wolf, Juliane Hollender, Steffen Neumann, MetFrag relaunched: incorporating strategies beyond in silico fragmentation, J. Chemin.-. 8 (1) (2016) 1–16, https://doi.org/10.1186/s13321-016-0115-9.
[121] Luiz Bruno de Sousa Sabino, Elenilson Godoy Alves Filho, Fabiano Andre Narciso Fernandes, Edy Sousa de Brito, Ivanildo Jose da Silva Junior, Optimization of pressurized liquid extraction and ultrasound methods for recovery of anthocyanins present in jambolan fruit (Syzygium cumini L.), Food Bioprod. Process. 127 (2021) 77–89, https://doi.org/10.1016/j.fbp.2021.02.012.
[122] Nabeelah Bibi Sadeer, Gabriele Rocchetti, Biancamaria Senizza, Domenico Montesano, Gokhan Zengin, Ahmet Uysal, Rajesh Jeewon, Luigi Lucini, Mohamad Fawzi Mahomoodally, Untargeted metabolomic profiling, multivariate analysis and biological evaluation of the true mangrove (Rhizophora mucronata Lam.), Antioxidants 8 (10) (2019) 489, https://doi.org/10.3390/antiox8100489.
[123] Supradip Saha, Ja.Shbir Singh, Anindita Paul, Rohan Sarkar, Zareen Khan, Kaushik Banerjee, Anthocyanin profiling using uv–vis spectroscopy and liquid chromatography mass spectrometry, J. AOAC Int. 103 (1) (2020) 23–39, https://doi.org/10.5740/jaoacint.19-0201.
[124] Marco Santin, L. Lucini, A. Castagna, G. Chiodelli, M.T. Hauser, A. Ranieri, Post-harvest UV-B radiation modulates metabolite profile in peach fruit, Postharvest Biol. Technol. 139 (2018) 127–134, https://doi.org/10.1016/j.postharvbio.2018.02.001.
[125] Millena Cristina Barros Santos, Luciana Ribeiro da Silva Lima, Fabiana Ramos Nascimento, Talita Pimenta do Nascimento, Luiz Claudio Cameron, Mariana Simões Larraz Ferreira, Metabolomic approach for characterization of phenolic compounds in different wheat genotypes during grain development, Food Res. Int. 124 (2019) 118–128, https://doi.org/10.1016/j.foodres.2018.08.034.
[126] Giorgia Sarais, G. D'Urso, Carla Lai, Filippo Maria Pirisi, C. Pizza, P. Montoro, Targeted and untargeted mass spectrometric approaches in discrimination between Myrtus communis cultivars from Sardinia region, J. Mass Spectrom. 51 (9) (2016) 704–715, https://doi.org/10.1002/jms.3811.
[127] Alexandra C. Schrimpe-Rutledge, Simona G. Codreanu, Stacy D. Sherrod, John A. McLean, Untargeted metabolomics strategies—challenges and emerging directions, J. Am. Soc. Mass Spectrom. 27 (12) (2016) 1897–1905, https://doi.org/10.1007/s13361-016-1469-y.
[128] Sonia Scorrano, Maria Rosaria Lazzoi, Lucia Mergola, Maria Pia Di Bello, Roberta Del Sole, Giuseppe Vasapollo, Anthocyanins profile by Q-TOF LC/MS in Myrtus communis berries from Salento Area, Food Anal. Methods 10 (7) (2017) 2404–2411, https://doi.org/10.1007/s12161-017-0813-6.
[129] Jincheng Shen, Wanlu Shao, Zhaokui Du, Hongfei Lu, Junmin Li, Integrated metabolomic and transcriptomic analyses reveal differences in the biosynthetic pathway of anthocyanins in Fragaria nilgerrensis and Fragaria pentaphylla, Sci. Hortic. 271 (2020), 109476, https://doi.org/10.1016/j.scienta.2020.109476.
[130] Nan Shen, Tongfei Wang, Quan Gan, Sian Liu, Li Wang, Biao Jin, Plant flavonoids: Classification, distribution, biosynthesis, and antioxidant activity, Food Chem. (2022), 132531, https://doi.org/10.1016/j.foodchem.2022.132531.
[131] Bin Shi, Hongxia Wu, Bin Zheng, Minjie Qian, Aiping Gao, Kaibing Zhou, Analysis of Light-Independent Anthocyanin Accumulation in Mango (Mangifera indica L.), Horticulturae 7 (11) (2021) 423, https://doi.org/10.3390/horticulturae7110423.
[132] Ni Shi, Xiaoxin Chen, Tong Chen, Anthocyanins in Colorectal Cancer Prevention Review, Antioxidants 10 (10) (2021) 1600, https://doi.org/10.3390/antiox10101600.
[133] Qianqian Shi, Jiangtao Du, Dajun Zhu, Xi Li, Xingang Li, Metabolomic and transcriptomic analyses of anthocyanin biosynthesis mechanisms in the color mutant Ziziphus jujuba cv. Tailihong, J. Agric. Food Chem. 68 (51) (2020) 15186–15198, https://doi.org/10.1021/acs.jafc.0c05334.
[134] Liang Shuai, Huan Liu, Lingyan Liao, Tingting Lai, Ziying Lai, Xinxin Du, Zhenhua Duan, Zhenxian Wu, Tao Luo, Widely targeted metabolic analysis

revealed the changed pigmentation and bioactive compounds in the ripening Berchemia floribunda (Wall.) Brongn. fruit, Food Sci. Nutr. 9 (3) (2021) 1375–1387, https://doi.org/10.1002/fsn3.2093.
[135] Rui P.F.F. Silva, Teresa A.P. Rocha-Santos, Armando C. Duarte, Supercritical fluid extraction of bioactive compounds, TrAC Trends Anal. Chem. 76 (2016) 40–51, https://doi.org/10.1016/j.trac.2015.11.013.
[136] S. Silva, E.M. Costa, C. Calhau, R.M. Morais, M.E. Pintado, Anthocyanin extraction from plant tissues: A review, Crit. Rev. Food Sci. Nutr. 57 (14) (2017) 3072–3083, https://doi.org/10.1080/10408398.2015.1087963.
[137] Jashbir Singh, Khushbu Sharma, Suresh Walia, Supradip Saha, Anthocyanin profiling method based on isocratic elution for comparable speed, reproducibility, and quantitation with gradient elution, Food Anal. Methods 10 (1) (2017) 118–128, https://doi.org/10.1007/s12161-016-0561-z.
[138] Jianghao Sun, Long-ze Lin, Pei Chen, Study of the mass spectrometric behaviors of anthocyanins in negative ionization mode and its applications for characterization of anthocyanins and non-anthocyanin polyphenols, Rapid Commun. Mass Spectrom. 26 (9) (2012) 1123–1133, https://doi.org/10.1002/rcm.6209.
[139] Yangzhao Sun, Xiaojun Liao, Zhengfu Wang, Xiaosong Hu, Fang Chen, Optimization of Microwave-Assisted Extraction of Anthocyanins in Red Raspberries and Identification of Anthocyanin of Extracts Using High-Performance Liquid Chromatography - Mass Spectrometry, 2007, Eur. Food Res. Technol. 225 (2007) 511–523, https://doi.org/10.1007/s00217-006-0447-1.
[140] L. Sunil, Nandini P. Shetty, Biosynthesis and regulation of anthocyanin pathway genes, Appl. Microbiol. Biotechnol. (2022) 1–16, https://doi.org/10.1007/s00253-022-11835-z.
[141] Swathi Suresh, Rukaiah Fatma Begum, Ankul Singh, V. Chitra, Anthocyanin as a Therapeutic in Alzheimer's Disease: A Systematic Review of Preclinical Evidences, Ageing Res. Rev. (2022), 101595, https://doi.org/10.1016/j.arr.2022.101595.
[142] Jiaqi Tan, Yanmei Han, Bo Han, Xiangmei Qi, Xu Cai, Shaoqin Ge, Hongkun Xue, Extraction and Purification of Anthocyanins: A Review, J. Agric. Food Res. 8 (2022), 100306, https://doi.org/10.1016/j.jafr.2022.100306.
[143] Matthew D. Teegarden, Steven J. Schwartz, Jessica L. Cooperstone, Profiling the impact of thermal processing on black raspberry phytochemicals using untargeted metabolomics, Food Chem. 274 (2019) 782–788, https://doi.org/10.1016/j.foodchem.2018.09.053.
[144] Noelia Tena, Agustin G. Asuero, Up-to-date analysis of the extraction methods for anthocyanins: Principles of the techniques, optimization, technical progress, and industrial application, Antioxidants 11 (2) (2022) 286, https://doi.org/10.3390/antiox11020286.
[145] Ketti Toffali, Anita Zamboni, Andrea Anesi, Matteo Stocchero, Mario Pezzotti, Marisa Levi, Flavia Guzzo, Novel aspects of grape berry ripening and post-harvest withering revealed by untargeted LC-ESI-MS metabolomics analysis, Metabolomics 7 (3) (2011) 424–436, https://doi.org/10.1007/s11306-010-0259-y.
[146] Takayuki Tohge, Federico Scossa, Regina Wendenburg, Pierre Frasse, Ilse Balbo, Mutsumi Watanabe, Saleh Alseekh, et al., Exploiting natural variation in tomato to define pathway structure and metabolic regulation of fruit polyphenolics in the lycopersicum complex, Mol. Plant 13 (7) (2020) 1027–1046, https://doi.org/10.1016/j.molp.2020.04.004.
[147] Gamze Toydemir, Esra Capanoglu, Maria Victoria Gomez Roldan, Ric C.H. de Vos, Dilek Boyacioglu, Robert D. Hall, Jules Beekwilder, Industrial processing effects on phenolic compounds in sour cherry (Prunus cerasus L.) fruit, Food Res. Int. 53 (1) (2013) 218–225, https://doi.org/10.1016/j.foodres.2013.04.009.
[148] Maria Urrutia, Wilfried Schwab, Thomas Hoffmann, Amparo Monfort, Genetic dissection of the (poly) phenol profile of diploid strawberry (Fragaria vesca) fruits using a NIL collection, Plant Sci. 242 (2016) 151–168, https://doi.org/10.1016/j.plantsci.2015.07.019.
[149] Joshua VanderWeide, Ilce G. Medina-Meza, Tommaso Frioni, Paolo Sivilotti, Rachele Falchi, Paolo Sabbatini, Enhancement of fruit technological maturity and alteration of the flavonoid metabolomic profile in Merlot (Vitis vinifera L.) by early mechanical leaf removal, J. Agric. Food Chem. 66 (37) (2018) 9839–9849, https://doi.org/10.1021/acs.jafc.8b02709.
[150] Patricia Velásquez, Daniela Bustos, Gloria Montenegro, Ady Giordano, Ultrasound-Assisted Extraction of Anthocyanins Using Natural Deep Eutectic Solvents and Their Incorporation in Edible Films, Molecules 26 (4) (2021) 984, https://doi.org/10.3390/molecules26040984.
[151] Ana A. Vilas-Boas, D.ébora A. Campos, Catarina Nunes, S.ónia Ribeiro, João Nunes, Ana Oliveira, Manuela Pintado, Polyphenol extraction by different techniques for valorisation of non-compliant portuguese sweet cherries towards a novel antioxidant extract, Sustainability 12 (14) (2020) 5556, https://doi.org/10.3390/su12145556.
[152] Fei Wang, Jaanus Liigand, Siyang Tian, David Arndt, Russell Greiner, David S. Wishart, CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification, Anal. Chem. 93 (34) (2021) 11692–11700, https://doi.org/10.1021/acs.analchem.1c01465.
[153] Lei Wang, Xiaoliang Sun, Jakob Weiszmann, Wolfram Weckwerth, System-level and granger network analysis of integrated proteomic and metabolomic dynamics identifies key points of grape berry development at the interface of primary and secondary metabolism, Front. Plant Sci. 8 (2017) 1066, https://doi.org/10.3389/fpls.2017.01066.
[154] Ru Wang, Sangram Keshari Lenka, Varun Kumar, Noga Sikron-Persi, Irena Dynkin, David Weiss, Avichai Perl, Aaron Fait, Michal Oren-Shamir, A Synchronized Increase of Stilbenes and Flavonoids in Metabolically Engineered Vitis vinifera cv. Gamay Red Cell Culture, J. Agric. Food Chem. 69 (28) (2021) 7922–7931, https://doi.org/10.1021/acs.jafc.1c02119.
[155] Yuwei Wang, Ying Ye, Le Wang, Wei Yin, Jian Liang, Antioxidant Activity and Subcritical Water Extraction of Anthocyanin from Raspberry Process Optimization by Response Surface Methodology, Food Biosci. 44 (2021), 101394, https://doi.org/10.1016/j.fbio.2021.101394.
[156] Chandré M. Willemse, Maria A. Stander, André De Villiers, Hydrophilic Interaction Chromatographic Analysis of Anthocyanins, J. Chromatogr. A 1319 (2013) 127–140, https://doi.org/10.1016/j.chroma.2013.10.045.
[157] David S. Wishart, An.Chi Guo, Eponine Oler, Fei Wang, Afia Anjum, Harrison Peters, Raynard Dizon, et al., HMDB 5.0: the Human Metabolome Database for 2022, Nucleic Acids Res. 50 (D1) (2022) D622–D631, https://doi.org/10.1093/nar/gkab1062.
[158] Yaqiong Wu, Chunhong Zhang, Zhengjin Huang, Lianfei Lyu, Jing Li, Weilin Li, Wenlong Wu, The color difference of rubus fruits is closely related to the composition of flavonoids including anthocyanins, LWT 149 (2021), 111825, https://doi.org/10.1016/j.lwt.2021.111825.
[159] Feng Xian, Christopher L. Hendrickson, Alan G. Marshall, High Resolution Mass Spectrometry, 2012, Anal. Chem. 84 (2) (2012) 708–719, https://doi.org/10.1021/ac203191t.
[160] Ming Xin, Changbao Li, Xuemei He, Li Li, Ping Yi, Yayuan Tang, Jiemin Li, Guoming Liu, Jinfeng Sheng, Jian Sun, Integrated metabolomic and transcriptomic analyses of quality components and associated molecular regulation mechanisms during passion fruit ripening, Postharvest Biol. Technol. 180 (2021), 111601, https://doi.org/10.1016/j.postharvbio.2021.111601.
[161] Jidi Xu, Jinjiao Yan, Wenjie Li, Qianying Wang, Caixia Wang, Junxing Guo, Dali Geng, Qingmei Guan, Fengwang Ma, Integrative analyses of widely targeted metabolic profiling and transcriptome data reveals molecular insight into metabolomic variations during apple (Malus domestica) fruit development and ripening, Int. J. Mol. Sci. 21 (13) (2020) 4797, https://doi.org/10.3390/ijms21134797.
[162] Ranran Xu, Limin Wang, Kexin Li, Jiankang Cao, Zhilei Zhao, Integrative transcriptomic and metabolomic alterations unravel the effect of melatonin on mitigating postharvest chilling injury upon plum (cv. Friar) fruit, Postharvest Biol. Technol. 186 (2022), 111819, https://doi.org/10.1016/j.postharvbio.2021.111819.
[163] Hongkun Xue, Hao Xu, Xiaorui Wang, Liuyang Shen, Han Liu, Chai Liu, Qingyu Qin, Xianzhe Zheng, Qingyan Li, Effects of microwave power on extraction kinetic of anthocyanin from blueberry powder considering absorption of microwave energy, J. Food Qual. 2018 (2018), https://doi.org/10.1155/2018/9680184.
[164] Oscar Yanes, Ralf Tautenhahn, Gary J. Patti, Gary Siuzdak, Expanding Coverage of the Metabolome for Global Metabolite Profiling, Anal. Chem. 83 (6) (2011) 2152–2161, https://doi.org/10.1021/ac102981k.
[165] Bohan Yang, Shuang He, Yuan Liu, Buchun Liu, Yanlun Ju, Dengzhao Kang, Xiangyu Sun, Yulin Fang, Transcriptomics integrated with metabolomics reveals the effect of regulated deficit irrigation on anthocyanin biosynthesis in Cabernet Sauvignon grape berries, Food Chem. 314 (2020), 126170, https://doi.org/10.1016/j.foodchem.2020.126170.
[166] Haiying Yang, Changping Tian, Shujun Ji, Fengzhu Ni, Xinguang Fan, Yanqing Yang, Chanchan Sun, Hansheng Gong, Aidi Zhang, Integrative analyses of metabolome and transcriptome reveals metabolomic variations and candidate genes involved in sweet cherry (Prunus avium L.) fruit quality during development and ripening, PloS One 16 (11) (2021), e0260004, https://doi.org/10.1371/journal.pone.0260004.
[167] Heejung Yang, Hyun Woo Kim, Yong Soo Kwon, Ho Kyong Kim, Sang Hyun Sung, Fast and Simple Discriminative Analysis of Anthocyanins-Containing Berries Using LC/MS Spectral Data, Phytochem. Anal. 28 (5) (2017) 416–423, https://doi.org/10.1002/pca.2689.
[168] Zhendong Yang, Weiwei Zhai, Optimization of Microwave-Assisted Extraction of Anthocyanins from Purple Corn (Zea mays L.) Cob and Identification with HPLC–MS, Innov. Food Sci. Emerg. Technol. 11 (3) (2010) 470–476, https://doi.org/10.1016/j.ifset.2010.03.003.
[169] Debao Yi, Hongna Zhang, Biao Lai, Liqin Liu, Xiaolu Pan, Zhiling Ma, Yicheng Wang, Jianghui Xie, Shengyou Shi, Yongzan Wei, Integrative analysis of the coloring mechanism of red longan pericarp through metabolome and transcriptome analyses, J. Agric. Food Chem. 69 (6) (2020) 1806–1815, https://doi.org/10.1021/acs.jafc.0c05023.
[170] Min Yu, Yuping Man, Rui Lei, Xuemei Lu, Yanchang Wang, Metabolomics study of flavonoids and anthocyanin-related gene analysis in kiwifruit (Actinidia chinensis) and kiwiberry (Actinidia arguta, Plant Mol. Biol. Report. 38 (3) (2020) 353–369, https://doi.org/10.1007/s11105-020-01200-7.
[171] Sara Zenoni, Marianna Fasoli, Flavia Guzzo, Silvia Dal Santo, Alessandra Amato, Andrea Anesi, Mauro Commisso, et al., Disclosing the molecular basis of the postharvest life of berry in different grapevine genotypes, Plant Physiol. 172 (3) (2016) 1821–1843, https://doi.org/10.1104/pp.16.00865.
[172] Aiai Zhang, Jing Zheng, Xuemiao Chen, Xueyin Shi, Huaisong Wang, Qiushi Fu, Comprehensive analysis of transcriptome and metabolome reveals the flavonoid metabolic pathway is associated with fruit peel coloration of melon, Molecules 26 (9) (2021) 2830, https://doi.org/10.3390/molecules26092830.
[173] J. Zhang, R. Singh, S.Y. Quek, Extraction of anthocyanins from natural sources—Methods and commercial considerations. Anthocyanins from Natural Sources: Exploiting Targeted Delivery for Improved Health, Royal Society of Chemistry, London, 2019, pp. 77–105, https://doi.org/10.1039/9781788012614-00077.

[174] Lixia Zhang, Gongjian Fan, Muhammad Ammar Khan, Zheng Yan, Trust Beta, Ultrasonic-assisted enzymatic extraction and identification of anthocyanin components from mulberry wine residues, Food Chem. 323 (2020), 126714, https://doi.org/10.1016/j.foodchem.2020.126714.
[175] Xianan Zhang, Mingshen Su, Jihong Du, Huijuan Zhou, Xiongwei Li, Xin Li, Zhengwen Ye, Comparison of phytochemical differences of the pulp of different peach [Prunus persica (L.) Batsch] cultivars with alpha-glucosidase inhibitory activity variations in China using UPLC-Q-TOF/MS, Molecules 24 (10) (2019) 1968, https://doi.org/10.3390/molecules24101968.
[176] Yajun Zhou, Xutong Zhao, Hui Huang, Effects of Pulsed Electric Fields on Anthocyanin Extraction Yield of Blueberry Processing By-Products, J. Food Process. Preserv. 39 (6) (2015) 1898–1904, https://doi.org/10.1111/jfpp.12427.

Capítulo 2 - Perfil metabolômico integrado direcionado e não direcionado de espécies de baunilha da Mata Atlântica: Revelando o potencial bioeconômico da *Vanilla cribbiana*

O Capítulo 2 desta tese corresponde a um artigo original de metabolômica aplicada a espécies nativas do gênero *Vanilla* ocorrentes no Brasil, publicado na revista *Food Chemistry* em outubro de 2024 (Oliveira et al., 2025), e constitui a aplicação empírica direta dos fundamentos conceituais discutidos no capítulo anterior. Enquanto o Capítulo 1 estabelece a metabolômica como ferramenta central para a caracterização de classes químicas complexas, com foco em antocianinas, este capítulo transpõe essa abordagem para o estudo de polifenóis estruturalmente associados às principais moléculas aromáticas das baunilhas, incluindo derivados fenólicos, ácidos cinâmicos e compostos relacionados à biossíntese da vanilina. Inserido no eixo estruturante da tese, o capítulo materializa o papel da metabolômica como estratégia de bioprospecção orientada por dados, voltada à geração de perfis químicos de alto rendimento capazes de sustentar análises comparativas, funcionais e preditivas em etapas subsequentes do projeto. O estudo baseia-se na seleção de espécies nativas da Mata Atlântica, com destaque para *Vanilla cribbiana*, e na aplicação integrada de abordagens direcionadas e não direcionadas por LC–HRMS, combinadas a fluxos robustos de processamento, anotação molecular e classificação química, permitindo a exploração simultânea da diversidade metabólica global e da quantificação de metabólitos-chave associados ao aroma de baunilha. Os resultados revelam uma elevada complexidade química nas espécies nativas analisadas, com a anotação de um conjunto expressivo de metabólitos pertencentes a diferentes classes químicas, incluindo compostos fenólicos e polifenóis relevantes para o perfil sensorial e funcional, além da identificação de metabólitos ainda não descritos previamente para o gênero *Vanilla*. As análises comparativas evidenciam perfis metabólicos distintos entre *V. cribbiana* e a espécie comercialmente dominante, *Vanilla planifolia*, ao mesmo tempo em que demonstram similaridades quantitativas relevantes em moléculas centrais para o aroma característico da baunilha, como a vanilina e seus precursores biossintéticos, reforçando o potencial bioeconômico de espécies nativas brasileiras. A discussão dos dados integra os resultados metabolômicos com evidências da literatura sobre composição aromática, percepção sensorial e aceitação do consumidor, sustentando a interpretação de que *V. cribbiana* reúne atributos químicos comparáveis aos da baunilha comercial, associados a características adicionais que ampliam sua singularidade química. Para além do caráter descritivo, o conjunto de dados gerado neste capítulo é concebido como informação estruturante e reutilizável, passível de integração em etapas

subsequentes de descoberta de potencial bioativo por meio de estratégias *in silico*, como *docking* molecular e modelagem farmacocinética, bem como de análises ecológicas e espaciais orientadas à conservação e ao uso sustentável da biodiversidade. Dessa forma, o Capítulo 2 consolida-se como o núcleo empírico químico da tese, demonstrando como a metabolômica integrada direcionada e não direcionada pode revelar o valor oculto de recursos vegetais nativos e fornecer a base analítica necessária para pipelines bioeconômicos que conectam diversidade química, bioatividade e sustentabilidade socioambiental.

Food Chemistry 464 (2025) 141650

Contents lists available at ScienceDirect

Food Chemistry

journal homepage: www.elsevier.com/locate/foodchem

Integrated targeted and untargeted metabolomics profiling of *Vanilla* species from the Atlantic Forest: Unveiling the bioeconomic potential of *Vanilla cribbiana*

Joana Paula da Silva Oliveira [a], Rafael Garrett [b], Maria Gabriela Bello Koblitz [c], Andrea Furtado Macedo [a,*]

[a] *Integrated Laboratory of Plant Biology (LIBV), Institute of Biosciences, Federal University of the State of Rio de Janeiro – UNIRIO. Av. Pasteur, 458 Urca, Rio de Janeiro, RJ, Brazil*
[b] *Metabolomics Laboratory (LabMeta), Institute of Chemistry, Federal University of Rio de Janeiro – UFRJ. Av. Horacio Macedo, 1281 - Polo de Química – Cidade Universitária, Rio de Janeiro, RJ, Brazil*
[c] *Food and Nutrition Graduate Program (PPGAN), Federal University of the State of Rio de Janeiro – UNIRIO. Av. Pasteur, 296 Urca, Rio de Janeiro, RJ, Brazil*

ARTICLE INFO

Keywords:
Vanilla crop wild relatives
vanilla flavor
Vanillin
Orchidaceae
Sustainability
LC-HRMS

ABSTRACT

In this study, we employ both targeted and untargeted approaches to explore the metabolomic profiles of *Vanilla* spp., with a particular focus on *V. cribbiana* (VCR) and its comparison with *V. planifolia* (VP). We also examine *V. bahiana* and *V. chamissonis* using targeted approaches. Through advanced analytical techniques, our untargeted LC-HRMS approach led to the annotation of 60 metabolites, revealing a complex chemical composition with 34 novel compounds in the *Vanilla* genus in VCR and VP. These findings highlight significant flavoring compounds and lay the foundation for a subsequent quantitative estimation approach. Our targeted analysis, which measured key molecules, underscores VCR's potential in producing vanillin and acetovanillone at levels comparable to the commercially valuable VP and even higher levels of vanillic acid. This research enriches our understanding of flavor composition in vanilla species and emphasizes the importance of exploring wild relatives of vanilla crop for sustainable production and biodiversity conservation.

1. Introduction

Primarily derived from *Vanilla planifolia*, vanilla is a globally significant commodity with multiple uses. Its commercial value stems from its widespread use as a flavoring agent, especially in food and beverages, where it is essential in products such as ice cream, baked goods, chocolates, and various drinks (Baqueiro-Peña & Guerrero-Beltrán, 2017). Beyond culinary uses, vanilla is also valuable in the fragrance industry for perfumes and in pharmaceutical industry for flavoring medicines (Banda Ortiz et al., 2021). The vanilla market is projected to grow at a compound annual growth rate (CAGR) of 4.7 % from 2022 to 2030, potentially reaching a valuation of USD 1.9 billion by the end of this period (Market Research Future, 2024). The Asia-Pacific region dominates the market, driven by rising demand for vanilla across various industries. Key market drivers include vanilla's flavoring properties, product versatility, and growing consumer preference for natural and organic ingredients (Market Research Future, 2024).

Vanilla species are native to neotropical regions, particularly Mexico and Central America. *Vanilla planifolia* is indigenous to Mexico, where it was first cultivated by Totonac people and later by the Aztecs (de Oliveira et al., 2022). However, Madagascar is now the largest vanilla producer, followed by Indonesia, with Mexico, Papua New Guinea, and

Abbreviations: AGC, Automatic Gain Control; ANOVA, Analysis of Variance; CAGR, Compound Annual Growth Rate; CID, PubChem Compound Identifier; CWR, Crop Wild Relatives; dd-MS2, Data-Dependent Acquisition Mass Spectrometry; ESI, Electrospray Ionization; GC-MS, Gas Chromatography-Mass Spectrometry; HPLC, High-Performance Liquid Chromatography; IT, Injection Time; LC-HRMS, Liquid Chromatography-High Resolution Mass Spectrometry; LOQ, Limit of Quantification; LOD, Limit of Detection; MAD, Median Absolute Deviation; m/z, Mass-to-Charge Ratio; NCE, Normalized Collision Energy; PRM, Parallel Reaction Monitoring; SPME, Solid Phase Microextraction; tR, Retention Time; VB, *Vanilla bahiana*; VC, *Vanilla chamissonis*; VCR, *Vanilla cribbiana*; VP, *Vanilla planifolia*..

* Corresponding author at: Integrated Laboratory of Plant Biology, Institute of Biosciences, Federal University of the State of Rio de Janeiro– UNIRIO. Av. Pasteur, 458, Urca, Rio de Janeiro 22290-240, Brazil.
E-mail addresses: joanapaulasoliveira@edu.unirio.br (J.P. da Silva Oliveira), rafael_garrett@iq.ufrj.br (R. Garrett), maria.koblitz@unirio.br (M.G.B. Koblitz), andrea.macedo@unirio.br (A.F. Macedo).

https://doi.org/10.1016/j.foodchem.2024.141650
Received 4 June 2024; Received in revised form 8 October 2024; Accepted 13 October 2024
Available online 15 October 2024
0308-8146/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

Tahiti also being significant producers. Madagascar's dominance in the vanilla market prevails, as it supplies the majority of the world's vanilla (de Oliveira et al., 2022). However, cultivating vanilla outside its native regions presents several challenges. Non-native regions often lack optimal environmental conditions for vanilla orchids, leading to increased susceptibility to diseases and pests, resulting in lower yields and higher production costs (Al-Khayri et al., 2019). Additionally, the need for artificial pollination, due to the absence of natural pollinators in the regions where vanilla is not native, adds to the labor intensity and cost of cultivation. Furthermore, introducing vanilla to non-native ecosystems can potentially disrupt local biodiversity and may pose ecological imbalances (Al-Khayri et al., 2019). Diseases affecting *V. planifolia* have a significant impact on the vanilla market due to their detrimental effects on both the quality and quantity of vanilla pod production (Kadir et al., 2019), reducing its availability in the market and potentially driving up prices (Chambers et al., 2021; Kadir et al., 2019).

Vanilla × *tahitensis*, a hybrid likely derived from *V. planifolia* and *V. odorata*, has a distinct flavor compared to the more common *V. planifolia*. Primarily cultivated in French Polynesia, it features a unique anise and caramel taste due to its high content of anisyl compounds. Understanding its volatile composition is vital for quality control and for distinguishing it from *V. planifolia*, which is known for its phenolic, woody, and smoky flavor (Brunschwig et al., 2016). The U.S. Food and Drug Administration has approved only *Vanilla planifolia* Andrews and *Vanilla* × *tahitensis* Moore for use in food dressings and flavorings (Food & Administration, 2023). Despite this, *V. pompona*, also known as West Indian vanilla, has emerged as a viable alternative for vanilla essence due to its unique chemical profile, which includes high concentrations of vanillin, 4-hydroxybenzyl alcohol, and anisyl alcohol. Although less commercialized than *V. planifolia*, *V. pompona* offers a distinct flavor and aroma profile, potentially diversifying the vanilla industry (Maruenda et al., 2013; Pérez-Silva et al., 2021).

Neotropical vanilla species, such as *V. bahiana* and *V. chamissonis* from the Atlantic Forest, have also emerged as important alternatives in the vanilla industry. These species have been identified as having significant bioeconomic potential due to their flavor compounds, which are essential for vanilla quality (da Silva Oliveira et al., 2022; Franco et al., 2024; Lopes et al., 2019). They contain various compounds, such as vanillin, 4-hydroxybenzoic acid, 4-vinylguaiacol, and vanillyl alcohol, traditionally associated with vanilla flavor. However, they also possess unique compounds not typically linked to commercial vanilla flavors, contributing to a broader array of flavor descriptors. These compounds may be associated with over 58 flavor notes, ranging from almond, anise, and caramel to woody, smoky, and spicy (da Silva Oliveira et al., 2022). Recent evaluations of *V. cribbiana* cultivated in Guatemala showed levels of aromatic compounds, particularly vanillin, comparable to *V. planifolia*, and significantly higher levels of 4-hydroxybenzoic acid (Galeana, 2024). Additionally, sensory analysis and consumer acceptance studies indicate that these native species have the potential to be used as substitutes for commercial species, with *V. cribbiana* from the Atlantic Forest showing particular promise for market success due to its favorable sensory profile and consumer acceptance (Franco et al., 2024). The aroma and sweetness of *V. cribbiana* closely match those of *V. planifolia*, indicating its potential as a natural alternative to vanilla flavoring. Exploring these species is essential for diversifying the sources of vanilla flavor sources and improving the genetics of vanilla crops (Chambers et al., 2021; Franco et al., 2024).

Numerous studies have shown that liquid chromatography and mass spectrometry (LC-MS) play a crucial role in vanilla research. High-Performance Liquid Chromatography (HPLC) with Ultra-Violet detection was utilized to quantify glucovanillin and other aromatic compounds in ripe fruits of *Vanilla planifolia* and related species, allowing for comparisons of aromatic profiles across different genotypes (Pérez-Silva et al., 2021). Untargeted LC-MS/MS metabolic profiling helped annotate essential flavor compounds in two Atlantic Forest species of *Vanilla*, revealing distinct flavor descriptors for these wild species (da Silva Oliveira et al., 2022). Furthermore, an LC-MS method was developed and validated for determining coumarin, vanillin, and ethyl vanillin in vanilla extract products, providing a reliable analytical technique for quality control (De Jager et al., 2007). This method was also compared with headspace-SPME-GC–MS for detecting and quantifying key flavor compounds in vanilla extracts (De Jager et al., 2008). Comparative metabolomics studies have utilized LC-MS to examine vanillin biosynthesis during curing, comparing metabolites in vanilla pods and beans (Gu et al., 2017; Yang et al., 2017). These studies underscore the vital role of LC-MS in analyzing vanilla samples, encompassing the quantification of flavor compounds, recovery of valuable substances, metabolic profiling of species, and quality control in vanilla product manufacturing.

This study aims to expand vanilla research by combining untargeted and targeted metabolomics using mass spectrometry. *V. cribbiana* was investigated using untargeted metabolomics, complementing previous work on the chemical composition of Atlantic Forest vanillas (*V. bahiana*, *V. chamissonis,* and *V. planifolia*) (da Silva Oliveira et al., 2022). Additionally, a targeted experiment was conducted in the same injection to quantify and compare key metabolites related to vanilla flavor among these species (*V. cribbiana, V. bahiana, V. planifolia* and *V. chamissonis*).

2. Material and methods

2.1. Chemicals

The chemical standards acetovanillone, 4-coumaric acid, 4-hydroxybenzaldehyde, 4-hydroxybenzoic acid, vanillic acid, and vanillin were purchased from Sigma-Aldrich (St. Louis, MO, USA). Unless otherwise stated, ethanol and methanol were of LC-MS purity grade (Honeywell Fluka™, Germany). Ultrapure water was obtained using a Milli-Q® Direct 8 Water Purification System (Millipore, Bedford, USA). Standard stock solutions were prepared at a concentration of 30 $\mu g.mL^{-1}$ by dissolving each standard in water:acetonitrile (3:7, *v*/v). An aliquot of each stock solution was mixed to achieve standard mixtures ranging from 0.001 to 30 $\mu g.mL^{-1}$ in water:acetonitrile (3:7, v/v). A calibration curve with triplicate calibrator points was used to quantify the compounds in the vanilla extracts.

2.2. Plant material

Three replicate samples of ripe pods from *V. bahiana* (VB), *V. chamissonis* (VC), *V. cribbiana* (VCR) and *V. planifolia* (VP) were collected from distinct locations within the Brazilian Atlantic Forest. Ripe VP pods were collected at the Botanic Garden of Rio de Janeiro. Each flower across all species underwent manual pollination, and all pod samples were collected in the morning within a consistent time frame, to ensure a similar stage of fruit maturity. By the eighth month, a noticeable yellowing appeared at the fruit tips without any signs of dehiscence. Upon collection in the nineth month, the pods were sectioned and preserved in a − 80 °C freezer within conical tubes. Representative vouchers for each species were deposited at the herbaria of both the Research Institute of Rio de Janeiro Botanic Garden (RB) and the Prof. Jorge Pedro Pereira Carauta Herbarium (HUNI), Federal University of the Rio de Janeiro State as follows: *V. bahiana* at RB (RB01111540), *V. chamissonis* at HUNI (4402), *V. planifolia* at RB (RB01400209), and *V. cribbiana* at HUNI (6715). The research concerning *Vanilla* spp. has been registered under the National System for the Management of Genetic Heritage and Associated Traditional Knowledge (SISGEN) with the number A41237C.

2.3. Enzymatic treatment and extraction

Triplicate samples each weighing 100 mg and derived from pooled

pods representing each species were ground in liquid nitrogen until a fine white powder was obtained. The subsequent enzymatic treatment followed a modified protocol based on Ruiz-Terán and collaborators (Ruiz-Terán et al., 2001). The pulverized pod material was dispersed and vortexed with 600 μL of a 0.05 M citrate-phosphate buffer (pH 5) Emulsin. One milliliter of almond ß-glucosidase (4 $U.mg^{-1}$, Sigma-Aldrich) prepared at 1 $mg.mL^{-1}$ was added to each sample, and the combined mixture was incubated in a water bath at 37 °C for 4 h. Then, the samples were stored at −80 °C until analysis. After thawing, samples were dried using a vacuum centrifuge (Savant, Thermo Scientific) and extracted with methanol. Four milliliters of methanol were added to every sample, which was then vortexed and submitted to Ultrasound-Assisted Extraction (UAE) utilizing an ultrasonic probe (Disruptor 500 W, Ultronique) operating at 80 % power for 8 min. This step was repeated three times, and the supernatant was collected and dried. The dry extracts were reconstituted in a mixture of methanol:acetonitrile: water (1:1:1, *v*/v/v) at a concentration of 10 $mg.mL^{-1}$. This solution was further diluted to obtain a 1 $mg.mL^{-1}$ solution in methanol:acetonitrile: water (1:1:2, v/v/v), which was then filtered through a PTFE syringe filter (13 mm, 0.22 μm pore diameter, Analytica) before LC-HRMS analysis. Pooled Quality Control (QC) samples were created by combining 30 μL of each filtered sample into a vial to evaluate the reproducibility of the LC-HRMS system and injected at regular intervals. Blank samples from extraction and injection solvents were also analyzed within the sample batch.

2.4. LC-HRMS data acquisition

LC-HRMS analyses were performed on a UHPLC Dionex Ultimate 3000 system coupled to a Q-Exactive Plus™ Orbitrap mass spectrometer (Thermo Fisher Scientific, USA) equipped with an electrospray ion source (ESI) operating in both positive (ESI+) and negative (ESI–) ionization modes. The chromatographic column was an ACQUITY UPLC® BEH C18 130 Å (2.1 × 100 mm, 1.7 μm particle size) (Waters, UK) maintained at 40 °C. The mobile phases consisted of (A) 0.1 % formic acid in Milli-Q water and (B) 0.1 % formic acid in acetonitrile. The elution gradient at a flow rate of 0.35 $mL.min^{-1}$ was structured as follows: 0–1.0 min B 5 %, 1.0–3.0 min B 5–35 %, 3.0–13.0 min B 35–95 %, 13.0–15.0 min B 95 %, 15.0–15.1 min B 5 %, and 15.1–18.0 min B 5 %. The injection volume was 8 μL.

Two types of MS acquisition experiments were performed. Samples and standard calibrants were analyzed using a combined targeted and untargeted experiment in the same injection consisting of: i) Full scan with polarity switch (positive/negative ion modes) at a resolution of 35,000 FWHM (full width at half maximum) within the *m/z* range of 100–850 (Full MS), AGC 1e6, and an injection time (IT) of 120 ms; and ii) Parallel Reaction Monitoring (PRM) acquisition with an inclusion list of *m/z* values from vanilla flavor compounds previously screened by our group and used for the targeted quantitation as follows: vanillin (*m/z* 153.0546201, t_R 5.13 min; [M + H]+), acetovanillone (*m/z*: 167.0702702, t_R: 5.56 min; [M + H]+), vanillic acid (*m/z*: 169.0495347, t_R: 3.82 min; [M + H]+), 4-hydroxybenzaldehyde (*m/z*: 121.0295034, t_R: 4.3 min; [M-H]-), 4-hydroxybenzoic acid (*m/z*: 137.024418, t_R: 2.55 min; [M-H]-), and 4-coumaric acid (*m/z*: 163.0400681, t_R: 5.19 min; [M-H]-). These molecules were selected due to their importance in flavor and fragrance, characteristic of commercial vanillas (Gallage et al., 2018; Pérez Silva et al., 2011; Toth, 2018). The capillary temperature was set at 300 °C, with spray voltages of 3600 V for ESI+ and 3300 V for ESI–, S-lens RF level 50. Sheath and auxiliary gas settings were 45 and 15, respectively. This Full scan experiment, combined with PRM, yielded between 7 and 10 data points per targeted peak used for quantitation. For compound annotation purposes only, the pooled quality control (QC) samples were used to acquire MS/MS fragmentation data via a separate Full scan experiment with polarity switch (positive/negative ion modes) combined with data-dependent acquisition (dd-MS2 top 3) at a resolution of 17,500 FWHM, AGC 2e5, IT 100 ms, NCE of 35 in ESI+ and 30 in ESI–. Extraction and solvent blank samples were applied to measure background noise and assess the instrumental response in the absence of the analytes, to ensure that any detected signals could be attributed to the target compounds.

The data files from LC-HRMS in the vendor-specific format were transformed into Analysis Base File format using the ABF converter tool (http://www.reifycs.com/AbfConverter). The converted files were then imported into the MS-DIAL software 4.9.221218 (Tsugawa et al., 2019). Processing parameters were: MS1 mass range of 100–1000 *m/z*; MS1 tolerance set at 0.01 and MS2 tolerance at 0.025; linear weighted moving average as the smoothing method (smoothing level of 3, minimum peak width 5, and minimum peak height 50,000). Peak spotting parameters were defined with a mass slice width of 0.1. Parameters for deconvolution were a sigma window value of 0.5 and isotopes retained until 0.5 Da. MS/MS identification settings were MS1 accurate mass tolerance of 0.01 Da and MS2 of 0.05 Da. The MS/MS spectral libraries utilized for annotations in .MSP format were the publicly accessible ESI+ and ESI– databases from the MS-Dial platform (https://systemsomicslab.github.io/compms/msdial/main.html, accessed on January 2024), version 17 (Matsuzawa et al., 2022). Alignment parameters were retention time tolerance of 0.5 min; MS1 tolerance of 0.015 Da; both retention time factor and MS1 factor at 0.5. Annotated features were filtered using detection in a at least two replicates within at least one group and a minimum sample-to-blank ratio of 10. For the annotation of molecules previously found in *Vanilla* spp., a custom database was created based on published papers annotations from the literature (da Silva Oliveira et al., 2022; Pérez-Silva et al., 2006; Pérez-Silva et al., 2021; Zhang & Mueller, 2012). The MS-FINDER software (version 3.60) (Tsugawa et al., 2016) was used for in silico annotation of compounds, with search parameters configured for formula prediction and structure elucidation: precursor-oriented spectral search; MS1 and MS2 tolerances set as 0.01 and 0.05 Da, respectively; isotopic ratio tolerance set at 20 % for formula calculation; in silico MS/MS fragmenter tree depth set at 2; with data sourced from FooDB, PlantCyc, and custom vanilla molecules databases. Only annotations scoring above 6.0, with a minimum of two explained fragmentation peaks, were accepted. The online portal FlavorDB (https://cosylab.iiitd.edu.in/flavordb) was used to assign putative flavor descriptors to annotated compounds (Garg et al., 2018). These descriptors, once assigned, were integrated to the discussion section to highlight potential flavor differences and similarities between VCR and VP. Ontological chemical classification was performed through the Classyfire application (Djoumbou Feunang et al., 2016).

2.5. Targeted data processing and molecule identification

The compounds 4-hydroxybenzaldehyde, 4-hydroxybenzoic acid, vanillin, 4-coumaric acid, acetovanillone, and vanillic acid were quantitatively analyzed. The regression equation for each target compound, delineating the relationship between concentration and instrumental response, was constructed by subtracting the intensity values of the blank samples from the target compound data to eliminate background noise. This calibration curve was established across seven concentrations: 0.1, 0.5, 1.0, 2.5, 10.0, 20.0, and 30.0 $\mu g.mL^{-1}$ (Supplementary Table A.1). Blank samples were tested for normal distribution of peak areas using the Shapiro-Wilk normality test at a 5 % significance level. The limits of detection (LOD) and quantification (LOQ) were calculated using a non-parametric approach based on the median absolute deviation (MAD) from the blank samples' peak areas. This method shifted from using the standard deviation of blank samples—where "*s*" is the calibration curve slope and "σ" is the response's standard deviation—to employing MAD, for its robustness against outliers, providing reliable estimates even in the presence of a non-normal distribution of blank noise (Chen et al., 2021; Wulff & Mitchell, 2018; Zheng et al., 2021). LOD and LOQ were calculated for each target compound, using confidence levels of 95 % (analogous for 10 σ/*s*) and 99 % (analogous for 3.3 σ/*s*), respectively. Standard samples responses were computed for each

target, and the LOD and LOQ values were integrated into the dataset (Table 1). Data acquired from samples with concentrations of 0.001 μg.mL^{-1} and 0.01 μg.mL^{-1} were filtered out from the analysis after LOD/LOQ evaluation. The concentration (μg.mL^{-1}) of each target metabolite detected in the samples was calculated using the linear equation from the calibration curves. This concentration was then multiplied by the yield of the samples after drying the extracts, and the result was divided by 1000 to express the values in mg. Initially, LOD and LOQ values were expressed in μg.mL^{-1} during the screening stage to determine valid concentrations for quantification estimation before converting them to the unit of mg/100 mg of pods dry weight. This conversion ensured accurate representation and comparability of the analytical data across different stages of the analysis. The amount of each targeted molecule in the VB, VCH, VCR, and VP samples was then calculated in mg/100 mg of dried vanilla pod weight, considering each replicate's yield after enzymatic treatment and extraction. Although the targeted quantitation method has not been fully validated according to standard guidelines, its results offer valuable insights into compound abundances and demonstrate the biotechnological potential of the studied species.

Fragmentation patterns of target precursors at various concentrations were compared to sample features to confirm molecular identities, using spectral data preprocessed to remove non-significant peaks. The analysis involved comparing spectral files of standards and samples, with a 0.005 *m/z* tolerance for peak matching and a focus on peaks above 10,000 intensity units for significance. Matches required cosine similarity scores ≥0.9 based on matched peak intensities and with at least three matched peaks (Supplementary Table B.1).

2.6. Statistical analyses

Differential analysis was performed, and a volcano plot was generated to exhibit differences in metabolite intensities between VCR and VP groups using untargeted data. Only VCR was compared with the commercial species, as an untargeted study using the same methods for VC and VB had already been conducted (da Silva Oliveira et al., 2022). Metabolites with *p*-values (threshold ≤0.05) from two-sample *t*-tests and log2 fold changes (threshold ≤ −2 and ≥ 2) comparing group means informed the basis of the plot. Based on these thresholds, significant metabolites were classified as having either considerably higher or lower intensities, while others were classified as unchanged.

Based on targeted data, ANOVA and Tukey's tests (*p*-values ≤0.05) were applied to compare metabolite amounts across groups (VB, VCH, VCR, VP). *Post-hoc* analysis with Tukey's test, using the agricolae package's `HSD.test` function, identified significant differences, assigning letters to indicate statistically distinct groups. The standard error of the means was incorporated as error bar ranges in the generated bar plots.

3. Results

3.1. Untargeted metabolite annotation and differential analysis

The metabolic profiles of *Vanilla* spp. were investigated, specifically comparing VCR and VP, utilizing an untargeted approach and ontological chemical classification methods as defined by Djoumbou Feunang et al. (2016). In total, 60 metabolites were annotated in the mass spectrum data for VCR and VP (Table 2). Of these, 31 were identified using ESI– and 35 using ESI+, with 6 metabolites common to both ionization modes. Among these metabolites, 26 had previously been annotated in *Vanilla* spp., based on a custom database composed exclusively of compounds found in vanilla. In contrast, 34 metabolites were annotated as compounds not previously described for VCR and VP (Table 2). All annotations were classified as level 2 confidence (Malinowska & Viant, 2019), indicating putative annotations based on matching spectra with in silico fragmentation data from public libraries compounds. Of these, 39 % were confirmed by matching the mass spectra data against both experimental databases and in silico fragmentation processing. The remaining 61 % were annotated based solely on in silico fragmentation processing. Twenty different Classes were attributed to the annotated compounds: fatty acyls (15.15 %), organooxygen compounds (15.15 %), carboxylic acids and derivatives (13.64 %), benzene and substituted derivatives (9.09 %), flavonoids (7.58 %), hydroxy acids and derivatives (7.58 %), phenols (7.58 %), cinnamic acids and derivatives (4.55 %), cinnamaldehydes (3.03 %), dihydrofurans (3.03 %), furans (3.03 %), glycerolipids (1.52 %), homoisoflavonoids (1.52 %), organonitrogen compounds (1.52 %), prenol lipids (1.52 %), pteridines and derivatives (1.52 %), purine nucleosides (1.52 %), and pyrans (1.52 %).

Using the commercial *V. planifolia* as the benchmark for optimal quality in the production of vanilla flavor-associated molecules, a differential analysis was implemented to scrutinize the varying intensities of annotated metabolites and detected features between VCR and VP, employing statistical significance via *t*-test analysis (*p*-value ≤0.05) and Log2 fold change (Fig. 1). In ESI–, annotated metabolites, known in the *Vanilla* genus, demonstrated higher intensity (peak height) in VCR than commercial vanilla, including acetovanillone, L-asparagine, and 4-coumaric acid. Conversely, in VP, the metabolites of notable intensity were 4-hydroxybenzaldehyde, malic acid, quinic acid, sucrose, and vanillyl alcohol (Fig. 1). Additionally, the dereplication process revealed molecules previously unreported in these species, rendering the annotation of 3-hydroxysuberic acid; 7,8-dihydropteroic acid; 7-hydroxy-3-(3-hydroxy-4-methoxybenzyl)-5-methoxy-4-chromanone; galactonic acid; hesperetin; and 4-coumaraldehyde with higher intensity in VCR, and daucic acid in VP. In ESI+, metabolites that were more pronounced in VCR as opposed to commercial vanilla comprised cinnamaldehyde, cinnamic acid, and L-asparagine, while VP exhibited 4-hydroxybenzaldehyde, benzaldehyde, L-isoleucine (or L-leucine), methyl 2-furoate; and sucrose with higher intensities (Fig. 1). The sole dereplication in this ionization mode indicated a higher abundance in VP of 2-linoleoylglycerol and 7-hydroxy-2-methyl-4-oxo-4H-1-benzopyran-5-carboxylic acid 7-glucoside. Other vanilla-related compounds that showed no differential abundance relative to any of the species were, in ESI–: 3,5-dihydroxy-6-methyl-2,3-dihydropyran-4-one; 4-hydroxybenzoic acid; D-glucopyranose; 2,4-dihydroxy-2,5-dimethylfuran-3-one; homocitric acid; methyl furan-2-carboxylate; and vanillin, while in ESI+, they were: guaiacol; *m*-anisaldehyde; oleic acid; 4-coumaric acid; vanillic acid; vanillin; homocitric acid; and linoleic acid.

Untargeted analysis results highlighted molecules associated with vanilla's characteristic flavor (e.g., acetovanillone) and vanillin's biosynthesis (e.g., 4-coumaric acid), which were more prevalent in VCR.

Table 1
Analytical performance characteristics of targeted analytical standards in liquid chromatography-mass spectrometry (LC-MS) analysis. The table presents retention times (t_R, min), mass-to-charge ratios (*m/z*), coefficient of determination (r^2), limits of detection (LOD, μg.ml^{-1}), and limits of quantification (LOQ, μg.ml^{-1}) for each compound.

Standard	t_R (min)	*m/z*	Ion mode	r^2	LOD (μg.ml^{-1})	LOQ (μg.ml^{-1})
4-hydroxybenzaldehyde	4.37	121.0295	ESI–	0.99	0.00087	0.00115
4-hydroxybenzoic acid	2.55	137.0244	ESI–	0.99	0.00375	0.00493
acetovanillone	5.53	167.0703	ESI+	0.98	0.00242	0.00318
4-coumaric acid	5.19	163.0401	ESI–	0.98	0.01640	0.05466
vanillic acid	4.54	163.0401	ESI+	0.99	0.02351	0.03090
vanillin	5.14	153.0546	ESI+	0.99	0.00225	0.00296

Table 2
Molecule annotations from *V. cribbiana* and *V. planifolia* methanolic extracts mass spectrometry data acquisition. The table includes annotations of molecules detected in the methanolic extracts of *V. cribbiana* and *V. planifolia*. The columns provide the following information: ion mode, retention time (t_R), mass-to-charge ratio (*m/z*), molecule annotation, PubChem Compound Identifier (CID), chemical class, *p*-value (from a *t*-test assessing the significance of differences in molecule abundance between the two species), and descriptors from FlavorDB.

Ion mode	t_R (min)	*m/z*	Annotation	CID	Chemical class	*p*-value	FlavorDB descriptors
ESI–	0.62	131.0473	L-asparagine [a,b]	6267	Carboxylic acids and derivatives	0.00	odorless
ESI–	0.62	101.0254	succinic acid semialdehyde	1112	Fatty Acyls	0.00	
ESI–	0.62	179.0582	D-glucopyranose [a]	5793	Organooxygen compounds	0.00	sweet
ESI–	0.62	341.1111	sucrose [a,b]	5988	Organooxygen compounds	0.00	sweet
ESI–	0.62	161.0465	D-glucose [b]	5793	Organooxygen compounds	0.00	sweet
ESI–	0.63	195.0524	galactonic acid	128,869	Hydroxy acids and derivatives	0.00	sweet
ESI–	0.64	113.0257	norfuraneol	4,564,493	Dihydrofurans	0.00	bready, caramel, cotton, sweet, sugar, candy
ESI+	0.64	360.1537	sucrose [a,b]	5988	Organooxygen compounds	0.00	sweet
ESI–	0.66	143.0360	3,5-dihydroxy-6-methyl-2,3-dihydropyran-4-one [a]	119,838	Pyrans	0.00	
ESI+	0.66	127.0404	methyl 2-furoate [a]	11,902	Furans	0.00	tobacco, fungal, sweet, fruity, mushroom
ESI+	0.66	268.1061	adenosine [b]	60,961	Purine nucleosides	0.33	sweet-like
ESI–	0.67	133.0153	malic acid [a,b]	525	Hydroxy acids and derivatives	0.01	
ESI+	0.67	133.0611	L-asparagine [a,b]	6267	Carboxylic acids and derivatives	0.00	odorless
ESI+	0.68	229.0344	homocitric acid [a]	28,371	Carboxylic acids and derivatives	0.10	
ESI–	0.69	205.0374	homocitric acid [a]	28,371	Carboxylic acids and derivatives	0.05	
ESI–	0.70	203.0201	daucic acid	21,578,281	Hydroxy acids and derivatives	0.00	
ESI+	0.71	148.0609	L-glutamic acid [b]	33,032	Carboxylic acids and derivatives	0.00	yeast, bread, baked, mild
ESI–	0.74	187.0260	*cis*-2-methylaconitate	3,080,625	Carboxylic acids and derivatives	0.05	
ESI–	0.74	143.0357	2,4-dihydroxy-2,5-dimethylfuran-3-one [a]	538,757	Dihydrofurans	0.13	
ESI–	0.74	125.0252	methyl furan-2-carboxylate [a]	11,902	Furans	0.08	tobacco, fungal, sweet, fruity, mushroom
ESI+	0.75	132.1027	L-isoleucine (L-leucine) [a]	6306	Carboxylic acids and derivatives	0.03	odorless, bitter
ESI–	0.97	139.0417	5-methoxybenzene-1,3-diol	71,648	Phenols	0.07	
ESI+	0.99	166.0866	L-phenylalanine	6140	Carboxylic acids and derivatives	0.06	odorless, bitter
ESI+	1.05	220.1189	pantothenic acid [b]	6613	Carboxylic acids and derivatives	0.85	
ESI–	1.06	191.0581	quinic acid [a,b]	6508	Organooxygen compounds	0.03	
ESI+	1.95	137.0611	*m*-anisaldehyde [a]	11,569	Benzene and substituted derivatives	0.03	anisic
ESI–	2.40	205.0737	diethyl tartrate	62,333	Hydroxy acids and derivatives	0.60	mild, must, wine, fruity, earth
ESI+	2.42	139.0400	protocatechuic aldehyde	8768	Organooxygen compounds	0.69	almond, medical, bitter, dry
ESI–	2.44	137.0257	4-hydroxybenzoic acid [a,b]	135	Benzene and substituted derivatives	0.51	nutty, phenolic
ESI–	2.93	175.0615	2-isopropylmalic acid [b]	77	Fatty Acyls	0.00	
ESI+	3.99	169.0499	vanillic acid [a]	8468	Benzene and substituted derivatives	0.04	powdery, vanilla, bean, milky, sweet, creamy, dairy
ESI–	4.35	121.0307	4-hydroxybenzaldehyde [a]	126	Organooxygen compounds	0.01	woody, sweet, balsam, nutty, almond
ESI+	4.38	123.0450	4-hydroxybenzaldehyde [a,b]	126	Organooxygen compounds	0.01	woody, sweet, balsam, nutty, almond
ESI–	4.48	153.0572	vanillyl alcohol [a]	62,348	Phenols	0.03	vanilla, sweet, anise, creamy, balsam, tonka, phenolic
ESI+	4.89	133.0639	cinnamaldehyde [a,b]	637,511	Cinnamaldehydes	0.00	spice, warm, red hots, paint, clove, sweet, candy, cinnamon
ESI–	5.05	163.0417	4-coumaric acid [a,b]	637,542	Cinnamic acids and derivatives	0.00	balsam, balsamic
ESI+	5.15	153.0560	vanillin [a,b]	1183	Phenols	0.91	vanilla, chocolate, sweet, creamy
ESI+	5.16	134.0609	4-hydroxybenzene acetonitrile	26,548	Benzene and substituted derivatives	0.90	

(*continued on next page*)

Table 2 (*continued*)

Ion mode	t_R (min)	*m/z*	Annotation	CID	Chemical class	*p*-value	FlavorDB descriptors
ESI+	5.16	125.0609	guaiacol [a]	460	Phenols	0.92	spice, vanilla, smoky, sweet, woody, phenolic, medicine, medicinal, smoke
ESI–	5.17	151.0411	vanillin [a,b]	1183	Phenols	0.90	vanilla, chocolate, sweet, creamy
ESI+	5.25	165.0539	4-coumaric acid [a,b]	637,542	Cinnamic acids and derivatives	0.00	balsam, balsamic
ESI–	5.30	189.0776	3-hydroxysuberic acid [b]	22,328,017	Hydroxy acids and derivatives	0.00	
ESI+	5.36	107.0502	benzaldehyde [a]	240	Benzene and substituted derivatives	0.01	cherry, almond, sweet, burnt sugar, sharp, strong, bitter
ESI–	5.70	147.0454	4-coumaraldehyde [b]	641,301	Cinnamaldehydes	0.00	
ESI+	5.83	383.1011	7-hydroxy-2-methyl-4-oxo-4H-1-benzopyran-5-carboxylic acid 7-glucoside	73,981,794	Organooxygen compounds	0.04	
ESI–	5.92	329.1039	7-hydroxy-3-(3-hydroxy-4-methoxybenzyl)-5-methoxy-4-chromanone	21,676,258	Homoisoflavonoids	0.00	
ESI–	5.92	165.0569	acetovanillone [a]	2214	Organooxygen compounds	0.00	vanilla, sweet, vanillin, faint
ESI+	5.93	149.0586	cinnamic acid [a,b]	444,539	Cinnamic acids and derivatives	0.00	storax, sweet, honey, balsam
ESI–	5.97	313.1065	7,8-dihydropteroic acid	135,398,662	Pteridines and derivatives	0.00	
ESI–	6.05	287.0584	eriodictyol	440,735	Flavonoids	0.08	
ESI–	6.46	315.0517	isorhamnetin [b]	5,281,654	Flavonoids	0.00	bitter
ESI–	6.59	301.0738	hesperetin [b]	72,281	Flavonoids	0.00	bitter
ESI+	6.72	341.1424	8-prenylnaringenin	480,764	Flavonoids	0.04	
ESI–	6.76	335.0952	atalantoflavone	14,162,621	Flavonoids	0.24	
ESI+	9.12	288.2531	lauroyl diethanolamide [b]	8430	Fatty Acyls	0.23	
ESI–	11.32	295.2286	9,10-epoxy-18-hydroxy-octadecanoic acid	9,818,312	Fatty Acyls	0.12	
ESI+	11.63	295.2273	9(10)-EpODE	16,061,060	Fatty Acyls	0.28	
ESI+	13.71	355.2874	2-linoleoylglycerol	5,365,676	Fatty Acyls	0.03	
ESI+	13.81	279.2298	gamma-linolenic acid	5,280,933	Fatty Acyls	0.10	
ESI+	13.92	263.2373	farnesyl acetone	1,711,945	Prenol lipids	0.12	
ESI+	13.93	355.2804	1-linoleoyl-(2*S*)-glycerol	6,436,630	Fatty Acyls	0.14	
ESI+	14.00	326.3049	n-oleoylethanolamine [b]	5,283,454	Organonitrogen compounds	0.01	
ESI+	14.69	281.2448	linoleic acid [a,b]	5,280,450	Fatty Acyls	0.44	fatty, faint
ESI+	14.98	357.2962	2-oleoylglycerol	5,319,879	Glycerolipids	0.57	fatty, bland, waxy
ESI+	15.54	121.0655	tyramine [b]	5610	Benzene and substituted derivatives	0.59	mild, rubbery, meaty, dirty, cooked, phenolic
ESI+	15.65	283.2639	oleic acid [a]	445,639	Fatty Acyls	0.22	lard, fat, waxy, fatty, fried, faint

[a] Molecules previously annotated in *Vanilla* spp.
[b] Annotations with higher confidence level, based both on experimental and in silico fragmentation matches.

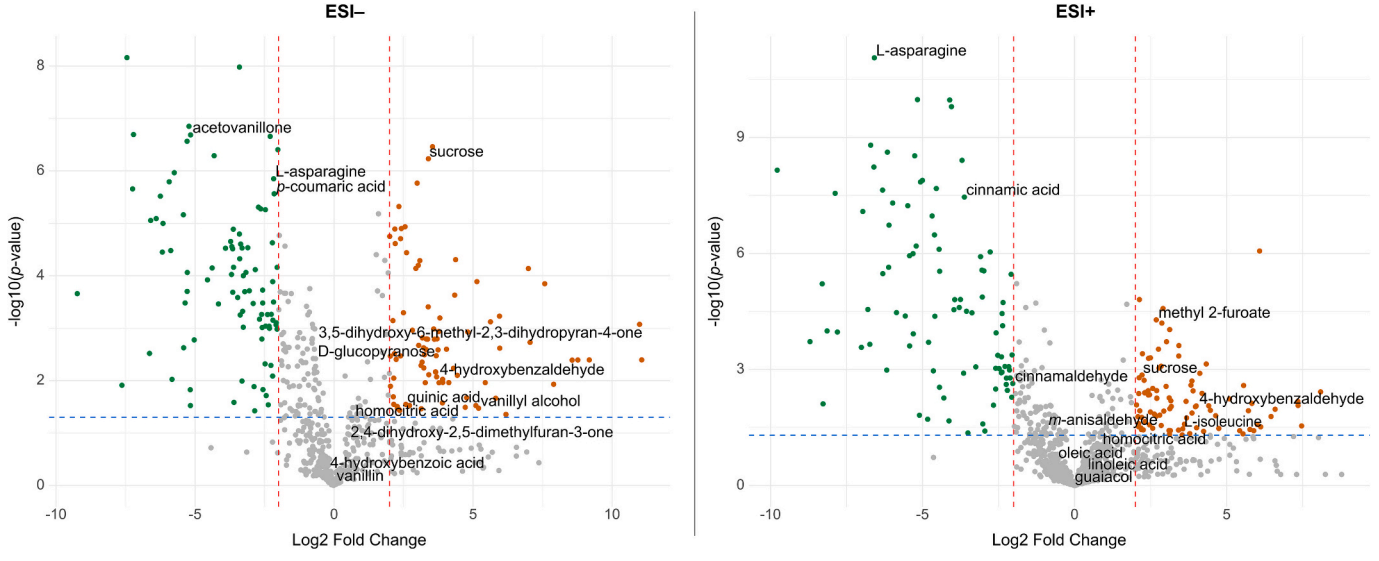


Fig. 1. Differential analysis based on Log2 fold change and t-test's *p*-values of molecules annotated in *V. cribbiana* and *V. planifolia*. Each point represents a different detected molecule. Orange points to the right side of the plots have *p*-value ≤ 0.05, Log fold change ≥ 2, and are more abundant in VP, while green points have the same *p*-value threshold, Log fold change < -2, and are more abundant in VCR. Gray points represent molecules that do not show considerably higher abundance in any species and/or *p*-values >0.05, based on untargeted LC-MS/MS metabolomics analysis data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

However, significant flavor compounds such as vanillin and 4-hydroxybenzoic acid showed similar intensities in both VCR and VP (Fig. 1). This observation is particularly noteworthy since VP's commercial value is primarily attributed to its elevated vanillin content. This preliminary exploratory analysis led to a subsequent target-oriented approach. Six molecules strongly linked to the cherished flavor of commercial vanilla were selected: 4-hydroxybenzoic acid, 4-hydroxybenzaldehyde, acetovanillone, 4-coumaric acid, vanillic acid, and vanillin. This approach aimed to understand the potential aromatic profile of Atlantic Forest VCR and compare it quantitatively with the commercial VP species.

3.2. Targeted analysis focused on the vanilla flavor

The quantitative comparative assessment across VB, VCH, VCR and VP revealed distinct profiles for each compound, with measurements expressed in mg per 100 mg of pods dry weight (Fig. 2). Notably, 4-hydroxybenzaldehyde's levels fell below the limit of quantification (< LOQ) across all samples except in VP (0.20 mg/100 mg dry weight). Conversely, 4-hydroxybenzoic acid demonstrated considerable variation, with notably higher levels measured in VCR (0.18 mg/100 mg dry weight) and VP (0.17 mg/100 mg dry weight) samples, significantly higher than VB (0.03 mg/100 mg dry weight), though not significantly different (p-value >0.05) from VCH (0.10 mg/100 mg dry weight). Furthermore, acetovanillone was < LOQ in the VB and VCH samples, while it was present in VCR (1.83 mg/100 mg dry weight) and most prominently in VP (2.15 mg/100 mg dry weight). In the quantification estimation of vanillic acid, VB and VCH both showed < LOQ levels, while VCR exhibited significantly higher concentration (0.19 mg/100 mg dry weight), followed by VP (0.08 mg/100 mg dry weight). Interestingly, no significant difference in the concentration of vanillin was observed between VCR (2.71 mg/100 mg dry weight) and VP (2.50 mg/100 mg dry weight), while in VB and VCH, this important compound was < LOQ (Fig. 2). The compound 4-coumaric acid was also analyzed in this targeted approach; however, it could not be quantified, as it was found to be below the level of quantification and detection in all species.

Hence, five molecules from the targeted analysis could be successfully measured, except for 4-coumaric acid. In the untargeted approach, acetovanillone was annotated in ESI– mode with m/z 165.0568 (Fig. 1) and in ESI+ mode with m/z 167.0715. Interestingly, the ESI+ mode showed a higher abundance in VP, contrary to the ESI– annotation. However, this feature was omitted from the ESI+ data during the noise baseline subtraction phase, informed by the blank samples data. Subsequent quantification estimation analysis, however, revealed that acetovanillone was more accurately detected in ESI+, showing significantly higher concentrations in VP samples (Fig. 2). Since in the untargeted approach this annotation in the negative mode relied exclusively on in silico fragmentation scores (Table 2), a more thorough targeted analysis of this compound in ESI– is required in future studies.

4. Discussion

In this study, we focused on native, non-cultivated *Vanilla* species from the Atlantic Forest, which presented unique sampling challenges. Our sample size was constrained by the time-intensive nature of manual pollination and the seasonal limitations of working with wild *Vanilla* species in their natural habitats. While these factors limited the number of samples, we believe our approach provides valuable insights into these understudied species, particularly in the context of their flavor profiles and potential commercial applications.

4.1. Characterization of flavor-associated molecules in vanilla wild crops and cultivars

The sensory potential of Brazilian vanilla species has been recently investigated. According to Silva et al. (2023) in a study that applied a Solid Phase Microextraction (SPME) coupled with Gas Chromatography-Mass Spectrometry (GC–MS) approach, *V. bahiana* exhibited only low and medium-intensity aroma-associated volatile compounds including 2,3-butanediol, ethyl octanoate, methyl salicylate, methyl-chavicol, octanoic acid, *p*-anisic alcohol, and vanillin with medium intensities. In *V. chamissonis*, *p*-anisic alcohol was associated with high intensity and *p*-anisaldehyde with medium intensity, while all other compounds had low intensities. All anisic compounds had low intensities in the *V. planifolia* samples (Silva et al., 2023). These findings suggest that these characteristics likely differentiate the sensory profile of these Brazilian species in comparison to the commercial *V. planifolia* and more closely aligns with the sensory profile of the commercial *V.* × *tahitensis*, characterized by strong anise and caramel notes, which are attributed to high levels of anisyl compounds (Brunschwig et al., 2016; Pérez-Silva et al., 2021). Galeana (2024) studied the aromatic profile of *V. cribbiana* cultivated in Coban, Guatemala, and also found vanillin levels comparable to those in *V. planifolia*, with *V. cribbiana* containing 1.68 mg/100

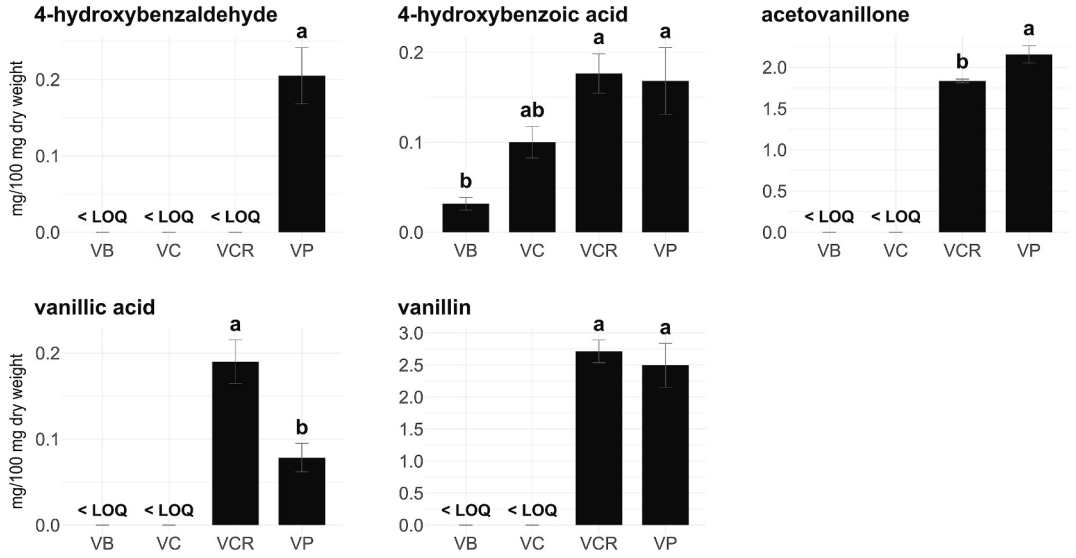


Fig. 2. Targeted analysis for the quantification estimation of characteristic flavoring molecules found in commercial vanilla, annotated in the untargeted approach of this study. The measurement is given in mg per 100 mg of dry weight of vanilla pods. < LOQ – below the limit of quantification. VB – *V. bahiana*; VC – *V. chamissonis*; VCR – *V. cribbiana*; VP – *V. planifolia*. The letters above the error bars resulted from the post-hoc analysis through the Tukey test.

mg of dry weight and *V. planifolia* containing 1.13–1.72 mg/100 mg of dry weight. In contrast to our findings, Galeana's research on Guatemalan *V. cribbiana* showed significantly lower levels of vanillic acid (0.12 mg/100 mg of dry weight) compared to *V. planifolia* (0.13–0.17 mg/100 mg of dry weight). Additionally, *V. cribbiana* exhibited significantly higher levels of 4-hydroxybenzaldehyde (0.33 mg/100 mg of dry weight) compared to *V. planifolia* (0.07–0.1 mg/100 mg of dry weight), a compound that was below the limit of quantification (LOQ) in VCR from the Atlantic Forest, in the present study. Furthermore, *V. cribbiana* showed significantly higher levels of 4-hydroxybenzoic acid (0.45 mg/100 mg of dry weight) compared to *V. planifolia* (0.03–0.04 mg/100 mg of dry weight), whereas the present study found that VCR and *V. planifolia* exhibited statistically similar levels of this compound. It is worth noting that a different instrumental approach was applied (HPLC-DAD) for the analysis of the aromatic compounds *V. cribbiana* and *V. planifolia* from Central America (Galeana, 2024), it has the same quantification intent and capacity as the present study, suggesting that enzymatically cured *V. cribbiana* from the Atlantic Forest has 1.6 times more vanillin content. These differences may be associated with different chemotypes or even with the different curing strategies (Ibarra-Cantún et al., 2018; Pardío et al., 2018; Salazar-Rojas et al., 2012). Depending on the cultivar or chemotype, *V. planifolia* is known to produce varying levels of flavor-associated molecules (Ibarra-Cantún et al., 2018; Salazar-Rojas et al., 2012). Meanwhile, enzymatic treatments applied to green vanilla pods have been reported to increase the levels of these compounds (Pardío et al., 2018).

On the other hand, Franco et al. (2024) explored the sensory characteristics more focused on the taste, flavor, and consumer acceptance of various vanilla species, comparing them to the commercial *V. planifolia* extract. *V. bahiana, V. chamissonis,* and *V. cribbiana* all displayed lower flavor detection thresholds, suggesting more intense flavors compared to the commercial *V. planifolia* extract. Despite this enhanced flavor profile, sensory tests showed no significant differences in the 'vanilla aroma' perception across these species and *V. planifolia*. However, consumer acceptance varied: milk flavored with *V. bahiana* and *V. chamissonis* received moderate overall liking scores, suggesting a moderate acceptance level. In contrast, milk flavored with *V. cribbiana* not only stood out for its strong and distinct vanilla taste but also received notably positive feedback and high scores for its 'sweet taste,' indicating a favorable taste profile among consumers (Franco et al., 2024).

We've demonstrated through targeted analysis that *V. bahiana* and *V. chamissonis* present < LOQ levels of vanillin, 4-hydroxybenzaldehyde, acetovanillone, and vanillic acid. In parallel, untargeted analysis exhibits *m*-anisaldehyde with higher intensity in *V. cribbiana* than in *V. planifolia* (*p*-value 0.03 and − 0.94 log2 fold change). Ultimately, our results confirm the expected levels of flavor-related molecules in *V. cribbiana*, closely aligning them with those found in *V. planifolia*. Additionally, our research suggests that Atlantic Forest *V. cribbiana* might incorporate the sought-after sensory characteristics of *V. planifolia* and the anisic qualities of *V.* × *tahitensis*, effectively merging the most desirable aspects of both species. This anisic component was also observed in the Gatemalan *V. cribbiana* (Galeana, 2024). Nonetheless, more profound elucidation targeting the anisyl compounds in this species from the Atlantic Forest is still needed.

Several other known flavoring molecules were found in higher concentrations in *V. cribbiana* compared to *V. planifolia*. For instance, cinnamaldehyde is associated with the sensory attributes of "spicy" and "sweet aromatic spicy, cinnamic and cassia-like, balsamic" when detected in *V.* × *tahitensis* and Bourbon and Ugandan *V. planifolia* cultivars respectively (Brunschwig et al., 2016; Zhang & Mueller, 2012); cinnamic acid is reportedly associated with the "sweet aromatic, balsamic, somewhat cinnamic like" descriptors in *V. planifolia*, with small and medium odor strength in Bourbon and Ugandan cultivars (Zhang & Mueller, 2012). It is present in this commercial species with an amount approximately two thousand times lower than vanillin (Pérez-Silva et al., 2006). It is worth noting that compounds present in low abundance in vanilla pods can exhibit sensory intensities comparable to those in high abundance. Pérez-Silva et al. (2006) demonstrated that specific compounds, including guaiacol, 4-methylguaiacol, acetovanillone, and vanillyl alcohol, despite their significantly lower concentrations in vanilla beans compared to vanillin, possess equivalent sensory impacts. Notably, methyl salicylate, present at less than 1 ppm, was perceived with an intensity akin to vanillin, found in concentrations over 1000 times greater. This evidence highlights that a compound's sensory effect is not directly proportional to its concentration (Pérez-Silva et al., 2006). Therefore, minor constituents play a crucial role in the overall aroma profile of foods, underscoring the complexity of flavor perception and the necessity of considering both predominant and trace compounds regarding both traditional and novel flavoring ingredients in vanilla.

In *V. cribbiana*, compounds that share similar intensities with those in *V. planifolia* have been annotated, hinting at comparable sensory attributes between the two species. For example, guaiacol in *V. cribbiana* is linked to "aromatic, phenolic, burnt" and "chemical, sweet spicy" descriptors, as identified in *V. planifolia* by Zhang and Mueller (2012) and Pérez-Silva et al. (2006), respectively. Additionally, linoleic acid is suggested to contribute to the characteristic flavor of *V. cribbiana*. While its intensity mirrors that found in *V. planifolia*, its specific sensorial contributions in the commercial species have not yet been documented. This suggests a potential aspect of *V. cribbiana*'s flavor profile that warrants further exploration.

Compared to the flavor descriptor database in FlavorDB (Garg et al., 2018), the compounds identified through a untargeted approach are linked with various descriptors for molecules previously unrecognized in commercial vanilla (Table 2). These compounds emerge as potential sensory markers that might underscore the similarities between *V. cribbiana* and *V. planifolia* and highlight distinctive descriptors. Nevertheless, a targeted analysis is imperative for the structural elucidation of these molecules to verify their association with the putative sensory descriptors conclusively.

5. Conclusion

The discovery that Brazilian *V. cribbiana* produces vanillin at levels similar to commercial *V. planifolia* is a major finding. This finding not only unveils a potential sustainable source of natural vanilla flavor but also highlights the untapped potential of vanilla crop wild relatives. Remarkably, *V. cribbiana* potentially exhibits an amalgamation of desirable traits, encompassing the sought-after sensory profile of *V. planifolia* and the distinctive anisic qualities reminiscent of *V.* × *tahitensis*. Adapted to its native Brazilian climate and free from reported pathogen issues, *V. cribbiana* is a promising candidate for future cultivation endeavors. While *V. bahiana* and *V. chamissonis* exhibited immeasurable levels of vanillin and other essential flavor compounds, *V. cribbiana* exhibited a comprehensive array of molecules associated with the characteristic vanilla aroma, including vanillin, acetovanillone, vanillic acid, and 4-hydroxybenzoic acid. Moreover, compounds like cinnamaldehyde, cinnamic acid, and guaiacol in *V. cribbiana* suggest a unique and potentially superior flavor profile to commercial vanilla. This groundbreaking discovery underscores the imperative of preserving and investigating vanilla crop wild relatives, as they harbor invaluable genetic resources with the potential to revolutionize sustainable vanilla production, ensure flavor diversity, and contribute to the conservation of neotropical biodiversity.

Funding

This work was supported by the National Council for Scientific and Technological Development of Brazil (CNPq) (CNPq/MCTI N° 10/2023, grant number 404700/2023–3); the Foundation for Research of Rio de Janeiro State (FAPERJ) (AFM – E-13/2023 – APQ1, grant number E-26/210.438/2024 and MGBK – E-26/111.482/2014); the Coordination of

Superior Level Staff Improvement (CAPES) and FAPERJ, which granted fellowships (J.P.S·O – previously CAPES 001, grant number 88887.685735/2022–00, and is currently FAPERJ E-26/202.379/2024 (300759) scholar at The Graduate Program in Biological Sciences (Neotropical Biodiversity)—PPGBIO); and the Federal University of the State of Rio de Janeiro (UNIRIO) which granted financial support (Edital INOVA-UNIRIO grant number 01/2021 and 01/2022); Edital PPQ-UNIRIO grant number 03/2022).

CRediT authorship contribution statement

Joana Paula da Silva Oliveira: Writing – original draft, Visualization, Investigation, Formal analysis, Data curation. **Rafael Garrett:** Writing – review & editing, Methodology, Investigation, Formal analysis. **Maria Gabriela Bello Koblitz:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization. **Andrea Furtado Macedo:** Writing – review & editing, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used OpenAi's ChatGPT4 model in order to improve English writing and readability. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Andrea Furtado Macedo, Maria Gabriela Bello Koblitz, Rafael Garret and Joana Paula da Silva Oliveira have patent #PCT/BR2022/050468 pending to Federal University of the State of Rio de Janeiro (UNIRIO) & Federal University of Rio de Janeiro (UFRJ). If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

We want to thank the Botanic Garden Research Institute of Rio de Janeiro, especially Dr. Delfina de Araujo, for donating the fruits of *Vanilla planifolia*.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodchem.2024.141650.

References

Al-Khayri, J. M., Jain, S. M., Johnson, D. V., & (Eds.).. (2019). *Advances in plant breeding strategies: Industrial and Food crops. 6*. Springer International Publishing. https://doi.org/10.1007/978-3-030-23265-8

Banda Ortiz, H., Garza Morales, R., & Vázquez Trujillo, K. V. T. (2021). Supply chain and economic viability of Vanilla planifolia Andrew production: A case study. *Agro Productividad, 14*(1). https://doi.org/10.32854/agrop.v14i1.1795

Baqueiro-Peña, I., & Guerrero-Beltrán, J.Á. (2017). Vanilla (Vanilla planifolia Andr.), its residues and other industrial by-products for recovering high value flavor molecules: A review. *Journal of Applied Research on Medicinal and Aromatic Plants, 6*, 1–9. https://doi.org/10.1016/j.jarmap.2016.10.003

Brunschwig, C., Rochard, S., Pierrat, A., Rouger, A., Senger-Emonnot, P., George, G., & Raharivelomanana, P. (2016). Volatile composition and sensory properties of *Vanilla ×tahitensis* bring new insights for vanilla quality control. *Journal of the Science of Food and Agriculture, 96*(3), 848–858. https://doi.org/10.1002/jsfa.7157

Chambers, A., Cibrián-Jaramillo, A., Karremans, A. P., Moreno Martinez, D., Hernandez-Hernandez, J., Brym, M., … Hu, Y. (2021). Genotyping-by-sequencing diversity analysis of international Vanilla collections uncovers hidden diversity and enables plant improvement. *Plant Science, 311*, Article 111019. https://doi.org/10.1016/j.plantsci.2021.111019

Chen, Z., Song, S., Wei, Z., Fang, J., & Long, J. (2021). Approximating median absolute deviation with bounded error. *Proceedings of the VLDB Endowment, 14*(11), 2114–2126. https://doi.org/10.14778/3476249.3476266

De Jager, L. S., Perfetti, G. A., & Diachenko, G. W. (2007). Determination of coumarin, vanillin, and ethyl vanillin in vanilla extract products: Liquid chromatography mass spectrometry method development and validation studies. *Journal of Chromatography A, 1145*(1–2), 83–88. https://doi.org/10.1016/j.chroma.2007.01.039

De Jager, L. S., Perfetti, G. A., & Diachenko, G. W. (2008). Comparison of headspace-SPME-GC–MS and LC–MS for the detection and quantification of coumarin, vanillin, and ethyl vanillin in vanilla extract products. *Food Chemistry, 107*(4), 1701–1709. https://doi.org/10.1016/j.foodchem.2007.09.070

Djoumbou Feunang, Y., Eisner, R., Knox, C., Chepelev, L., Hastings, J., Owen, G., … Wishart, D. S. (2016). ClassyFire: Automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics, 8*(1), Article 1. https://doi.org/10.1186/s13321-016-0174-y

Food, U. S., & Administration, D. (2023). CFR - Code of Federal Regulations Title 21. October 17 https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/cfrsearch.cfm?fr=169.3.

Franco, J. L., Ayres, E. M. M., De Oliveira, D., Martins, I. B. A., Macedo, A. F., Deliza, R., & Koblitz, M. G. B. (2024). Exploring the potential of the *Vanilla* species from the Brazilian Atlantic Forest: Sensory description and consumer acceptance. *Journal of Sensory Studies, 39*(1), Article e12896. https://doi.org/10.1111/joss.12896

Galeana, A. O. (2024). *Establecimiento de un Protocolo para la Determinación del Perfil de Textura y Evaluación Fisicoquímica y Aromática de Vainilla Beneficiadad Diferentes Especies [Master's, Instituto Tecnológico de Tuxtepec]* (http://51.143.95.221/handle/TecNM/7768).

Gallage, N. J., JÃrgensen, K., Janfelt, C., Nielsen, A. J. Z., Naake, T., Duński, E., Dalsten, L., Grisoni, M., & MÃller, B. L. (2018). The intracellular localization of the vanillin biosynthetic machinery in pods of Vanilla planifolia. Plant and Cell Physiology, 59 (2), 304–318. doi:https://doi.org/10.1093/pcp/pcx185.

Garg, N., Sethupathy, A., Tuwani, R., Nk, R., Dokania, S., Iyer, A., Gupta, A., Agrawal, S., Singh, N., Shukla, S., Kathuria, K., Badhwar, R., Kanji, R., Jain, A., Kaur, A., Nagpal, R., & Bagler, G. (2018). FlavorDB: A database of flavor molecules. *Nucleic Acids Research, 46*(D1), D1210–D1216. https://doi.org/10.1093/nar/gkx957

Gu, F., Chen, Y., Hong, Y., Fang, Y., & Tan, L. (2017). Comparative metabolomics in vanilla pod and vanilla bean revealing the biosynthesis of vanillin during the curing process of vanilla. *AMB Express, 7*(1), 116. https://doi.org/10.1186/s13568-017-0413-2

Ibarra-Cantún, D., Delgado-Alvarado, A., Herrera-Cabrera, B. E., Soto-Hernández, R. M., Salazar-Rojas, V. M., Aguilar, M. I., … Aguilar, M. I. (2018). Effect of the environmental condition of Vanilla planifolia jacks. Ex Andrews cultivation on phytochemical concentration. *Revista Chapingo. Serie Horticultura, 24*(2), 151–165. https://doi.org/10.5154/r.rchsh.2017.08.031

Kadir, N. A., Naher, L., & Sidek, N. (2019). Economical important phytopathogenic diseases in Vanilla planifolia: A review paper. *Journal of Tropical Resources and Sustainable Science (JTRSS), 7*(2), Article 2. https://doi.org/10.47253/jtrss.v7i2.513

Lopes, E. M., Linhares, R. G., De Oliveira Pires, L., Castro, R. N., Souza, G. H. M. F., Koblitz, M. G. B., … Macedo, A. F. (2019). Vanilla bahiana, a contribution from the Atlantic Forest biodiversity for the production of vanilla: A proteomic approach through high-definition nanoLC/MS. *Food Research International, 120*, 148–156. https://doi.org/10.1016/j.foodres.2019.02.038

Malinowska, J. M., & Viant, M. R. (2019). Confidence in metabolite identification dictates the applicability of metabolomics to regulatory toxicology. *Current Opinion in Toxicology, 16*, 32–38. https://doi.org/10.1016/j.cotox.2019.03.006

Market Research Future. (2024). Vanilla market size, growth, forecast, and industry share to 2030. https://www.marketresearchfuture.com/reports/vanilla-market-7806.

Matsuzawa, Y., Tsugawa, H., Nishida, K., Takahashi, M., & Buyantogtokh, B. (2022). MS-DIAL ESI(+/−)-MS/MS from standards+bio+in silico. https://systemsomicslab.github.io/compms/msdial/main.html.

de Oliveira, R. T., da Silva Oliveira, J. P., & Macedo, A. F. (2022). Vanilla beyond Vanilla planifolia and Vanilla × tahitensis: Taxonomy and historical notes, reproductive biology, and metabolites. *Plants, 11*(23), Article 23. https://doi.org/10.3390/plants11233311

Pardío, V. T., Flores, A., López, K. M., Martínez, D. I., Márquez, O., & Waliszewski, K. N. (2018). Effect of endogenous and exogenous enzymatic treatment of green vanilla beans on extraction of vanillin and main aromatic compounds. *Journal of Food Science and Technology, 55*(6), 2059–2067. https://doi.org/10.1007/s13197-018-3120-3

Pérez Silva, A., Gunata, Z., Lepoutre, J.-P., & Odoux, E. (2011). New insight on the genesis and fate of odor-active compounds in vanilla beans (*Vanilla planifolia* G. Jackson) during traditional curing. *Food Research International, 44*(9), 2930–2937. https://doi.org/10.1016/j.foodres.2011.06.048

Pérez-Silva, A., Nicolás-García, M., Petit, T., Dijoux, J. B., de los Ángeles Vivar-Vera, M., Besse, P., & Grisoni, M. (2021). Quantification of the aromatic potential of ripe fruit of Vanilla planifolia (Orchidaceae) and several of its closely and distantly related

species and hybrids. *European Food Research and Technology, 247*(6), 1489–1499. https://doi.org/10.1007/s00217-021-03726-w

Pérez-Silva, A., Odoux, E., Brat, P., Ribeyre, F., Rodriguez-Jimenes, G., Robles-Olvera, V., … Günata, Z. (2006). GC–MS and GC–olfactometry analysis of aroma compounds in a representative organic aroma extract from cured vanilla (Vanilla planifolia G. Jackson) beans. *Food Chemistry, 99*(4), 728–735. https://doi.org/10.1016/j.foodchem.2005.08.050

Ruiz-Terán, F., Perez-Amador, I., & López-Munguia, A. (2001). Enzymatic extraction and transformation of Glucovanillin to vanillin from Vanilla green pods. *Journal of Agricultural and Food Chemistry, 49*(11), 5207–5209. https://doi.org/10.1021/jf010723h

Salazar-Rojas, V. M., Herrera-Cabrera, B. E., Delgado-Alvarado, A., Soto-Hernández, M., Castillo-González, F., & Cobos-Peralta, M. (2012). Chemotypical variation in Vanilla planifolia Jack. (Orchidaceae) from the Puebla-Veracruz Totonacapan region. *Genetic Resources and Crop Evolution, 59*(5), 875–887. https://doi.org/10.1007/s10722-011-9729-y

Silva, F. N. D., Vieira, R. F., Bizzo, H. R., Gama, P. E., Brumano, C. N., Vidigal, M. C. T. R., … Minim, V. P. R. (2023). Chemical characterization and sensory potential of Brazilian vanilla species. *Pesquisa Agropecuária Brasileira, 58*, Article e03308. https://doi.org/10.1590/s1678-3921.pab2023.v58.03308

da Silva Oliveira, J. P., Garrett, R., Bello Koblitz, M. G., & Furtado Macedo, A. (2022). Vanilla flavor: Species from the Atlantic forest as natural alternatives. *Food Chemistry, 375*, Article 131891. https://doi.org/10.1016/j.foodchem.2021.131891

Toth, S. (2018). Volatile compounds in Vanilla. In D. Havkin-Frenkel, & F. C. Belanger (Eds.), *Handbook of Vanilla science and technology* (pp. 285–347). John Wiley & Sons Ltd.. https://doi.org/10.1002/9781119377320.ch17

Tsugawa, H., Kind, T., Nakabayashi, R., Yukihira, D., Tanaka, W., Cajka, T., … Arita, M. (2016). Hydrogen rearrangement rules: Computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Analytical Chemistry, 88*(16), 7946–7958. https://doi.org/10.1021/acs.analchem.6b00770

Tsugawa, H., Nakabayashi, R., Mori, T., Yamada, Y., Takahashi, M., Rai, A., … Saito, K. (2019). A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nature Methods, 16*(4), 295–298. https://doi.org/10.1038/s41592-019-0358-2

Wulff, J. E., & Mitchell, M. W. (2018). A comparison of various normalization methods for LC/MS metabolomics data. *Advances in Bioscience and Biotechnology, 09*(08), 339–351. https://doi.org/10.4236/abb.2018.98022

Yang, H., Barros-Rios, J., Kourteva, G., Rao, X., Chen, F., Shen, H., … Dixon, R. A. (2017). A re-evaluation of the final step of vanillin biosynthesis in the orchid Vanilla planifolia. *Phytochemistry, 139*, 33–46. https://doi.org/10.1016/j.phytochem.2017.04.003

Zhang, S., & Mueller, C. (2012). Comparative analysis of volatiles in traditionally cured Bourbon and Ugandan Vanilla bean (Vanilla planifolia) extracts. *Journal of Agricultural and Food Chemistry, 60*(42), 10433–10444. https://doi.org/10.1021/jf302615s

Zheng, J., Johnson, M., Mandal, R., & Wishart, D. S. (2021). A comprehensive targeted metabolomics assay for crop plant sample analysis. *Metabolites, 11*(5), 303. https://doi.org/10.3390/metabo11050303

Capítulo 3 - Triagem virtual baseada em reutilização de dados metabolômicos indica potenciais alvos moleculares dos metabólitos foliares de *Vanilla planifolia*

O Capítulo 3 desta tese corresponde a um artigo original, a ser submetido a um *Journal* da área de bioinformática, que investiga o uso de estratégias *in silico* para inferência de bioatividade de metabólitos naturais a partir da reutilização sistemática de dados de metabolômica não direcionada previamente publicados. Inserido no eixo integrador da tese, o capítulo aprofunda o conceito de sustentabilidade informacional ao indicar que dados químicos já disponíveis podem ser reinterpretados funcionalmente para a geração de hipóteses biologicamente relevantes, sem a necessidade de experimentação laboratorial adicional, em consonância com princípios de ciência aberta e uso eficiente de recursos científicos. Diferentemente dos capítulos empíricos anteriores, os dados metabolômicos originais gerados no âmbito desta tese não foram reutilizados neste estudo específico devido a restrições associadas à propriedade intelectual e à proteção de ativos estratégicos em desenvolvimento; essa limitação foi incorporada de forma deliberada ao desenho do trabalho como reforço conceitual do potencial analítico de dados abertos independentes. O estudo baseia-se na reutilização de um conjunto curado de dados de LC–MS de metabólitos foliares de *Vanilla planifolia*, espécie amplamente estudada e não nativa, utilizada aqui como modelo para validação do pipeline computacional proposto e para embasar sua aplicação futura a espécies nativas de baunilha. A abordagem metodológica fundamenta-se na seleção criteriosa de metabólitos estruturalmente anotados, representativos de compostos abundantes e quimicamente diversos ao longo do desenvolvimento foliar, seguidos por uma triagem racional baseada em *docking* molecular reverso para análise de potenciais de interações molécula–alvo. Adicionalmente, propriedades físico-químicas, farmacocinéticas e toxicológicas foram avaliadas por meio de triagem ADMET *in silico*, permitindo refinar as predições e evitar interpretações superestimadas baseadas exclusivamente em afinidade de ligação. Um elemento central do capítulo é o trabalho de readaptação, atualização e reestruturação do pipeline LigTMap, originalmente desenvolvido em ambientes computacionais obsoletos, para torná-lo compatível com versões atuais de Python e bibliotecas amplamente utilizadas. Essa atualização assegura maior reprodutibilidade, transparência e acessibilidade do fluxo analítico, ampliando o acesso a ferramentas avançadas de predição de alvos moleculares e reduzindo barreiras técnicas à incorporação de abordagens *in silico* em estudos de bioprospecção baseados em metabolômica. Os resultados demonstram que metabólitos foliares de *V. planifolia* apresentam perfis de

interação distintos ao longo do desenvolvimento, com mudanças consistentes nas classes de alvos moleculares preditos, refletindo a dinâmica metabólica vegetal. Compostos não associados ao perfil sensorial clássico da baunilha exibiram afinidade predita por alvos biologicamente relevantes, enquanto a triagem ADMET permitiu discriminar candidatos com baixa plausibilidade de exposição sistêmica. Essa integração revelou que a ontogenia foliar influencia diretamente o espaço funcional potencial acessível a partir de dados metabolômicos, reforçando a importância de considerar o contexto biológico dos dados na inferência de bioatividade. Do ponto de vista conceitual, o capítulo demonstra como o reuso estruturado de dados metabolômicos pode transformar inventários químicos descritivos em hipóteses funcionalmente orientadas, capazes de sustentar decisões estratégicas em pesquisa, desenvolvimento e inovação. Ao evidenciar o valor potencial de metabólitos de baunilha para além de seu papel sensorial tradicional, este estudo fornece a base científica necessária para as etapas subsequentes da tese, incluindo a integração com dados ecológicos, ambientais e espaciais voltados à valorização sustentável da biodiversidade. Assim, o Capítulo 3 consolida o papel das abordagens *in silico* como componentes centrais de pipelines modernos de bioprospecção, demonstrando que a reutilização inteligente de dados químicos é não apenas viável, mas estratégica para acelerar a geração de conhecimento funcional e reduzir redundâncias experimentais.

Title Page

Virtual Screening Strategy Based on Untargeted Metabolomics Data Reuse Indicated Potential Molecular Targets of *Vanilla planifolia* Leaf Metabolites

Joana Paula da Silva Oliveira[1], Murilo Lamim Bello[2]

[1]Institute of Biosciences, Federal University of the State of Rio de Janeiro (UNIRIO), Rio de Janeiro, RJ, Brazil.

[2]Pharmaceutical Planning and Computer Simulation Laboratory, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, RJ, Brazil.

Abstract

Untargeted metabolomics studies of *Vanilla planifolia* leaves have revealed extensive chemical diversity beyond vanillin, yet most annotated metabolites remain functionally unexplored. This study reused openly available LC–MS metabolomics data from *V. planifolia* leaves across developmental stages to evaluate the bioactivity potential of abundant metabolites using an integrated *in silico* workflow. Structurally annotated compounds were subjected to reverse docking using an adapted LigTMap pipeline, followed by ADMET-oriented screening with ADMETlab 3.0. The analyses revealed stage-dependent shifts in predicted target profiles, with early leaf stages dominated by hydrolase and kinase interactions and late stages showing enrichment toward methyltransferase-related targets. Several non-vanillin metabolites exhibited strong predicted binding to biologically relevant protein classes, while ADMET constraints substantially refined the set of candidates compatible with realistic exposure scenarios. These results demonstrate that metabolomics data reuse, when coupled with complementary computational layers, can transform descriptive plant metabolite inventories into biologically informed hypotheses, offering a scalable strategy to prioritize natural products for future experimental validation.

Keywords

Vanilla planifolia; leaf metabolomics; reverse docking; ADMET modeling; natural product bioactivity

1. Introduction

Vanilla spp. are globally recognized for their aromatic and economic value, with the global vanilla market projected to reach USD 1.9 billion by 2030, driven by widespread applications in food, fragrance, and pharmaceutical industries (Market Research Future, 2024; Banda Ortiz et al., 2021), yet their chemical complexity extends far beyond vanillin and related phenylpropanoids traditionally associated with cured pods (Havkin-Frenkel & Belanger, 2018). Commercially dominated by *Vanilla planifolia*, which underpins the majority of global supply predominantly sourced from Madagascar, the genus faces persistent sustainability challenges including disease pressure, monoculture dependency, and price instability, reinforcing the relevance of expanding knowledge on vanilla beyond its fruit chemistry (Al-Khayri et al., 2019; Market Research Future, 2024). Over the last decade, metabolomics approaches have substantially expanded the catalog of secondary metabolites reported in *Vanilla* species, revealing diverse classes of phenolics, alkaloids, terpenoids, and other specialized metabolites (Palama et al., 2010; Pérez-Silva et al., 2021). However, the majority of vanilla metabolomics studies remain strongly focused on fruits and curing processes, reflecting industrial priorities rather than whole-plant chemical ecology (Gu et al., 2017; Palama et al., 2010). In contrast, vegetative organs such as leaves have received comparatively limited attention, despite their central roles in plant defense, development, and environmental interactions (Hernández-Peña et al., 2025). Available evidence, however, points to a biologically meaningful chemical matrix in vanilla vegetative tissues: leaf metabolomic profiles in *V. planifolia* have been shown to shift under biotic stress conditions, including viral infection, showing responsiveness of the vegetative metabolome to environmental challenges (Palama et al., 2010); leaf-based metabolite signatures have been used to discriminate *Vanilla* species, reinforcing leaves as a biologically informative matrix (Leyva et al., 2021); and phenylpropanoid-related networks detected in leaves have been linked to vanillin biosynthesis pathways, suggesting functional continuity between vegetative and reproductive organ chemistry (Hernández-Peña et al., 2025). Together, these findings indicate that vanilla leaves harbor structurally diverse metabolites with plausible roles in plant defense and stress responses, representing an underexplored source of potential bioactive compounds beyond their sensory relevance. Recent high-resolution metabolomics analyses of *V. planifolia* vegetative tissues have showed pronounced developmental and organ-specific variation in metabolite composition, substantially increasing the number of structurally annotated compounds available for downstream analyses (Hernández-Peña et al., 2025). Despite this growing body of openly accessible metabolomics data, most reported studies remain restricted to qualitative or semi-quantitative annotation, with limited exploration of the potential biological activities or molecular targets associated with the detected compounds (Palama et al., 2010; Pérez-Silva et al., 2006). The reuse of curated metabolomics datasets represents an increasingly important strategy in natural products research, enabling new biological inferences without additional experimental burden and aligning with broader principles of open science and data sustainability (Mons et al., 2017; Wilkinson et al., 2016). Within this context, vanilla metabolomics datasets derived from vegetative tissues constitute an underexploited resource for hypothesis generation regarding the bioactivity and systemic relevance of plant-derived small molecules beyond their traditional sensory roles (Hernández-Peña et al., 2025).

The rapid expansion of plant metabolomics has generated large numbers of structurally annotated small molecules, yet experimental characterization of their molecular targets and systemic relevance remains impractical at scale due to cost, time, and ethical constraints (Atanasov et al., 2021). *In silico* target prediction methods, particularly reverse docking approaches, have therefore emerged as valuable hypothesis-generating tools to infer putative protein–ligand interactions starting from chemical structure rather than predefined biological assays (Rognan, 2010; Shaikh et al., 2021). Reverse docking workflows enable the systematic screening of a single compound against multiple protein targets, offering a ligand-centric perspective that is well suited to exploratory analyses of chemically diverse natural product datasets (Cereto-Massagué et al., 2015). Among available platforms, LigTMap implements a hybrid strategy that combines ligand similarity search, structure-based docking, and binding interaction fingerprint analysis, achieving competitive performance across multiple target classes while maintaining broad applicability to structurally heterogeneous

compounds (Shaikh et al., 2021). Despite their utility, reverse docking predictions are inherently limited by simplified scoring functions, incomplete structural coverage of protein space, and the absence of pharmacokinetic context, which restricts direct biological interpretation of predicted interactions (Luis et al., 2020).

ADMET modeling approaches provide a complementary layer of inference by estimating absorption, distribution, metabolism, excretion, and toxicity properties that constrain whether predicted molecular interactions are likely to be biologically accessible in vivo (Jones and Rowland-Yeo, 2013; Fu et al., 2024). *In silico* ADMET platforms enable rapid, standardized evaluation of physicochemical and pharmacokinetic parameters directly from molecular structure, facilitating the prioritization of compounds with plausible systemic exposure profiles (Daina et al., 2017; Fu et al., 2024). ADMETlab 3.0 integrates multi-task deep learning architectures with curated experimental data to predict a broad range of pharmacokinetic and toxicity endpoints, offering improved coverage, uncertainty estimation, and scalability for batch analyses of large compound sets (Fu et al., 2024). However, ADMET predictions alone do not provide mechanistic insight into molecular targets or modes of action, limiting their interpretive power when applied in isolation (Jones & Rowland-Yeo, 2013).

When combined, reverse docking and ADMET modeling constitute a complementary computational framework in which structural interaction plausibility and systemic exposure constraints jointly inform the prioritization of candidate bioactive molecules (Cereto-Massagué et al., 2015; Ekins et al., 2007). This integrative strategy is particularly relevant for natural product research, where chemical diversity is high, experimental validation is resource-intensive, and hypothesis generation must precede targeted biological testing (Atanasov et al., 2021). Applied to curated plant metabolomics datasets, such workflows enable the reinterpretation of annotated metabolites not merely as chemical features, but as candidates with differential likelihoods of engaging biologically meaningful targets within realistic pharmacokinetic boundaries (Fu et al., 2024; Luis et al., 2020).

Despite the growing availability of annotated plant metabolomics datasets, the systematic reuse of such data to infer molecular targets and pharmacokinetic constraints within a unified computational framework remains uncommon, particularly for complex plant-derived metabolite sets (Atanasov et al., 2021; Wilkinson et al., 2016). The central hypothesis of this study is that developmentally regulated metabolites detected in *Vanilla planifolia* leaves exhibit distinct, target-specific interaction profiles that can be computationally prioritized when reverse docking predictions are contextualized by ADMET-related constraints (Fu et al., 2024; Shaikh et al., 2021). Based on this premise, the primary objective is to apply an integrated reverse-docking and ADMET informed workflow to structurally annotated metabolites previously reported for *V. planifolia* leaves across developmental stages, enabling the prioritization of compounds with both plausible molecular interactions and feasible systemic exposure profiles (Hernández-Peña et al., 2025; Jones & Rowland-Yeo, 2013). The selection of *V. planifolia* was dictated by data availability: Hernández-Peña et al. (2025) provide both SMILES-resolved structural annotations and semi-quantitative abundance estimates across developmental stages — the two inputs required by the pipeline. Datasets from other species currently lack this paired structure-abundance resolution for vegetative organs, but the workflow is directly replicable as equivalent datasets become available. By coupling open metabolomics data reuse with complementary *in silico* inference layers, this work aims to demonstrate a scalable strategy for generating biologically meaningful hypotheses from plant metabolomes without additional experimental input (Wilkinson et al., 2016; Mons et al., 2017). The central innovation of this approach is the systematic repurposing of an existing, openly accessible dataset to generate functional hypotheses — transforming abundance-resolved SMILES annotations into ranked candidates with predicted molecular interactions and pharmacokinetic plausibility. Rather than treating metabolomics data as a static chemical inventory, this workflow operationalizes it as a discovery resource, establishing a transferable methodological precedent for data-driven natural product prioritization within and beyond the genus Vanilla.

2. Materials and Methods

Two complementary *in silico* methodologies were employed to evaluate the intermolecular interactions and ADMET properties of the compounds investigated. First, a reverse-docking strategy based on an adapted Python implementation of the LigTMap (Shaikh et al., 2021) workflow was used to indicate potential protein targets, ligand–protein interaction patterns, and binding affinities. Second, physiochemical, pharmacokinetic, and toxicity profiles were obtained by high-throughput virtual screening using ADMETlab 3.0 (Fu et al., 2024). Together, these approaches provide a combined molecular structural and pharmacokinetic assessment to support compound prioritization.

The metabolite data analyzed in this study were obtained from the integrative metabolomic investigation conducted by Hernández-Peña et al. (2025), which applied untargeted LC–MS to characterize the chemical composition of *Vanilla planifolia* leaves across developmental stages. Leaf tissues were analyzed to capture vegetative metabolic profiles associated with phenylpropanoid-related pathways, complementing fruit-centered studies of vanilla chemistry (Hernández-Peña et al., 2025). LC–MS analysis enabled the annotation of structurally diverse metabolites, including phenylpropanoids, benzenoids, organic acids, and glycosylated compounds, generating a curated dataset suitable for computational bioactivity inference (Hernández-Peña et al., 2025). The annotated leaf metabolite list and corresponding structural information derived from this dataset were used as the exclusive source of compounds for reverse-docking and pharmacokinetic analyses in the present study (Hernández-Peña et al., 2025).

2.1. Relationship to the original LigTMap workflow

The workflow implemented here is derived from the open-source LigTMap target-prediction method (Shaikh et al., 2021). LigTMap is a hybrid ligand- and structure-based workflow consisting of five sequential steps: (1) a ligand similarity search to shortlist potential targets, (2) docking of the query ligand into the shortlisted targets, (3) calculation of a binding similarity between the docked pose and the co-crystallized ligand, (4) prediction of class-specific binding activity using machine-learning models, and (5) consolidation of the results and ranking of targets. The developers of LigTMap showed that combining multiple fingerprints (Morgan, MACCS, and Daylight) for the ligand similarity search improves performance and adopted an average Tanimoto cutoff of 0.4. The final LigTMap score was defined as a weighted sum of the ligand fingerprint score and the binding fingerprint score using weights of 0.7 and 0.3 respectively (Shaikh et al., 2021).

2.2. Adaptations for Python 3 and additional modifications

The original LigTMap command-line script was written for Python 2.7 and relied on the PyChem library for fingerprint generation and on several older dependencies. To make the workflow compatible with modern environments and reproducible without proprietary tools, we implemented the following adaptations and enhancements:

1. Python 3 migration. All print statements were converted to function calls, and file opening modes were adjusted to distinguish between binary and text data. The code now adheres to Python 3 syntax and ensures Unicode-safe string handling.

2. Fingerprint generation with RDKit. In the absence of PyChem, which is no longer actively maintained, we provide a lightweight shim that implements the necessary fingerprint functions (CalculateMorganFingerprint, CalculateMACCSFingerprint, CalculateDaylightFingerprint and CalculateSimilarity) using RDKit. This change preserves the original behaviour while eliminating a dependency. Morgan fingerprints of radius 2, MACCS keys, and RDKit Daylight fingerprints are

computed exactly as in the original method, and similarities are measured using RDKit's Tanimoto implementation (rdkit.org).

3. Environment-controlled thresholds. The similarity cutoff and the number of docked candidates are now configurable via environment variables (LIGTMAP_TANIMOTO_CUTOFF and LIGTMAP_FALLBACK_TOPN). The default cutoff in this adaptation is higher (0.85) to reduce the number of false positives, although the original LigTMap paper recommends a cutoff of 0.4–0.5 (Shaikh *et al.,* 2021); users can revert to the original threshold by setting LIGTMAP_TANIMOTO_CUTOFF=0.4.

4. Fallback for 3D structure generation. Open Babel's --gen3d option is still used to generate initial 3D coordinates (openbabel.org), but if Open Babel fails, RDKit now generates a 3D conformer, adds hydrogens and optimizes geometry using the Universal Force Field. This fallback ensures robustness across platforms.

5. Error handling and reporting. The script now logs missing fingerprints and docking errors to separate files to aid troubleshooting. When no candidate passes the similarity threshold, placeholder output files are produced so that downstream processes remain consistent.

6. Standardized output and scaling. Predicted activities are generated using pre-trained scikit-learn models for each target class. Features are normalized to zero mean and unit variance using StandardScaler (scikit-learn.org). Output files (output.csv and IFP_result.csv) include compound identifiers, names, SMILES strings and predicted affinities in addition to the fields provided in the original implementation.

These modifications retain the methodological essence of LigTMap while enabling execution on contemporary Python 3 environments and providing additional control over thresholds and outputs. Users should still cite the original LigTMap publication when using this adapted workflow in scientific work (Shaikh *et al.,* 2021). The adapted pipeline is publicly available as a fork of the original LigTMap repository at https://github.com/joanapaulaso/LigTMap_JPSO/, enabling reproducibility and further reuse by the community.

2.3. Materials

Software and hardware. All analyses were performed in Python (v3.8) on a Linux workstation. Ligand representations were handled with RDKit (release 2023.09) and the Open Drug Discovery Toolkit (ODDT v0.7.2). Conformers were generated and PDBQT files were prepared using Open Babel (v3.1.1) and MGLTools (v1.5.7). Machine-learning models were implemented with scikit-learn (v1.2.2). Data manipulation used NumPy (v1.23), SciPy (v1.10) and pandas (v1.5). Unless stated otherwise, default parameters from the respective packages were used.

Input molecules. The pipeline accepts an input file containing a single SMILES string for the ligand of interest. The first non-empty line is parsed to extract the SMILES; fields separated by tabs or commas are tolerated. If a compound identifier or name is provided it is propagated to the output; otherwise, the file name is used.

Reference databases. For each target class, pre-computed indices were organized into a root directory containing (i) a CSV file mapping Protein Data Bank identifiers (PDB IDs) to canonical SMILES and experimental affinities, (ii) bitstring files for three fingerprint types (Morgan, MACCS and Daylight) for every PDB ligand, (iii) a docked structure of each protein in PDB format and (iv) pre-computed interaction fingerprints. A previously trained regression model in pickle format (Model/<class>.sav) stored the

relationship between EState fingerprints and binding affinity for each target class. Separate directories held the docking configuration and shell scripts used to perform the docking.

2.4. Methods

2.4.1. Ligand similarity search

For the input SMILES, an RDKit molecule was generated and its three fingerprints were computed. A circular Morgan fingerprint of radius 2 (equivalent to ECFP4) was generated as a sparse bit vector (rdkit.org). Public MACCS keys were calculated using RDKit's 166-bit implementation; these keys ask whether predefined substructures are present and produce a binary fingerprint (janiceto.github.io). An RDKit Daylight fingerprint (2048-bit hashed path-based fingerprint) was also generated. For each fingerprint, similarity to every pre-computed fingerprint in the selected database was measured using the Tanimoto index. The RDKit implementation defines the Tanimoto similarity between two bit vectors as the ratio of the number of common bits to the total number of bits set in either fingerprint (rdkit.org). When more than one fingerprint type was used, the three similarity values were averaged to form a ligand similarity score; otherwise the mean of Morgan and MACCS similarities was used. Candidates with an average similarity below 0.85 were discarded, and only the top 20 highest-scoring ligands were retained to limit the number of docking calculations. Although the original LigTMap validation established 0.4 as the empirically optimal cutoff for the MMD combination, balancing target recovery against false-positive inclusion, a more conservative threshold was applied here to prioritize interaction specificity over coverage, given that the goal was hypothesis generation rather than exhaustive target fishing. This choice reduces the risk of propagating low-confidence structural matches into the docking and ADMET steps, at the cost of potentially excluding structurally dissimilar but biologically relevant targets.

2.4.2. Molecular docking

The retained ligands were docked against their respective protein structures. The input SMILES was first converted to a three-dimensional structure using Open Babel with the --gen3d option, which automatically generates 3D coordinates when converting SMILES to PDB (openbabel.org). If Open Babel failed, RDKit was used as a fallback: hydrogens were added, a conformer was embedded using a distance-geometry algorithm, and the geometry was optimized with the Universal Force Field. The resulting PDB file was converted to the PDBQT format required by AutoDock using the prepare_ligand4.py script from MGLTools. This script adds hydrogens, assigns Gasteiger partial charges and removes unnecessary atoms as recommended for AutoDock4 docking (sebastianraschka.com). Docking was performed by an external shell script (dock.sh) that calls the AutoDock/Vina back end for the corresponding target; ten runs were performed using particle-swarm optimization parameters defined in the script (exhaustiveness = 1 and number of poses = 1). The search space was not blind: for each PDB target, a precomputed config file specified the grid center and box dimensions, constraining sampling to the known binding cavity rather than scanning the entire protein surface. Docking scores were extracted from the log file produced by the script.

2.4.3. Activity prediction with EState fingerprints

To estimate the binding affinity independent of docking, an electrotopological state (EState) fingerprint was computed for each docked ligand using RDKit's FingerprintMol function. The EState fingerprint encodes topological and electronic information by combining atomic contributions into a 79-bit vector and is commonly used to describe molecular properties. Fingerprint bits were converted into NumPy arrays; if explicit bit vectors were returned, they were converted to integer arrays by interpreting the bitstring as a sequence of 0/1 values. A pre-trained regression model specific for the target class, deserialized from a pickle file, was used to predict the pKd (negative logarithm of the dissociation constant). The model was trained on molecules from the same database with affinities measured experimentally and was applied to the test set without further tuning.

Features were scaled to zero mean and unit variance using scikit-learn's StandardScaler; this normalization shrinks the range of each feature to improve numerical stability (scikit-learn.org).

2.4.4. Binding similarity search

For each docked complex, the binding mode was compared to the crystallographic complex stored in the reference database using interaction fingerprints. Protein–ligand interaction fingerprints were calculated with ODDT's InteractionFingerprint method, which encodes eight types of interactions per residue, including hydrophobic contacts, aromatic stacking, hydrogen bonds, salt bridges and metal coordination (oddt.readthedocs.io). The fingerprint produced a binary array with a fixed length equal to the number of residues times eight; pre-computed fingerprints were read from binary files. Binding similarity was computed using a Tanimoto coefficient tailored to binary arrays; it measures the intersection of bits divided by the union. LigTMap scores were derived by combining the ligand similarity score (LigandScore) with the binding fingerprint similarity (BindingSimilarityScore) using a weighted sum: $0.7 \times \text{LigandScore} + 0.3 \times \text{BindingSimilarityScore}$.

2.4.5. Output files

The pipeline produced two CSV files. output.csv contains the list of candidates PDB IDs with their ligand similarity scores, docking scores, and predicted affinities. IFP_result.csv contains the same information together with the computed binding similarity and the final LigTMap score. Results are sorted by the LigTMap score to prioritize targets with high ligand and binding similarity as well as favorable docking and predicted affinity.

2.4.6. PyMol images rendering

The interaction analysis pipeline employed distance cutoffs consistent with structural surveys and interaction analyses of protein–ligand complexes. Hydrogen bonds were defined using a 3.3 Å donor–acceptor distance cutoff, which lies within the 2.5–3.5 Å length scale for protein hydrogen bonds and close to the typical clustering around ~2.9 Å with donor–H–acceptor angles ≥100° reported for backbone and side-chain hydrogen bonds (Tan et al., 2020). Hydrophobic contacts were identified when a ligand carbon (or halogen) atom was within 4.0 Å of a protein carbon or sulfur atom, following the geometric definition used in the published PDB-wide analysis of protein–ligand atomic interactions (Ferreira De Freitas & Schapira, 2017). π-stacking (aromatic) contacts were assigned using a 4.8 Å cutoff between aromatic ring centroids, consistent with the 4.5–7.0 Å centroid–centroid distance range described for aromatic π-networks and stacking motifs in protein environments (Lanzarotti et al., 2011). Together, these criteria provide a literature-based operational definition of hydrogen bonds, hydrophobic contacts, and π-stacking interactions in the analyzed protein–ligand complexes.

2.5. Statistical considerations

The Tanimoto index was used for all similarity calculations because it is widely used to compare binary fingerprints and scales from 0 (no similarity) to 1 (identical). Weighted averaging of scores emphasized ligand similarity (70 %) while still considering binding similarity (30 %), reflecting the assumption that similar ligands generally bind to the same protein. The fallback threshold for the number of docked complexes (20) ensured computational feasibility. Affinity predictions were based on regression models trained on hundreds to thousands of ligands per target class; however, the accuracy of these predictions depends on the representativeness of the training data and should be interpreted qualitatively.

2. 6. ADMET Screening (ADMETlab 3.0)

Physicochemical, pharmacokinetic, and toxicological profiles of the candidate molecules were predicted using ADMETlab 3.0 (https://admetlab3.scbdd.com), an integrated ADMET evaluation platform developed by the Center of Drug Design and Discovery (SCBDD) (Fu et al., 2024). The Batch Screening module (https://admetlab3.scbdd.com/server/screening) was used to process a list of compounds provided as SMILES strings, enabling automated high-throughput virtual evaluation across all models available in the service. The server performs standardized cheminformatics preprocessing, including structure sanitization, validation, and canonicalization before model execution. ADMETlab 3.0 integrates multiple machine learning models trained on curated datasets to generate predictions for a wide range of drug-likeness, physicochemical, absorption, distribution, metabolism, excretion, and toxicity (ADMET) endpoints. Predicted categories include, but are not limited to: Lipinski and Veber drug-likeness rules, oral bioavailability, water solubility, intestinal absorption, P-gp substrate probability, skin permeability, plasma protein binding, volume of distribution, BBB penetration, CYP450 isoform inhibition or substrate likelihood, renal and hepatic clearance parameters, acute toxicity, hERG cardiotoxicity risk, and environmental toxicity indices (Fu et al., 2024). For each submitted molecule, the platform generates a complete ADMET profile with numerical predictions, categorical labels (e.g., high/low risk or optimal/suboptimal), and model-specific confidence metrics where available. The Batch Screening tool returns results in tabular format, which were downloaded directly from the server and used without further modification. All predictions correspond to ADMETlab 3.0 model versions and parameters available on the platform at the time of analysis (Fu et al., 2024).

3. Results

3.1. Overview of the compound panel and data provenance

Virtual target profiling and docking were carried out for ten small molecules extracted from *Vanilla planifolia* leaf extracts as published by Hernández-Peña and collaborators (Hernández-Peña et al., 2025). Throughout the following section each compound is referred to by its PubChem compound identifier (CID) for brevity. The CIDs and their corresponding compound names are listed in (Table 1 and Figure 1). A multi-step workflow implemented in ligtmap.py performed (i) ligand-similarity searches using multiple RDKit fingerprints, (ii) docking of the top similarity hits, (iii) machine-learning–based affinity prediction and (iv) interaction-fingerprint (IFP) comparisons between docked and crystallographic complexes. An auxiliary script, compile_ifp.py, concatenated per-compound IFP results into global rankings by LigTMapScore and by DockingScore and generated summary plots. The LigTMapScore is defined as a weighted combination of ligand-based similarity and binding-fingerprint similarity; when both components are available, the score is computed as 0.7 × LigandSimilarityScore + 0.3 × BindingSimilarityScore.

Table 1. Composition of the virtual screening panel. Each compound is identified by its PubChem CID; these identifiers are used in the remainder of the Results section for conciseness.

CID	Common name
5280459	Quercitrin
12005192	2-[6,7-dimethoxy-3-(3-methoxyphenyl)-2,4-dioxo-1-quinazolinyl]-N-(5-methyl-1H-pyrazol-3-yl)acetamide
12308704	Quercetin 3-arabinoside

45359684	5-{4-[4-(hexopyranosyloxy)-3,5-dimethoxyphenyl]tetrahydro-1H,3H-furo[3,4-c]furan-1-yl}-2-methoxyphenyl hexopyranoside
13964506	Ptelatoside A
10073778	Scroside D
5488152	3,5,8-trihydroxy-3',4',7-trimethoxyflavone
737375	Benzylthiourea (Carpasemine)
12241336	*o*-Cresol, dihydrogen phosphate
70678861	S-adenosyl-L-homocysteinate

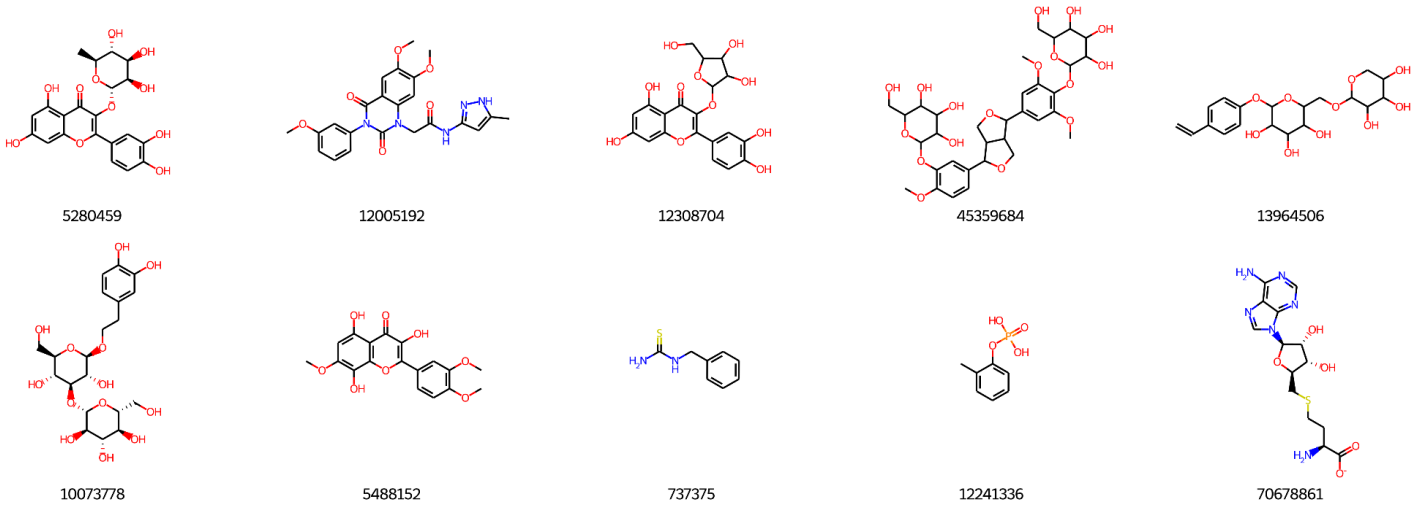


Figure 1. Two-dimensional depictions of the ten study ligands (PubChem CIDs: 5280459, 12005192, 12308704, 45359684, 13964506, 10073778, 5488152, 737375, 12241336, 70678861) drawn from canonical SMILES. Bond aromatization follows RDKit defaults; structures were geometry-optimized in 2D for clarity.

The selected ligands sit in the upper tail of each stage's abundance distribution (dataset tops: 1-5T20=1836, 2T20=2436, 4T20=1662, 6T20=1718, 8T20=1775; ~32–38 compounds per stage exceed 60 of relative abundance). The two strongest ligands (45359684, 5488152) are consistently top-5 per stage (≈22–30% of the stage maxima) (Figure 2). Mid-pack ligands (737375, 12241336, 12005192, 12308704) rank ~#6–12 across stages, still well above the 60 cutoff. The remaining three (5280459, 10073778, 70678861, 13964506) stay above the filter but drop to ranks ~#15–32 (among 727 compound annotations) depending on stage (Figure 2). Since all chosen ligands are among the most abundant at each stage, observed signals are unlikely to be driven by rare, noisy detections; stage-specific differences for the mid/low end of this set (e.g., 5280459 dips at 2T20) should be viewed with slightly more caution, but any strong effects remain biologically meaningful given their high baseline abundance.

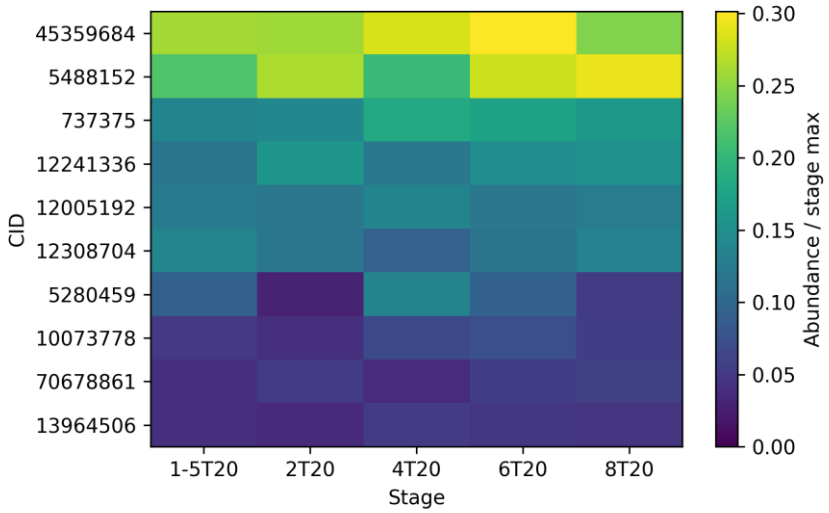


Figure 2. Abundance heatmap of selected ligands across developmental stages (1-5T20, 2T20, 4T20, 6T20, 8T20), showing each CID's abundance normalized to the stage-specific maximum.

3.2. Global composition of the prioritized hit sets

The consolidated docking-selected set (top10_docking_combined.csv) contained six distinct CIDs (5280459, 12005192, 12308704, 45359684, 10073778 and 13964506), whereas the LigTMap Score-selected set (top10_ligtmap_combined.csv) contained six different CIDs (5280459, 12308704, 5488152, 737375, 12241336 and 70678861). Only CIDs 5280459 and 12308704 were shared between these two selections, underscoring that energy-based docking and similarity/IFP-based scoring capture complementary regions of chemical space. No missing Docking Score values were observed in either data set, confirming that all docked complexes were scored successfully.

3.3. Stage-resolved docking results

The stage column in the aggregated tables describes the phenological stage of the leaves from which the compounds were isolated (1-5T20, 2T20, 4T20, 6T20 and 8T20), making it important to assess whether predicted targets shift as the plant matures (Table 2 and Figure 1). Across stages, the compound–target pairs with the lowest Docking Scores range narrowly from −11.16 to −10.75, indicating consistently tight predicted binding. In the early and intermediate stages (1-5T20 through 6T20), CID 5280459 bound most strongly to the hydrolase cGMP-specific 3′,5′-cyclic phosphodiesterase (PDB 2h44), with Docking Scores between −11.16 and −11.11. PyMOL analysis of this complex revealed a densely packed pocket comprising 19 residues within 4.0 Å, generating 140 atom–atom contacts dominated by Phe820, Leu765, Gln817, Tyr612 and Ile665, each contributing over ten interactions. Eleven hydrogen-bond candidates were identified (≤3.5 Å), involving Gln817, Tyr612, Ser663, Ile665, Ser668 and Asp764, with Tyr612 and Ser668 recurrently acting as both donor and acceptor partners, consistent with a stabilizing H-bond network that anchors the glycoside within the catalytic groove. The complex also displayed 35 hydrophobic C–C contacts, supporting interactions at the bottom of the cavity. In contrast, at the latest stage (8T20) CID 12308704 supplanted 5280459 as the top docking hit against the same phosphodiesterase (Docking Score −10.75), suggesting that the relative abundance or bioavailability of quercetin-derived glycosides (represented by 5280459) may diminish in mature leaves, permitting another glycoside (12308704) to dominate the hydrolase interaction profile. All other compounds showed substantially weaker docking at every stage.

Table 2. Stage-specific top docking hits. For each maturation stage, the compound-target pair with the lowest Docking Score is shown.

Stage	CID	PDB ID	Target (class)	Docking Score*
1-5T20	**5280459**	2h44	cGMP-specific 3′,5′-cyclic phosphodiesterase (Hydrolase)	−11.163
2T20	**5280459**	2h44	cGMP-specific 3′,5′-cyclic phosphodiesterase (Hydrolase)	−11.153
4T20	**5280459**	2h44	cGMP-specific 3′,5′-cyclic phosphodiesterase (Hydrolase)	−11.154
6T20	**5280459**	2h44	cGMP-specific 3′,5′-cyclic phosphodiesterase (Hydrolase)	−11.115
8T20	**12308704**	2h44	cGMP-specific 3′,5′-cyclic phosphodiesterase (Hydrolase)	−10.752

* Minor score differences among otherwise identical complexes arise from stochastic elements of the workflow—randomized initial ligand conformers and the non-deterministic search heuristics of the docking engine—so small fluctuations reflect numerical variability rather than real affinity changes.

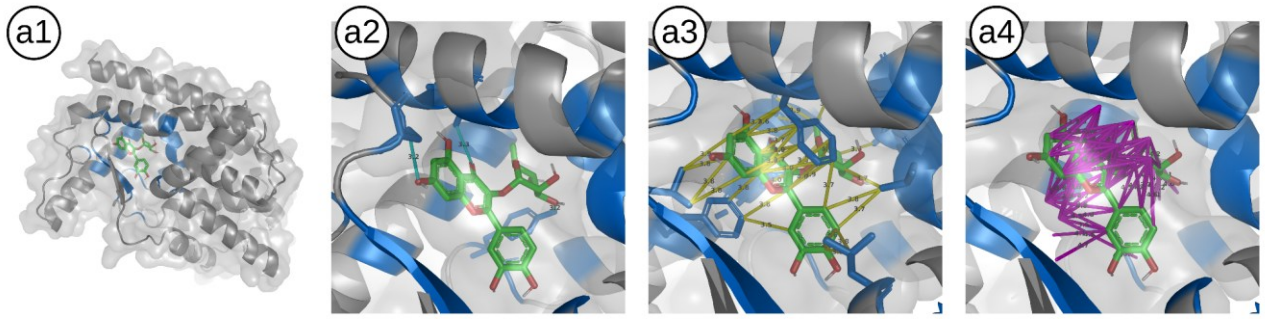


Figure 3. Molecular docking of CID 5280459 into the active site of cGMP-specific 3′,5′-cyclic phosphodiesterase (PDE), using the PDB structure 2H44. (a1) Overall view of the protein–ligand complex, with the protein rendered as a semi-transparent surface and gray cartoon, and the binding-site region highlighted in blue. The ligand is shown as element-colored sticks. (a2) Hydrogen-bond interactions detected using a 3.3 Å cutoff (cyan dashed lines). (a3) Hydrophobic contacts identified within 4.0 Å of the ligand (yellow dashed lines). (a4) Aromatic contacts involving ligand aromatic rings and pocket residues, based on a 4.8 Å centroid cutoff (magenta dashed lines). A clipping window was applied to expose the internal binding cavity, improving visualization of interaction geometries. Binding-site residues are shown as blue sticks, and ligand atoms follow standard CPK coloring conventions.

Beyond the hydrolase class, several other targets exhibited intermolecular interactions across stages (Table 4). CID 12005192 consistently achieved favorable interaction energies with protein kinase, transferase, tuberculosis and estrogen receptor targets, reflecting its polypharmacological potential. CID 13964506 provided the best docking energy for the HIV protease class (Docking Score −10.22, stage 4T20). These per-class observations are discussed further below.

3.4. Stage-resolved LigTMap Score rankings

LigTMap Score integrates ligand similarity and IFP-based binding similarity. Table 3 reports, for each stage, the compound-target pair with the highest LigTMap Score. The range of LigTMap Score values (0.65-0.75) reflects substantial similarity to known ligands. The pattern differs markedly from the docking-based ranking:

during the early (1-5T20) and mid (4T20 and 6T20) stages the kinase death-associated protein kinase 1 (DAPK1) was the predominant high-similarity target for CID 5280459, whereas in stage 2T20 the top hit shifted to CID 5488152 targeting a bacterial dehydratase enzyme. In the latest stage (8T20) the methyltransferase indolethylamine N-methyltransferase (PDB 2a14) emerged as the top LigTMap hit with CID 70678861; this compound exhibited very high ligand similarity to S-adenosyl-like ligands (LigandSimilarityScore 0.95) but only moderate docking energy (Docking Score −7.89), illustrating that similarity-driven prioritization can highlight new scaffold classes not evident from docking alone.

Table 3. Stage-specific top LigTMap Score hits. For each maturation stage, the compound-target pair with the highest LigTMap Score is shown together with its component similarity scores and Docking Score.

Stage	CID	PDB ID	Target (class)	LigTMap	Ligand Similarity	Binding Similarity	Docking
1-5T20	5280459	5auw	DAPK1 (Kinase)	0.653	0.718	0.500	−9.885
2T20	5488152	3cf8	(3R)-hydroxymyristoyl-ACP dehydratase (*H. pylori*)	0.669	0.833	0.600	−4.988
4T20	5280459	5auy	DAPK1 (Kinase)	0.667	0.736	0.556	−9.709
6T20	5280459	5auu	DAPK1 (Kinase)	0.667	0.735	0.556	−9.686
8T20	70678861	2a14	Indolethylamine N-methyltransferase (Transferase)	0.754	0.949	0.300	−7.889

3.5. Best docking interactions by target class

Across all stages, the compound–target pairs with the lowest Docking Scores within each target class highlight clear trends in predicted ligand affinity (Table 4 and Figure 4). Hydrolase, kinase and transferase targets exhibited the most favorable docking energies, frequently involving CID 5280459 or CID 12005192. CID 5280459 showed the strongest predicted affinity against cGMP-specific phosphodiesterase (PDE; −11.16), an enzyme that regulates intracellular cyclic nucleotide signaling and is a validated target in cardiovascular and inflammatory diseases (Omori & Kotera, 2007); the interaction involved an extensive network of 19 pocket residues with multiple hydrogen-bond contacts at Tyr612, Ser663 and Ile665. CID 12005192 displayed consistently high-affinity binding across multiple proteins, including eIF2α kinase (−11.13), a central regulator of the integrated stress response involved in cell survival and protein synthesis control (Wek, 2018); Tankyrase-2 (−11.06), a poly-ADP-ribosylase implicated in Wnt/β-catenin signaling and a target of interest in cancer and fibrosis (Mariotti et al., 2016); pantothenate synthetase (−10.80), a bacterial enzyme essential for coenzyme A biosynthesis and a validated anti-tuberculosis target (Wang & Eisenberg, 2003); the estrogen receptor (−10.50), a nuclear receptor governing reproductive, metabolic and bone homeostasis (Nilsson et al., 2001); and PPARγ (−10.17), a ligand-activated transcription factor central to adipogenesis, insulin sensitization and inflammatory modulation (Ahmadian et al., 2013), consistent with the polypharmacological potential of its heterocyclic quinazoline scaffold. In these complexes, the ligand occupied well-defined hydrophobic pockets, i.e. contacting 15 residues in PPARγ with several H-bond candidates involving Ser289 and Arg288, engaging 19 pocket residues including Leu346 and Thr347 in the estrogen receptor with three H-

bond candidates to Asp351 and Leu346, interacting with 24 residues in pantothenate synthetase with four H-bond candidates involving His47 and Tyr82, forming 18-pocket-residue contacts and nine H-bond candidates in Tankyrase-2 with prominent interactions to His1031 and Tyr1060, and establishing 18 pocket contacts with nine H-bond candidates in eIF2α kinase, notably involving Asp955 and Phe956. For the anticoagulant class, CID 12308704 showed strong predicted affinity for coagulation factor VII (−10.45), a serine protease that initiates the extrinsic coagulation cascade and is a target for antithrombotic therapy (Eigenbrot, 2002), with 18 pocket residues and 13 H-bond candidates dominated by Ser190 and Lys192. For the HIV class, CID 13964506 docked to HIV protease (−10.22), an aspartyl protease indispensable for viral polyprotein processing and a cornerstone of antiretroviral drug design (Kohl et al., 1988), engaging 17 residues with 10 H-bond candidates centered on Asp29, Asp30 and Gly49.

Table 4. Best docking hits per target class. For each class the row reports the stage, compound, PDB entry, target name, Docking Score and corresponding LigTMap Score.

Target class	Stage	CID	PDB ID	Target	Docking Score	LigTMap Score
Anticoagulant	6T20	12308704	2flr	Coagulation factor VII	−10.449	0.336
Diabetes	4T20	12005192	1zeo	Peroxisome proliferator-activated receptor γ	−10.171	0.376
Estrogen	2T20	12005192	2iog	Estrogen receptor	−10.505	0.353
HIV	4T20	13964506	3ok9	HIV protease	−10.223	0.311
Hydrolase	1-5T20	5280459	2h44	cGMP-specific phosphodiesterase	−11.163	0.575
Kinase	1-5T20	12005192	4x7h	eIF2α kinase	−11.133	0.485
Transferase	4T20	12005192	4tk5	Tankyrase-2	−11.063	0.439
Tuberculosis	2T20	12005192	3le8	Pantothenate synthetase	−10.804	0.456

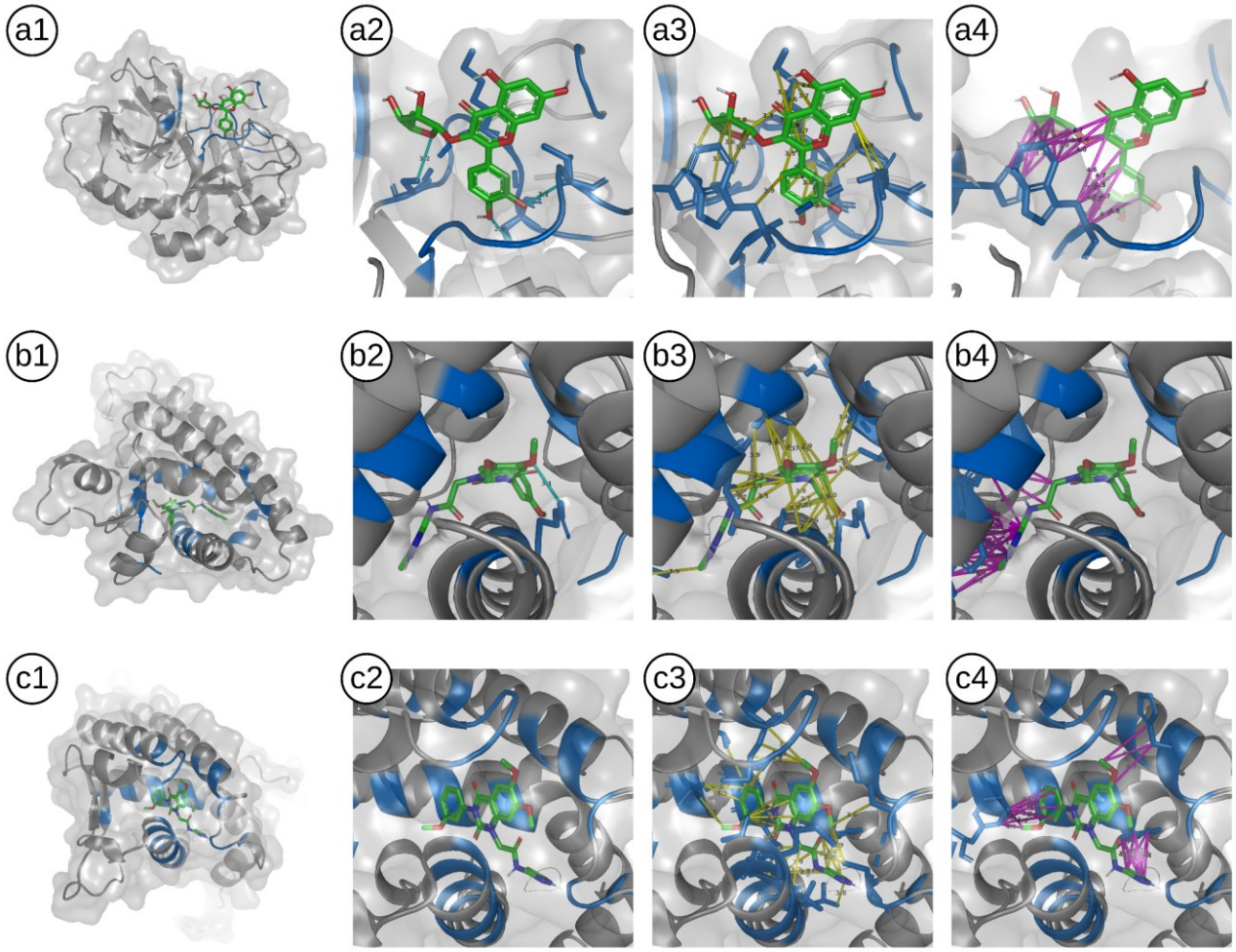


Figure 4. Molecular docking of the three highest-scoring ligand–target complexes identified across distinct biological target classes. Panels (a1–a4) show docking of CID 12308704 into coagulation factor VII (PDB ID: 2FLR; anticoagulant target, stage 6T20); panels (b1–b4) show docking of CID 12005192 into peroxisome proliferator-activated receptor γ (PDB ID: 1ZEO; diabetes-related target, stage 4T20); and panels (c1–c4) show docking of CID 12005192 into the estrogen receptor (PDB ID: 2IOG; estrogen-related target, stage 2T20). For each row, (1) depicts the overall protein–ligand complex with the protein rendered as a semi-transparent surface and gray cartoon and the binding site highlighted in blue; (2) shows hydrogen-bond interactions detected using a 3.3 Å cutoff (cyan dashed lines), with no hydrogen bonds observed in panel c2; (3) displays hydrophobic contacts within 4.0 Å of the ligand (yellow dashed lines); and (4) highlights aromatic contacts based on a 4.8 Å centroid cutoff (magenta dashed lines). Binding-site residues are shown as blue sticks, ligand atoms are rendered as element-colored sticks following CPK conventions, and a clipping window was applied to expose the internal binding cavities.

3.6. Best LigTMap Score interactions by target class

The compound–target pairs with the highest LigTMap Scores within each target class reveal cases in which ligand similarity signals strong biological relevance even when docking energies are moderate (Table 5 and Figure 5). The methyltransferase-like enzyme classes (transferase and hydrolase) were dominated by CID 70678861, which achieved LigTMap Scores of 0.67–0.75 despite Docking Scores of −5.06 to −7.89, consistent with its behaviour as an adenosyl analogue capable of mimicking S-adenosyl substrates. In the hydrolase complex (1xsc), the ligand was accommodated by an 11-residue pocket forming 198 atom–atom contacts, with Lys94, Glu115 and Lys47 contributing most heavily and seven hydrogen-bond candidates stabilising the interaction against bis(5′-nucleosyl) tetraphosphatase, an enzyme involved in nucleotide pool regulation and cellular stress signaling (Guranowski, 2000). In the transferase complexes, CID 70678861 engaged the

catalytic regions of indolethylamine N-methyltransferase (2a14), an enzyme that catalyses the N-methylation of tryptamine derivatives and has been associated with endogenous psychoactive compound biosynthesis and stress response (Thompson & Szabo, 2020), forming 20 pocket contacts with strong contributions from Asp161, Leu19 and Val96, supported by 11 hydrogen-bond candidates involving Lys42, Val96, Glu64 and Asn144, while in SETD7 (3cbp) – a histone-lysine methyltransferase that regulates gene expression through H3K4 monomethylation and is implicated in cancer and metabolic disease (Oudhoff et al., 2013) – the ligand was enclosed by 12 residues, with Trp352, His297 and Asn296 dominating the 106 atom–atom contacts and seven H-bond candidates including interactions with Asn282, Asn296 and His297. CID 70678861 likewise formed a well-buried pose in pantothenate synthetase (3iob), a bacterial enzyme essential for coenzyme A biosynthesis and a validated anti-tuberculosis target (Wang & Eisenberg, 2003), engaging 20 residues with strong contributions from Phe86, Tyr11 and Tyr20, and forming 11 hydrogen-bond candidates with Thr87, Tyr20, Asp85 and Asp142. Within the kinase class, the top LigTMap hit was CID 12308704 against DAPK1 (LigTMap Score 0.688; BindingSimilarityScore 0.73; Docking Score −9.87), a calcium/calmodulin-regulated kinase that modulates apoptosis, autophagy and tumor suppression (Bialik & Kimchi, 2006); here, the ligand contacted seven pocket residues with hydrophobic contacts dominated by Phe59, Pro112 and Lys62, consistent with a compact ATP-site binding mode. For the isomerase class, CID 737375 emerged as the leading candidate targeting macrophage migration inhibitory factor (LigTMap Score 0.647; Docking Score −5.41), a pleiotropic cytokine with tautomerase activity that promotes inflammation and has been implicated in autoimmune disease and cancer progression (Bloom & Bennett, 1966; Leng & Bucala, 2006), with binding supported by an eight-residue pocket dominated by Pro2, Tyr37 and Phe114 (together accounting for > 70 % of close-range contacts) and two H-bond candidates involving Pro2 and Tyr37. Together, these results illustrate how similarity-based methods can highlight biologically meaningful interactions that may not be apparent from docking scores alone.

Table 5. Best LigTMap Score hits per target class. For each class the row reports the stage, compound, PDB entry, target name, LigTMap Score, its components and Docking Score.

Target class	Stage	CID	PDB ID	Target	LigTMap	Ligand Similarity	Binding Similarity	Docking
Estrogen	8T20	**70678861**	3cbp	SETD7 (Histone-lysine N-methyltransferase)	0.708	0.87	0.46	−6.487
H. pylori	2T20	**5488152**	3cf8	(3R)-hydroxymyristoyl-ACP dehydratase	0.669	0.833	0.60	−4.988
Hydrolase	8T20	**70678861**	1xsc	Bis(5′-nucleosyl) tetraphosphatase	0.669	0.86	n.a.	−5.057
Isomerase	1-5T20	**737375**	3wnr	Macrophage migration inhibitory factor	0.647	0.80	0.60	−5.412
Kinase	8T20	**12308704**	5auw	DAPK1	0.688	0.72	0.73	−9.868
Transferase	8T20	**70678861**	2a14	Indolethylamine N-methyltransferase	0.754	0.949	0.30	−7.889

Tuberculosis	8T20	**70678861**	3iob	Pantothenate synthetase	0.701	0.86	0.50	−8.178

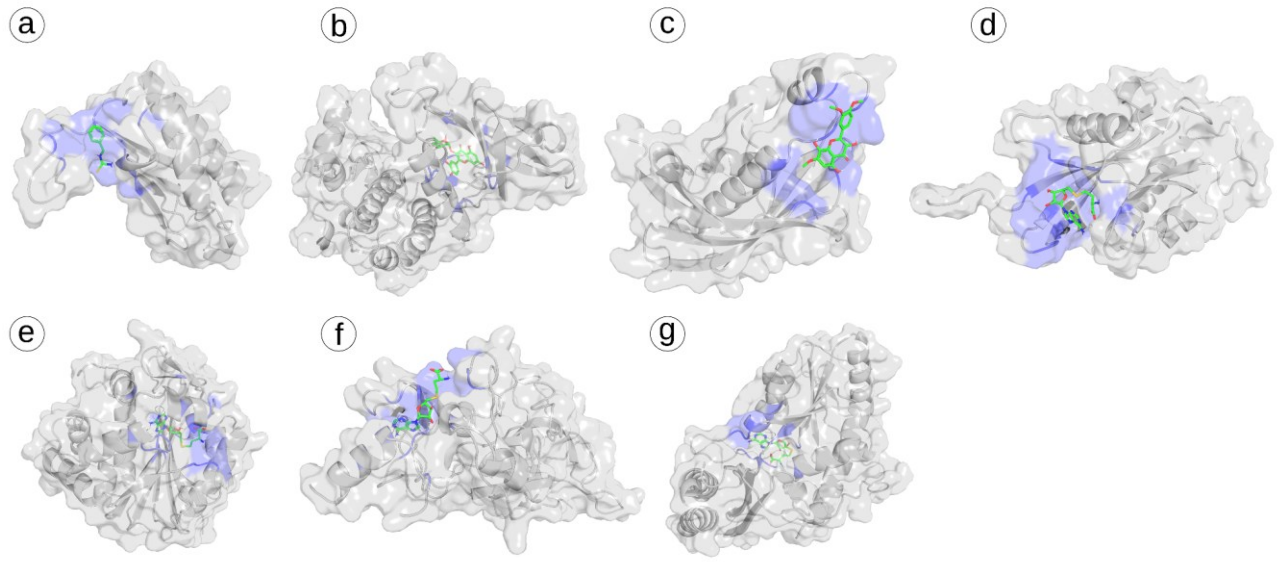


Figure 5. Predicted binding modes of the top-scoring ligand–target complexes ranked by LigTMap Score. Representative docking poses are shown for the highest-scoring compound–target pair within each target class identified in the LigTMap analysis. Proteins are displayed as semi-transparent surfaces with cartoon secondary structure, and ligand–interaction regions are highlighted in blue. Ligands are shown as sticks (carbon in green). (a) CID 737375 bound to macrophage migration inhibitory factor (PDB 3wnr, Isomerase class; stage 1-5T20); (b) CID 5280459 bound to bis(5′-nucleosyl) tetraphosphatase (PDB 5auw, Hydrolase class; stage 8T20); (c) CID 5488152 bound to (3R)-hydroxymyristoyl-ACP dehydratase (PDB 3cf8, *H. pylori* class; stage 2T20); (d) CID 70678861 bound to indolethylamine N-methyltransferase (PDB 2a14, Transferase class; stage 8T20); (e) CID 70678861 bound to SETD7 histone-lysine N-methyltransferase (PDB 3cbp, Estrogen class by classification; stage 8T20); (f) CID 12308704 bound to DAPK1 (PDB 5auw, Kinase class; stage 8T20); and (g) CID 70678861 bound to pantothenate synthetase (PDB 3iob, Tuberculosis class; stage 8T20).

3.7. Correlation between docking energy and similarity-based scoring

Because Docking Score (an energy-based metric) and LigTMap Score (a similarity/IFP-based metric) reflect different information sources, their relationship was examined quantitatively. In the docking-selected set, LigTMap Score and Docking Score displayed a modest anticorrelation (Pearson $r \approx -0.26$), whereas in the LigTMap-selected set the correlation was weakly positive ($r \approx 0.19$). Predicted affinities (from the machine-learning model) were essentially uncorrelated with Docking Score ($r \approx -0.08$), confirming that the learning model captures features not trivially encoded by docking energies. These observations corroborate the complementary nature of docking and similarity-based ranking: docking prioritises low-energy binders such as CIDs 5280459 and 12005192, while LigTMap Score highlights analogues with high ligand and interaction similarity, notably CID 70678861.

3.8. Implications for plant maturation stages

The combined Docking Score- and LigTMap Score-based analyses reveal clear maturation-dependent shifts in the predicted target landscape of *V. planifolia* leaf metabolites (Figures 6 and 7). When examining the minimum DockingScore per stage, early developmental phases (1–5T20 through 6T20) are dominated by high-affinity interactions with hydrolases—particularly the cGMP-specific 3′,5′-cyclic phosphodiesterase—showing the most favorable docking energies across the heatmap. This persistent early-stage enrichment

suggests that the metabolic profile of young leaves favors glycosylated flavonoids capable of stably occupying phosphodiesterase catalytic pockets. Additional early-stage signals include strong binding to eIF2α kinase, pantothenate synthetase, and Tankyrase-2, all of which appear with DockingScores near −11.0, indicating potential functional redundancy or broad molecular reach among phenolic constituents in early ontogeny.

As the plant progresses into later stages, a broader set of targets emerges. By 8T20, the top DockingScore hits shift toward kinases (death-associated protein kinase 1), proteases, and coagulation-factor-like enzymes, accompanied by a noticeable decline in phosphodiesterase dominance. This transition aligns with the chemical diversification observed in maturing leaves, where quercetin-derived glycosides and related polyphenols appear comparatively reduced, allowing alternative metabolite classes to govern the predicted target spectrum.

The LigTMap Score heatmap provides a complementary perspective rooted in ligand similarity rather than binding energy. Here, a striking pattern emerges at the 8T20 stage: methyltransferase-related enzymes—histamine N-methyltransferase, spermidine synthase, SETD7, and indolethylamine N-methyltransferase—consistently register the highest LigTMap Scores (≥0.70). These targets are repeatedly associated with CID 70678861, whose adenosyl-like scaffold strongly resembles known S-adenosyl-substrate ligands and thus receives high similarity contributions despite moderate docking energies. The appearance of this cluster only at the latest maturation stage suggests a shift toward metabolites capable of mimicking SAM-dependent cofactors, potentially reflecting late-stage biochemical transitions involving amino-group transfer or methylation reactions.

Intermediate stages (2T20 and 4T20) exhibit mixed profiles: (3R)-hydroxymyristoyl-ACP dehydratase and pantothenate synthetase contribute prominent LigTMap signals, while moderate DockingScore signals arise from acetylcholinesterase and estrogen receptor. These transitional patterns likely reflect the co-occurrence of early-stage flavonoids with diversifying secondary metabolites that become biosynthetically accessible as leaves mature.

Taken together, these heatmaps indicate that *V. planifolia* leaf ontogeny produces a dynamic molecular matrix that alters its predicted pharmacological space. Early stages emphasize tight-binding flavonoid interactions with hydrolases and kinases, whereas late stages introduce metabolite scaffolds that resemble canonical cofactors and shift the predicted target spectrum toward methyltransferases and more specialized enzymatic classes. This maturation-linked redistribution of predicted targets provides a mechanistic framework for understanding how phenological timing influences the biological potential of *V. planifolia* extracts.

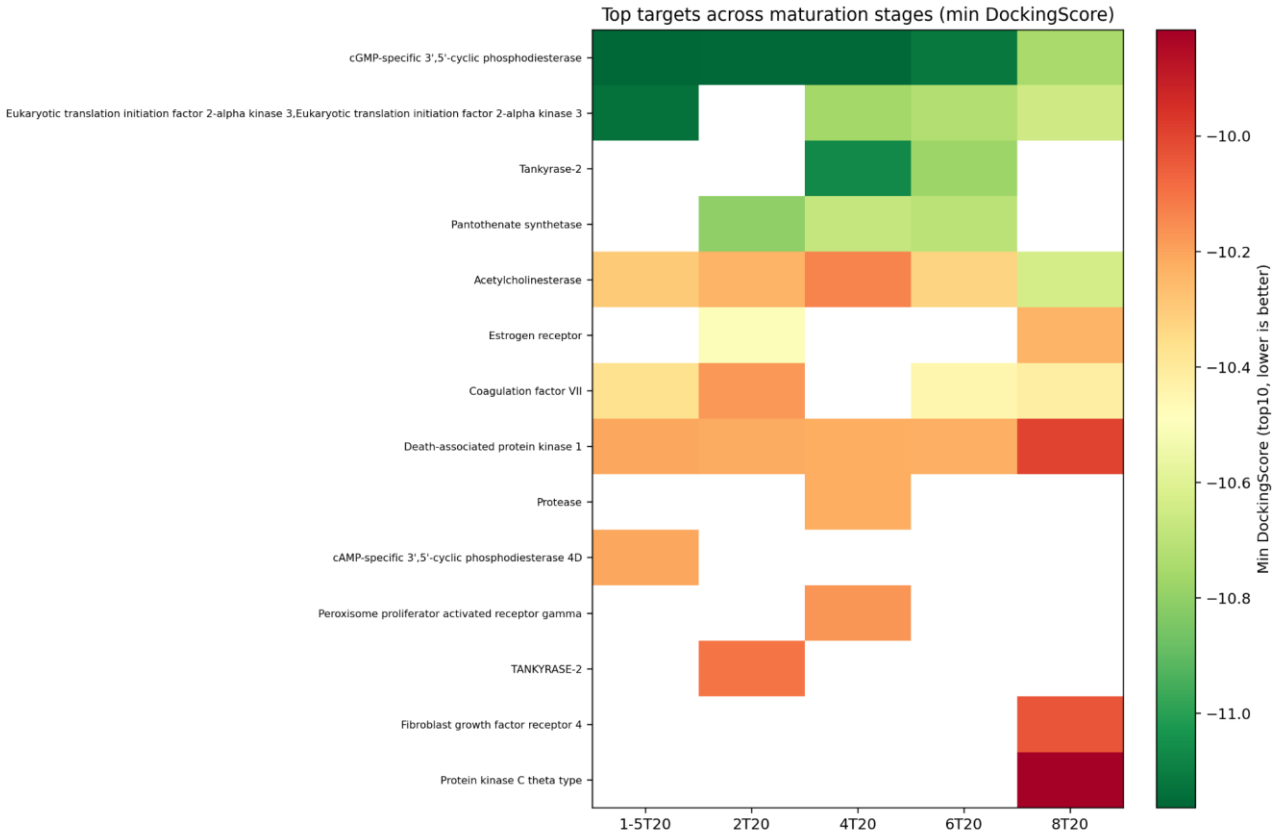


Figure 6. Heatmap of the most strongly predicted protein targets across maturation stages based on minimum Docking Score values. Darker green tones indicate more favorable predicted binding (lower Docking Score), highlighting the dominance of phosphodiesterases and kinases in early stages and a diversification of top targets at 8T20.

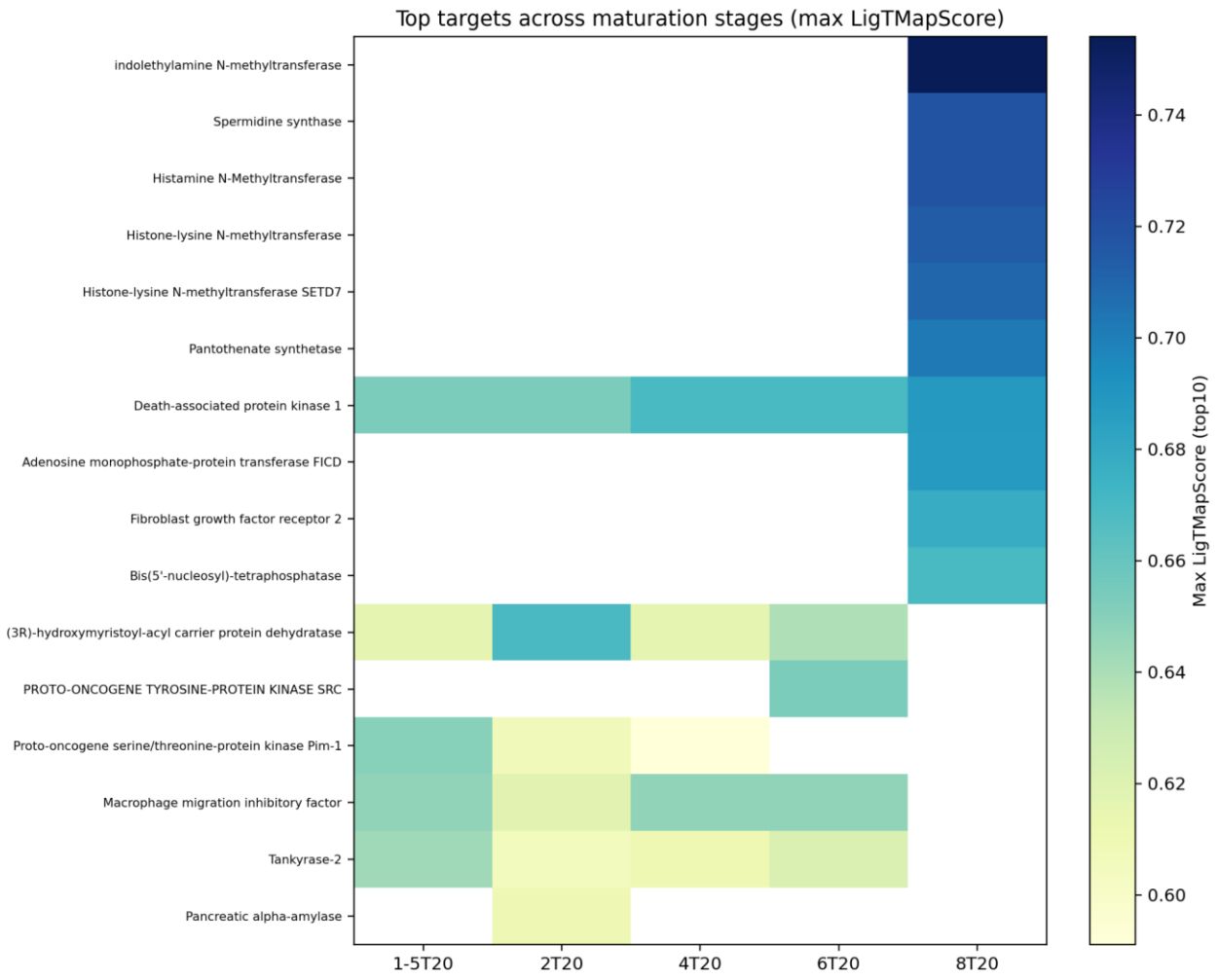


Figure 7. Heatmap of the highest LigTMapScores across maturation stages, illustrating similarity-based target enrichment. Late-stage leaves (8T20) show strong convergence toward methyltransferase-related enzymes, reflecting the emergence of metabolites with high resemblance to SAM-dependent cofactors.

3.9. Predicted ADMET profile of docked *V. planifolia* leaf metabolites

To complement the docking-based prioritization, the six metabolites selected for further analysis (CIDs 5280459, 12005192, 12308704, 45359684, 10073778 and 13964506) were submitted to the batch prediction workflow of ADMETlab 3.0. This platform delivers 119 pharmacokinetic and toxicity endpoints using multi-task graph neural network models, and the results were used to assess the developability of the docked hits. The endpoints are grouped below into absorption, distribution, metabolism, excretion and toxicity (ADMET), and Table 6 summarizes the qualitative assessment of key endpoints.

3.9.1 Absorption

Solubility and lipophilicity. Logarithmic aqueous solubility (logS) values between −4 and 0.5 log mol·L^{-1} are considered acceptable for oral drugs. All analyzed metabolites fell within this range (−3.84 to −0.82), indicating solubilities compatible with oral delivery. The logP values of quercitrin (CID 5280459) and the quinazolinone derivative (CID 12005192) were moderate (≈1.3–1.5), while the glycosylated metabolites were hydrophilic (logP ≈ –1.7 to –0.9). Since optimal logP values lie between 0 and 3, only the first two compounds balance permeability and solubility, whereas the highly polar glycosides may require transport-mediated uptake.

Membrane permeability. Caco-2 cell permeability is a widely used surrogate for intestinal absorption; compounds with predicted values greater than −5.15 log cm·s^{-1} are regarded as having good passive permeability. Only CID 12005192 (−5.08 log cm·s^{-1}) met this criterion; all glycosylated metabolites exhibited much lower permeability (≤ −6.04 log cm·s^{-1}), consistent with their high polar surface areas. Parallel artificial membrane permeability assay (PAMPA) predictions displayed the opposite pattern: most glycosides had high probabilities of permeation (> 0.95), whereas CID 12005192 scored poorly (≈0.01), highlighting that model definitions can vary depending on the assay.

Transporter interactions and oral bioavailability. None of the six metabolites were predicted to inhibit P-glycoprotein (all probabilities < 0.03), but several are likely to be efflux substrates. The lignan glycoside (CID 45359684) showed a high probability of being a P-gp substrate (0.83), whereas quercitrin and quercetin 3-arabinoside (CIDs 5280459 and 12308704) were predicted to be non-substrates. Probabilities for human intestinal absorption (HIA+) above 0.7 indicate poor absorption; the triglycosides (CIDs 45359684, 10073778 and 13964506) all exceeded 0.99, suggesting limited uptake in the gut, whereas quercitrin (0.10) and the quinazolinone (0.000003) were predicted to have good absorption. Likewise, the predicted probabilities that oral bioavailability would fall below 20 % (F20%+) showed that only CID 12005192 had a low risk of poor bioavailability (≈10^{-5}), while the glycosides displayed medium to high risk (0.62–0.99). Oral bioavailability prediction is a critical early filter in drug development, as compounds with F < 20% rarely achieve therapeutic plasma concentrations without formulation intervention (Daina et al., 2017).

3.9.2 Distribution

Plasma protein binding (PPB) values below 90 % are considered acceptable. All glycosides (CIDs 12308704, 45359684, 10073778 and 13964506) bound plasma proteins moderately (49–60 %), whereas CID 12005192 bound extensively (95 %), potentially limiting free drug concentrations. The fraction of unbound drug (Fu) was >5 % for all compounds except CID 12005192 (4 %), indicating that most metabolites are expected to circulate largely unbound. Predicted brain penetration was negligible: probabilities of crossing the blood-brain barrier were below 0.05 for all compounds, reflecting logBB ≤ –1 and categorizing them as BBB–. This suggests that the compounds are unlikely to cause central nervous system effects, which is desirable for non-CNS targets. BBB permeability assessment is particularly relevant when predicted molecular targets, such as kinases and nuclear receptors, are expressed in both peripheral and central tissues, since CNS exclusion can be either a safety advantage or a therapeutic limitation depending on the indication (Pajouhesh & Lenz, 2005).

3.9.3 Metabolism

The cytochrome P450 (CYP) models assign probabilities of being inhibitors or substrates for the major isoforms (1A2, 2C9, 2C19, 2D6 and 3A4). Quercitrin (CID 5280459) was predicted to be a strong inhibitor of CYP3A4 (0.94) but not a substrate, raising the possibility of drug–drug interactions if co-administered with CYP3A4-metabolized drugs. The quinazolinone derivative (CID 12005192) behaved conversely: it is likely a substrate for CYP3A4 (0.998) and CYP2C19 (0.91) with moderate inhibitory activity (≈0.25), suggesting rapid metabolic turnover. Quercetin 3-arabinoside (CID 12308704) displayed moderate CYP3A4 inhibition (0.45) and was predicted to be a substrate for CYP2D6 (0.47) and CYP2C9 (0.64). The lignan glycoside (CID 45359684) showed a high probability (0.99) of being metabolized by CYP2C9 but negligible inhibition of any isoform. Scroside D (CID 10073778) and ptelatoside A (CID 13964506) were predicted to neither inhibit nor be substrates for major CYPs, indicating a low propensity for metabolic drug–drug interactions. CYP inhibition and substrate profiling are standard early-stage flags for drug–drug interaction potential, as polyphenols are well-documented modulators of CYP3A4 and CYP2C9 activity in vivo (Pelkonen et al., 2008).

3.9.4 Excretion

Clearance values > 5 ml·min⁻¹·kg⁻¹ are considered acceptable. Only quercetin 3-arabinoside (CID 12308704) reached this threshold (≈5.2 ml·min⁻¹·kg⁻¹), while the remaining metabolites were predicted to have low clearance (1.2–4.3 ml·min⁻¹·kg⁻¹). These lower clearances translate into moderate half-lives; most metabolites showed half-life predictions around 2–3.7 h, whereas CID 12005192 had a short predicted half-life of 0.78 h. The half-life classification endpoint in ADMETlab considers compounds with half-lives ≤ 3 h as T1/2+ (undesirable); thus, CID 12005192 would be expected to clear quickly. Clearance and half-life predictions are central to dose regimen design, as compounds with low clearance and extended half-lives may accumulate to toxic levels, while those with very short half-lives require frequent dosing to maintain therapeutic concentrations; both extremes represent liabilities in early-stage candidate selection (Rowland & Tozer, 2011).

3.9.5 Toxicity

Cardiotoxicity and hepatotoxicity. All six metabolites had low probabilities (<0.22) of blocking hERG potassium channels, suggesting a low risk of QT prolongation. In contrast, drug-induced liver injury (DILI) probabilities were high for most compounds: quercitrin, the quinazolinone derivative, quercetin 3-arabinoside, lignan glycoside and scroside D all exceeded 0.70 (high risk), whereas ptelatoside A had a moderate DILI probability of 0.36. DILI remains the leading cause of drug withdrawal and clinical attrition, and in silico DILI prediction has demonstrated practical utility as a pre-experimental triage tool for natural product candidates (Chen et al., 2016).

Mutagenicity and carcinogenicity. ADMETlab models classify compounds as Ames positive when the probability exceeds 0.7. Quercitrin, scroside D and ptelatoside A were predicted to be mutagenic (probabilities 0.75, 0.92 and 0.77 respectively), whereas the quinazolinone and quercetin 3-arabinoside had medium mutagenicity risk (≈0.56–0.63). Only the quinazolinone derivative exhibited a high probability of carcinogenicity (0.81), while quercetin 3-arabinoside was intermediate (0.36) and the other compounds were low (0.03–0.12). Genotoxicity models similarly flagged quercitrin, the quinazolinone and quercetin 3-arabinoside as high-risk (probabilities > 0.94) but predicted low genotoxicity for scroside D and ptelatoside A. Neurotoxicity and haematotoxicity probabilities were generally low to medium except for the quinazolinone derivative, which showed high neurotoxicity (0.99) and haematotoxicity (0.72). Thus, CID 12005192, despite its favorable absorption, carries multiple toxicity liabilities. Ames mutagenicity and genotoxicity flags are routinely required by regulatory agencies prior to any in vivo testing, making their early computational assessment essential for candidate prioritization (Kirkland et al., 2011).

Table 6. Summary of key ADMETlab 3.0 predictions for selected *V. planifolia* metabolites, using abbreviated endpoints: Caco-2 permeability (Caco-2), P-glycoprotein substrate probability (P-gp Sub), human intestinal absorption class (HIA), plasma protein binding category (PPB), drug-induced liver injury risk (DILI), Ames mutagenicity test (Ames), carcinogenicity (Carc.), neurotoxicity (Neurotox.), and genotoxicity (Genotox.); Medium (Med.).

CID	Caco-2	P-gp Sub	HIA	PPB	DILI	Ames	Carc.	Neurotox.	Genotox.
5280459	Low	No	Good	Mod.	High	High	Low	Low	High
12005192	High	No	Good	High	High	Med.	High	High	High
12308704	Low	No	Med.	Mod.	Med.	Med.	Med.	Low	High

45359684	Low	Yes	Poor	Mod.	High	Med.	Low	Med.	Med.
10073778	Low	Med.	Poor	Mod.	High	High	Low	Low	Low
13964506	Low	Med.	Poor	Mod.	Med.	High	Low	Med.	Low

Note: qualitative assignments are based on ADMETlab threshold criteria: Caco-2 permeability > −5.15 log cm·s⁻¹ denotes high permeability; P-gp substrate probabilities ≥ 0.7 indicate high likelihood of efflux; HIA probabilities < 0.3 correspond to good absorption; PPB > 90 % is considered high; DILI, Ames, carcinogenicity, neurotoxicity and genotoxicity probabilities < 0.3 are rated low, 0.3–0.7 medium and > 0.7 high.

3.9.6 Implications for lead prioritization

The ADMET profiling underscores the trade-offs inherent to the docked hits. Quercitrin (CID 5280459) and quercetin 3-arabinoside (CID 12308704) exhibit acceptable solubility and limited BBB penetration but suffer from poor permeability and high risk of hepatic/genotoxic toxicity. The quinazolinone derivative (CID 12005192) combines good absorption and metabolism with extensive protein binding and numerous toxicity liabilities (carcinogenicity, neurotoxicity and haematotoxicity), raising safety concerns despite its excellent docking scores. The lignan glycoside (CID 45359684) and the triglycosides Scroside D (CID 10073778) and Ptelatoside A (CID 13964506) are predicted to be poorly absorbed, consistent with their high sugar content, but show relatively benign metabolic profiles. Among these, Ptelatoside A stands out with the lowest genotoxicity and carcinogenicity probabilities and moderate DILI risk, suggesting that late-stage metabolites may combine favorable ADMET with diverse target profiles.

4. Discussion

Vanilla research has historically centered on vanillin and pod curing, but the genus *Vanilla* is chemically richer than a single compound, making non-vanillin metabolites plausible contributors to bioactivity-relevant chemical space (Hernández-Peña et al., 2025). Leaf-focused metabolomics in *V. planifolia* has repeatedly shown that vegetative tissues contain broad metabolite classes (e.g., sugars, amino acids, organic acids, phenolics and related specialized metabolites) whose relative levels vary with leaf age, season, and diel cycle, supporting the idea that "bioactivity candidates" are not limited to pods (Palama et al., 2010). Metabolomics under biotic stress further supports functional chemical breadth in vegetative organs by demonstrating that viral infection can shift leaf metabolomic profiles, including changes in amino acids, sugars, and phenolic-associated signals in controlled in vitro settings (Palama et al., 2010). Cross-species leaf metabolomic profiling has also demonstrated that leaves can discriminate *Vanilla* species based on quantified metabolite signatures, reinforcing leaves as a biologically informative matrix rather than a passive background tissue (Leyva et al., 2021). More recently, expanded vegetative-organ metabolomics using LC–MS together with GC×GC–MS has provided a substantially broader structural inventory in *V. planifolia*, explicitly addressing earlier gaps in LC–MS/GC–MS coverage for vanilla leaves (Beer et al., 2025). Explored in the present study, integrative developmental datasets combining leaves with fruit/pod development have further strengthened the rationale that leaf chemistry participates in phenylpropanoid-related metabolic networks relevant to vanillin biosynthesis, while simultaneously providing many non-vanillin annotations amenable to downstream bioactivity inference (Hernández-Peña et al., 2025). Against this literature backdrop, computational prioritization that elevates non-vanillin leaf metabolites is consistent with the demonstrated chemical diversity and context-dependence of vegetative metabolomes in *V. planifolia* and related *Vanilla* species (Palama et al., 2010; Leyva et al., 2021; Beer et al., 2025). At the same time, the relative scarcity of mechanistic bioactivity studies explicitly tied to leaf-annotated vanilla metabolites implies that many predicted "beyond-vanillin" candidates are more likely to be under-investigated than contradicted by existing evidence, reinforcing the value of hypothesis-generating workflows anchored in vegetative metabolomics (Beer et al., 2025; Hernández-Peña et al., 2025).

Reverse-docking of *V. planifolia* leaf metabolites consistently prioritizes protein target classes that are already recognized in the literature as frequent interaction hubs for plant-derived phenolics, including phosphodiesterases, kinases, and nuclear receptors (Shaikh et al., 2021; Atanasov et al., 2021). Predicted high-affinity interactions with phosphodiesterases, PPARγ, estrogen receptors, and tankyrase-related enzymes align with extensive evidence that flavonoids and related phenylpropanoid-derived scaffolds can modulate these target families through conserved binding motifs (Ko et al., 2004; Sheng et al., 2022; Treml et al., 2025; Kiyama, 2023; Bolt et al., 2024; Narwal et al., 2013). The recurrence of these target classes across multiple vanilla leaf–derived metabolites supports the interpretation that the reverse-docking results are not arbitrary, but instead reflect known patterns of polyphenol–protein interaction reported across diverse plant systems (Pinzi and Rastelli, 2019; Atanasov et al., 2021).

In contrast, this analysis extends existing literature by demonstrating that several non-vanillin metabolites from vanilla leaves map onto the same biologically relevant target space as better-studied flavonoids, despite lacking prior experimental bioactivity characterization (Hernández-Peña et al., 2025). A second insight lies in constraining docking-based predictions through ADMET screening, which reveals that only a subset of high-scoring ligands remains compatible with realistic systemic exposure (Fu et al., 2024). The finding that aglycone-like structures exhibit more favorable predicted permeability than glycosylated counterparts is consistent with pharmacokinetic studies showing that flavonoid glycosylation generally reduces intestinal absorption and increases first-pass metabolism (Németh et al., 2003; Zhang et al., 2020; Hu et al., 2025). This exposure-based filtering provides a mechanistic explanation for why strong *in silico* binding affinities reported for polyphenols do not always translate into systemic bioactivity in vivo, a discrepancy repeatedly highlighted in pharmacokinetic and nutrition literature (Jones and Rowland-Yeo, 2013; Quesada-Vázquez et al., 2024). By integrating reverse docking with ADMET constraints, this work advances prior computational screenings by distinguishing between structurally plausible interactions and those that are additionally compatible with exposure-related limitations, thereby reducing overinterpretation of docking scores alone (Fu et al., 2024; Pinzi and Rastelli, 2019).

It should also be noted that the metabolite annotations underlying this workflow were derived from untargeted LC–MS data, which by nature represents a screening-level approach subject to annotation uncertainty. The structural identities of the compounds analyzed here have not been confirmed by targeted acquisition against authenticated reference standards, and some annotations may correspond to isomeric or isobaric structures. While this is an inherent limitation of non-targeted metabolomics, the field is rapidly maturing: advances in spectral library coverage, fragmentation scoring, and confidence-level reporting frameworks are progressively increasing the reliability of untargeted annotations for downstream functional inference (Huber et al., 2021). The results presented here should therefore be interpreted as hypothesis-generating, with targeted validation of candidate metabolite identities representing a necessary next step prior to experimental bioactivity testing.

This study contributes to natural product research by demonstrating that openly available plant metabolomics datasets can be repurposed to generate biologically oriented hypotheses without additional experimental acquisition, addressing a widely recognized bottleneck in natural product discovery pipelines (Wilkinson et al., 2016; Atanasov et al., 2021). Most plant metabolomics studies, including those focused on *Vanilla planifolia*, primarily report compound annotation and relative abundance patterns, while downstream functional interpretation is often left unexplored despite the availability of structural information suitable for *in silico* analyses (Wolfender et al., 2019; Hernández-Peña et al., 2025). By extracting curated metabolite annotations from LC–MS leaf metabolomics and integrating them with reverse docking and ADMET modeling, the present work illustrates how data reuse can extend the scientific value of existing datasets beyond their original biochemical or biosynthetic scope (Wilkinson et al., 2016; Mons et al., 2017). This approach aligns with recent calls in metabolomics and chemical ecology to move from descriptive metabolite cataloging

toward hypothesis-driven frameworks that connect chemical diversity to potential biological function (Wolfender et al., 2019; Perez de Souza et al., 2020).

In contrast to traditional natural product discovery strategies that rely on iterative extraction and bioassay-guided fractionation, computational reuse of metabolomics data enables rapid prioritization of candidate molecules while reducing experimental cost, material demand, and redundancy (Atanasov et al., 2021; Rodrigues et al., 2024). The results further suggest that vegetative-organ metabolomics datasets, which are often secondary to fruit- or product-centered analyses, represent an underutilized reservoir of structurally diverse compounds amenable to functional inference when combined with appropriate *in silico* tools (Beer et al., 2025; Hernández-Peña et al., 2025). These findings support the view that systematic metabolomics data reuse can act as a scalable entry point for natural product bioactivity exploration, complementing experimental validation rather than replacing it (Wilkinson et al., 2016; Wolfender et al., 2019).

The computational prioritization achieved here also opens prospective avenues beyond target identification. Predicted binding poses for high-scoring vanilla leaf metabolites provide structural starting points for ligand-based molecular design, in which scaffold features associated with favorable docking scores and ADMET profiles could guide the rational modification of natural product structures toward improved potency or selectivity (Schneider, 2018; Rodrigues et al., 2024). This de novo or semi-synthetic design perspective is increasingly tractable given the growing integration of generative models and structure-activity relationship tools with natural product scaffolds (Mouchlis et al., 2021). Furthermore, the static docking predictions reported here represent a first approximation of binding geometry that would benefit from molecular dynamics (MD) simulation to assess binding stability, induced-fit effects, and residence time — factors known to substantially influence the biological relevance of predicted interactions (Śledź and Caflisch, 2018). MD-based free energy perturbation approaches, such as MM-GBSA rescoring, have been shown to markedly improve the correlation between computational predictions and experimental binding affinities for plant-derived polyphenols (Genheden and Ryde, 2015). Taken together, molecular design and MD-based validation represent logical and technically feasible next steps for the candidate metabolites prioritized in this study, transforming the present hypothesis-generating workflow into an experimentally actionable pipeline.

5. Conclusion

This study showed that metabolomics data derived from *Vanilla planifolia* leaves can be systematically repurposed to generate biologically informed hypotheses on small-molecule bioactivity through the integration of reverse docking and ADMET screening. Computational prioritization highlighted multiple non-vanillin metabolites whose predicted target profiles converge on protein classes frequently implicated in polyphenol bioactivity, while ADMET constraints refined these predictions by distinguishing structurally plausible interactions from those compatible with realistic exposure scenarios. The results support the view that vegetative-organ metabolomics datasets, often secondary to fruit-centered analyses, represent an underexplored resource for functional inference when coupled with appropriate *in silico* workflows. More broadly, the work illustrates how structured data reuse can extend the scientific value of existing metabolomics studies beyond descriptive chemistry, offering a scalable and cost-effective entry point for natural product bioactivity exploration. Future perspectives include experimental validation of prioritized metabolite–target pairs and the application of similar integrative frameworks to other plant systems, thereby strengthening the connection between metabolomic diversity, biological plausibility, and functional discovery in natural product research.

6. References

Ahmadian, M., Suh, J. M., Hah, N., Liddle, C., Atkins, A. R., Downes, M., & Evans, R. M. (2013). PPARγ signaling and metabolism: the good, the bad and the future. *Nature Medicine*, 19(5), 557–566. https://doi.org/10.1038/nm.3159

Al-Khayri, J. M., Jain, S. M., & Johnson, D. V. (Eds.). (2019). *Advances in Plant Breeding Strategies: Industrial and Food Crops* (Vol. 6). Springer. https://doi.org/10.1007/978-3-030-23265-8

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., the International Natural Product Sciences Taskforce, Orhan, I. E., Banach, M., Rollinger, J. M., Barreca, D., Weckwerth, W., Bauer, R., Bayer, E. A., Majeed, M., Bishayee, A., Bochkov, V., Bonn, G. K., Braidy, N., Bucar, F., Cifuentes, A., D'Onofrio, G., … Supuran, C. T. (2021). Natural products in drug discovery: Advances and opportunities. *Nature Reviews Drug Discovery*, *20*(3), 200–216. https://doi.org/10.1038/s41573-020-00114-z

Banda Ortiz, H., Garza Morales, R., & Vázquez Trujillo, K. V. T. (2021). Supply chain and economic viability of *Vanilla planifolia* Andrew production: A case study. *Agro Productividad*, 14(1). https://doi.org/10.32854/agrop.v14i1.1795

Beer, M. F., Trappitt, A. C., Stander, E. A., & Fester, T. (2025). Comprehensive metabolome characterization of leaves, internodes, and aerial roots of *Vanilla planifolia* by untargeted LC–MS and GC×GC–MS. *Phytochemical Analysis*. https://doi.org/10.1002/pca.3506

Bialik, S., & Kimchi, A. (2006). The death-associated protein kinases: structure, function, and beyond. *Annual Review of Biochemistry*, 75, 189–210. https://doi.org/10.1146/annurev.biochem.75.103004.142615

Bloom, B. R., & Bennett, B. (1966). Mechanism of a reaction in vitro associated with delayed-type hypersensitivity. *Science*, 153(3731), 80–82. https://doi.org/10.1126/science.153.3731.80

Bolt, M. J., Oceguera, J., Singh, P. K., Safari, K., Abbott, D. H., Neugebauer, K. A., Mancini, M. G., Gorelick, D. A., Stossi, F., & Mancini, M. A. (2024). Characterization of flavonoids with potent and subtype-selective actions on estrogen receptors alpha and beta. *iScience*, *27*(3), 109275. https://doi.org/10.1016/j.isci.2024.109275

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Pujadas, G., & Garcia-Vallve, S. (2015). Tools for *in silico* target fishing. *Methods*, *71*, 98–103. https://doi.org/10.1016/j.ymeth.2014.09.006

Chen, M., Suzuki, A., Borlak, J., Andrade, R. J., & Lucena, M. I. (2016). Drug-induced liver injury: interactions between drug properties and host factors. *Journal of Hepatology*, 63(2), 503–514. https://doi.org/10.1016/j.jhep.2015.04.016

da Silva Oliveira, J. P., Garrett, R., Bello Koblitz, M. G., & Furtado Macedo, A. (2025). Integrated targeted and untargeted metabolomics profiling of *Vanilla* species from the Atlantic Forest: Unveiling the bioeconomic potential of *Vanilla cribbiana*. *Food Chemistry*, 464, 141650. https://doi.org/10.1016/j.foodchem.2024.141650

Daina, A., Michielin, O., & Zoete, V. (2017). SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports*, *7*, 42717. https://doi.org/10.1038/srep42717

Eigenbrot, C. (2002). Structure, function, and inhibition of factor VIIa. *Current Protein and Peptide Science*, 3(1), 119–127. https://doi.org/10.2174/1389203023380853

Ekins, S., Mestres, J., & Testa, B. (2007). *In silico* pharmacology for drug discovery: Applications to targets and beyond. *British Journal of Pharmacology*, *152*(1), 21–37. https://doi.org/10.1038/sj.bjp.0707306

Ferreira De Freitas, R., & Schapira, M. (2017). A systematic analysis of atomic protein–ligand interactions in the PDB. *MedChemComm*, *8*(10), 1970–1981. https://doi.org/10.1039/C7MD00381A

Fu, L., Shi, S., Yi, J., Wang, N., He, Y., Wu, Z., Peng, J., Deng, Y., Wang, W., Wu, C., Lyu, A., Zeng, X., Zhao, W., Hou, T., & Cao, D. (2024). ADMETlab 3.0: An updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. *Nucleic Acids Research*, *52*(W1), W422–W431. https://doi.org/10.1093/nar/gkae236

Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery*, 10(5), 449–461. https://doi.org/10.1517/17460441.2015.1032936

Gu, F., Chen, Y., Hong, Y., Fang, Y., & Tan, L. (2017). Comparative metabolomics in vanilla pod and vanilla bean revealing the biosynthesis of vanillin during the curing process of vanilla. *AMB Express*, *7*(1), 116. https://doi.org/10.1186/s13568-017-0413-2

Guranowski, A. (2000). Specific and nonspecific enzymes involved in the catabolism of mononucleoside and dinucleoside polyphosphates. *Pharmacology & Therapeutics*, 87(2–3), 117–139. https://doi.org/10.1016/S0163-7258(00)00046-3

Havkin-Frenkel, D., & Belanger, F. C. (Eds.). (2018). Vanilla in Perfumery and Beverage Flavors. In *Handbook of Vanilla Science and Technology* (1st ed., pp. 367–373). Wiley. https://doi.org/10.1002/9781119377320.ch19

Hernández-Peña, R., Lorenzo-Manzanarez, J. L., Cruz-Ramírez, L. A., Reyes-López, D., Hernández-Domínguez, C., Pascual-Ramírez, F., & Ordaz-Ortiz, J. J. (2025). Unravelling Vanillin Biosynthesis: Integrative Transcriptomic and Metabolomic Insights into Vanilla planifolia Pod Development. *Journal of Agricultural and Food Chemistry*, *73*(30), 19094–19106. https://doi.org/10.1021/acs.jafc.5c05293

Hu, L., Luo, Y., Yang, J., & Cheng, C. (2025). Botanical flavonoids: Efficacy, absorption, metabolism and advanced pharmaceutical technology for improving bioavailability. *Molecules*, *30*(5), 1184. https://doi.org/10.3390/molecules30051184

Huber, F., Ridder, L., Verhoeven, S., Spaaks, J. H., Diblen, F., Rogers, S., & Deckers, J. J. (2021). Spec2Vec: Improved mass spectral similarity scoring through learning of structural relationships. *PLOS Computational Biology*, 17(2), e1008724. https://doi.org/10.1371/journal.pcbi.1008724

Jones, H., & Rowland-Yeo, K. (2013). Basic Concepts in Physiologically Based Pharmacokinetic Modeling in Drug Discovery and Development. *CPT: Pharmacometrics & Systems Pharmacology*, *2*(8), e63. https://doi.org/10.1038/psp.2013.41

Kirkland, D., Pfuhler, S., Tweats, D., Aardema, M., Corvi, R., Darroudi, F., … Stamm, R. (2007). How to reduce false positive results when undertaking in vitro genotoxicity testing. *Mutation Research*, 628(1–2), 31–55. https://doi.org/10.1016/j.mrgentox.2006.11.008

Kiyama, R. (2023). Estrogenic flavonoids and their molecular mechanisms of action. *Journal of Nutritional Biochemistry*, *114*, 109250. https://doi.org/10.1016/j.jnutbio.2022.109250

Ko, W.-C., Shih, C.-M., Lai, Y.-H., Chen, J.-H., & Huang, H.-L. (2004). Inhibitory effects of flavonoids on phosphodiesterase isozymes from guinea pig and their structure–activity relationships. *Biochemical Pharmacology*, *68*(10), 2087–2094. https://doi.org/10.1016/j.bcp.2004.06.030

Kohl, N. E., Emini, E. A., Schleif, W. A., Davis, L. J., Heimbach, J. C., Dixon, R. A., … Sigal, I. S. (1988). Active human immunodeficiency virus protease is required for viral infectivity. *PNAS*, 85(13), 4686–4690. https://doi.org/10.1073/pnas.85.13.4686

Lanzarotti, E., Biekofsky, R. R., Estrin, D. A., Marti, M. A., & Turjanski, A. G. (2011). Aromatic–Aromatic Interactions in Proteins: Beyond the Dimer. *Journal of Chemical Information and Modeling*, *51*(7), 1623–1633. https://doi.org/10.1021/ci200062e

Leng, L., & Bucala, R. (2006). Insight into the biology of macrophage migration inhibitory factor (MIF) revealed by the cloning of its cell surface receptor. *Cell Research*, 16(2), 162–168. https://doi.org/10.1038/sj.cr.7310022

Leyva, E., Flores-Bocanegra, L., Lozada-Pérez, C. A., Pérez-Vásquez, A., & Osuna-Fernández, H. R. (2021). NMR-based leaf metabolic profiling of *V. planifolia* and three endemic *Vanilla* species from the Peruvian Amazon. *Phytochemistry*, 190, 112868. https://doi.org/10.1016/j.phytochem.2021.112868

Luis, J., Barros, R., Sousa, N., Muratov, E., Scotti, L., & Scotti, M. (2020). Virtual Screening of Natural Products Database. *Mini-Reviews in Medicinal Chemistry*, *20*. https://doi.org/10.2174/1389557520666200730161549

Mariotti, L., Templeton, C. M., Ranes, M., Paracuellos, P., Cronin, N., Beuron, F., … Bhatt, D. (2016). Tankyrase requires SAM domain-dependent polymerization to support Wnt-β-catenin signaling. *Molecular Cell*, 63(3), 498–513. https://doi.org/10.1016/j.molcel.2016.06.019

Market Research Future. (2024). *Vanilla market size, growth, forecast, and industry share to 2030*. https://www.marketresearchfuture.com/reports/vanilla-market-7806

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services and Use*, *37*(1), 49–56. https://doi.org/10.3233/ISU-170824

Mouchlis, V. D., Afantitis, A., Serra, A., Fratello, M., Papadiamantis, A. G., Aidinis, V., … Melagraki, G. (2021). Advances in de novo drug design: from conventional to machine learning methods. *International Journal of Molecular Sciences*, 22(4), 1676. https://doi.org/10.3390/ijms22041676

Narwal, M., Singh, N., Ulvila, J., Hytönen, V. P., & Lehtiö, L. (2013). Discovery of tankyrase inhibiting flavones with increased potency and isoenzyme selectivity. *Journal of Medicinal Chemistry*, 56(8), 3507–3517. https://doi.org/10.1021/jm400186w

Németh, K., Plumb, G. W., Berrin, J.-G., Juge, N., Jacob, R., Naim, H. Y., Williamson, G., Swallow, D. M., & Kroon, P. A. (2003). Deglycosylation by small intestinal epithelial cell β-glucosidases is a critical step in the absorption and metabolism of dietary flavonoid glycosides in humans. *European Journal of Nutrition*, *42*(1), 29–42. https://doi.org/10.1007/s00394-003-0397-3

Nilsson, S., Mäkelä, S., Treuter, E., Tujague, M., Thomsen, J., Andersson, G., … Gustafsson, J. Å. (2001). Mechanisms of estrogen action. *Physiological Reviews*, 81(4), 1535–1565. https://doi.org/10.1152/physrev.2001.81.4.1535

Omori, K., & Kotera, J. (2007). Overview of PDEs and their regulation. *Circulation Research*, 100(3), 309–327. https://doi.org/10.1161/01.RES.0000256354.95791.f1

Oudhoff, M. J., Freeman, S. A., Couzens, A. L., Antignano, F., Kuznetsova, E., Min, P. H., … Bhatt, D. (2013). Control of the hippo pathway by Set7/9 (Setd7) histone methyltransferase. *Molecular Cell*, 51(2), 213–225. https://doi.org/10.1016/j.molcel.2013.06.006

Pajouhesh, H., & Lenz, G. R. (2005). Medicinal chemical properties of successful central nervous system drugs. *NeuroRx*, 2(4), 541–553. https://doi.org/10.1602/neurorx.2.4.541

Palama, T. L., Fock, I., Choi, Y. H., Verpoorte, R., & Kodja, H. (2010). Biological variation of Vanilla planifolia leaf metabolome. *Phytochemistry*, *71*(5–6), 567–573. https://doi.org/10.1016/j.phytochem.2009.12.011

Pelkonen, O., Turpeinen, M., Hakkola, J., Honkakoski, P., Hukkanen, J., & Raunio, H. (2008). Inhibition and induction of human cytochrome P450 enzymes: current status. *Archives of Toxicology*, 82(10), 667–715. https://doi.org/10.1007/s00204-008-0332-8

Perez De Souza, L., Alseekh, S., Brotman, Y., & Fernie, A. R. (2020). Network-based strategies in metabolomics data analysis and interpretation: from molecular networking to biological interpretation. *Expert Review of Proteomics*, *17*(4), 243–255. https://doi.org/10.1080/14789450.2020.1766975

Pérez-Silva, A., Nicolás-García, M., Petit, T., Dijoux, J. B., de los Ángeles Vivar-Vera, M., Besse, P., & Grisoni, M. (2021). Quantification of the aromatic potential of ripe fruit of Vanilla planifolia (Orchidaceae) and several of its closely and distantly related species and hybrids. *European Food Research and Technology*, *247*(6), 1489–1499. https://doi.org/10.1007/s00217-021-03726-w

Pérez-Silva, A., Odoux, E., Brat, P., Ribeyre, F., Rodriguez-Jimenes, G., Robles-Olvera, V., García-Alvarado, M. A., & Günata, Z. (2006). GC–MS and GC–olfactometry analysis of aroma compounds in a representative organic aroma extract from cured vanilla (Vanilla planifolia G. Jackson) beans. *Food Chemistry*, *99*(4), 728–735. https://doi.org/10.1016/j.foodchem.2005.08.050

Pinzi, L., & Rastelli, G. (2019). Molecular docking: Shifting paradigms in drug discovery. *International Journal of Molecular Sciences*, 20(18), 4331. https://doi.org/10.3390/ijms20184331

Quesada-Vázquez, S., Eseberri, I., Les, F., Pérez-Matute, P., Herranz-López, M., Atgié, C., Lopez-Yus, M., Aranaz, P., Oteo, J. A., Escoté, X., Lorente-Cebrian, S., Roche, E., Courtois, A., López, V., Portillo, M. P., Milagro, F. I., & Carpéné, C. (2024). Polyphenols and metabolism: from present knowledge to future challenges. *Journal of Physiology and Biochemistry*, *80*(3), 603–625. https://doi.org/10.1007/s13105-024-01046-7

Rodrigues, T., Sieglitz, F., Somovilla, V. J., et al. (2024). Machine Learning Uncovers Natural Product Modulators of the 5-Lipoxygenase Pathway and Facilitates the Elucidation of Their Biological Mechanisms. *ACS Chemical Biology*, *19*(1), 217–229. https://doi.org/10.1021/acschembio.3c00725

Rognan, D. (2010). Structure-Based Approaches to Target Fishing and Ligand Profiling. *Molecular Informatics*, *29*(3), 176–187. https://doi.org/10.1002/minf.200900081

Rowland, M., & Tozer, T. N. (2011). *Clinical Pharmacokinetics and Pharmacodynamics: Concepts and Applications* (4th ed.). Lippincott Williams & Wilkins.

Schneider, G. (2018). Automating drug discovery. *Nature Reviews Drug Discovery*, 17(2), 97–113. https://doi.org/10.1038/nrd.2017.232

Shaikh, F., Tai, H. K., Desai, N., & Siu, S. W. I. (2021). LigTMap: Ligand and structure-based target identification and activity prediction for small molecular compounds. *Journal of Cheminformatics*, *13*(1), 44. https://doi.org/10.1186/s13321-021-00523-1

Sheng, J., Zhang, S., Wu, L., Kumar, G., Liao, Y., GK, P., & Fan, H. (2022). Inhibition of phosphodiesterase: A novel therapeutic target for the treatment of mild cognitive impairment and Alzheimer's disease. *Frontiers in Aging Neuroscience*, *14*, 1019187. https://doi.org/10.3389/fnagi.2022.1019187

Śledź, P., & Caflisch, A. (2018). Protein structure-based drug design: from docking to molecular dynamics. *Current Opinion in Structural Biology*, 48, 93–102. https://doi.org/10.1016/j.sbi.2017.10.010

Tan, K. P., Singh, K., Hazra, A., & Madhusudhan, M. S. (2020). Peptide bond planarity constrains hydrogen bond geometry and influences secondary structure conformations. *Current Research in Structural Biology*, *3*, 1–8. https://doi.org/10.1016/j.crstbi.2020.11.002

Thompson, M. A., & Szabo, A. (2020). Endogenous N,N-dimethyltryptamine and the enigma of its function. *Frontiers in Neuroscience*, 14, 289. https://doi.org/10.3389/fnins.2020.00289

Treml, J., Nykodýmová, D., & Kubatka, P. (2025). Structure activity relationship of flavonoids as PPARγ agonists. *Phytochemistry Reviews*. https://doi.org/10.1007/s11101-025-10121-w

Wang, S., & Eisenberg, D. (2003). Crystal structures of a pantothenate synthetase from *M. tuberculosis* and its complexes with substrates and a reaction intermediate. *Protein Science*, *12*(5), 1097–1108. https://doi.org/10.1110/ps.0241803

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Wek, R. C. (2018). Role of eIF2α kinases in translational control and adaptation to cellular stress. *Cold Spring Harbor Perspectives in Biology*, 10(7), a032870. https://doi.org/10.1101/cshperspect.a032870

Wolfender, J. L., Nuzillard, J. M., van der Hooft, J. J. J., Renault, J. H., & Bertrand, S. (2019). Accelerating metabolite identification in natural product research: Toward an ideal combination of hyphenated and spectroscopic techniques. *Analytical Chemistry*, 91(1), 704–742. https://doi.org/10.1021/acs.analchem.8b05112

Zhang, H., Hassan, Y. I., Liu, R., Mats, L., Yang, C., Liu, C., & Tsao, R. (2020). Molecular mechanisms underlying the absorption of aglycone and glycosidic flavonoids in a Caco-2 BBe1 cell model. *ACS Omega*, *5*(19), 10782–10793. https://doi.org/10.1021/acsomega.0c00379

Capítulo 4 - Modelagem da distribuição de espécies revela *hotspots* de biodiversidade para baunilha selvagem (*Vanilla* spp.) de valor bioeconômico no Brasil

O Capítulo 4 desta tese corresponde a um artigo original a ser submetido na revista *Perspectives in Ecology and Conservation* (PECON) dedicado à modelagem da distribuição de espécies nativas do gênero *Vanilla* no Brasil, integrando ferramentas de modelagem ecológica como componente final de um pipeline de bioprospecção orientado por sustentabilidade socioambiental. Inserido no eixo integrador do trabalho, este capítulo expande a análise dos níveis químico e funcional, explorados nos capítulos anteriores, para a escala espacial e territorial, reconhecendo que a valorização bioeconômica da biodiversidade depende não apenas do potencial químico e bioativo das espécies, mas também da compreensão das condições ambientais que sustentam sua ocorrência, persistência e viabilidade de uso sustentável. A abordagem baseia-se na compilação e curadoria de registros de ocorrência georreferenciados de espécies de *Vanilla* nativas, combinados a conjuntos selecionados de variáveis ambientais climáticas, edáficas, topográficas e de sensoriamento remoto, integrados em modelos de distribuição construídos a partir de múltiplos algoritmos amplamente utilizados em biogeografia e conservação. A utilização de estratégias de modelagem em conjunto permite estimar áreas de adequabilidade ambiental com elevado desempenho preditivo, bem como identificar padrões de convergência e divergência ecológica entre as espécies avaliadas. Os resultados revelam distribuições potenciais heterogêneas, com espécies apresentando nichos ambientais contrastantes e graus distintos de restrição espacial, ao mesmo tempo em que evidenciam áreas de sobreposição entre regiões de alta adequabilidade ambiental e hotspots de riqueza específica. Essas áreas de convergência espacial destacam-se como regiões estratégicas onde biodiversidade, potencial produtivo e viabilidade ecológica coexistem, particularmente em porções da Mata Atlântica e zonas de transição associadas a sistemas florestais e agroflorestais. A análise espacial também evidencia lacunas relevantes na cobertura por unidades de conservação, indicando que parte significativa das áreas ambientalmente adequadas para espécies de alto valor bioeconômico permanece fora de regimes formais de proteção. Ao integrar esses resultados com os dados metabolômicos, funcionais e preditivos apresentados nos capítulos anteriores, o Capítulo 4 demonstra que a modelagem da distribuição de espécies não atua apenas como ferramenta de conservação, mas como instrumento estratégico de planejamento territorial, capaz de orientar decisões sobre conservação in situ, coleta responsável, cultivo sustentável e priorização de investimentos

em cadeias produtivas baseadas na biodiversidade. Um elemento central deste capítulo é a ênfase na disseminação e acessibilidade dos resultados gerados, materializada na proposta de uma plataforma de acesso aberto que reúne os mapas de adequabilidade ambiental, hotspots de biodiversidade e resumos interpretativos do potencial bioeconômico das espécies modeladas. Essa plataforma é concebida como um instrumento de tradução do conhecimento científico para diferentes públicos, incluindo gestores públicos, produtores rurais, comunidades locais, formuladores de políticas e investidores, permitindo que decisões relacionadas à conservação, uso sustentável e desenvolvimento econômico sejam embasadas em informação científica integrada e transparente. Ao tornar explícita a conexão entre valor químico, potencial bioativo e viabilidade ecológica, o capítulo reforça que a acessibilidade aos dados é condição essencial para a construção de cadeias de valor sustentáveis e socialmente legitimadas. Como conclusão, o Capítulo 4 encerra o fluxo metodológico da tese ao demonstrar que a integração entre metabolômica, predições funcionais *in silico* e modelagem ecológica permite não apenas revelar o valor oculto da biodiversidade de *Vanilla* no Brasil, mas também fornecer uma base espacial concreta para ações coordenadas de conservação, bioeconomia e planejamento territorial. Dessa forma, a tese culmina em uma abordagem verdadeiramente integrativa de bioprospecção, na qual dados químicos, biológicos e ambientais convergem para sustentar estratégias de uso sustentável da biodiversidade alinhadas às demandas científicas, sociais e econômicas contemporâneas.

Species distribution modeling reveals biodiversity hotspots for bioeconomically valuable wild vanilla (*Vanilla* spp.) in Brazil

Joana Paula da Silva Oliveira; Maria Lucia Lorini; Andrea Furtado Macedo; Rodrigo Tardin

Introduction

Vanilla, extracted from the cured fruits of orchid species within the genus *Vanilla*, stands as one of the most economically significant spices globally, contributing to an industry valued at approximately USD 724.5 million as of 2025 projections (de Oliveira et al., 2022; Pérez-Silva et al., 2025). The spice's high commercial demand stems from its versatile applications in food, beverages, cosmetics, and pharmaceuticals, positioning it as the second most expensive spice after saffron, with production challenges leading to price fluctuations that underscore its vulnerability in global supply chains (Barragán et al., 2024; Watteyn et al., 2022). Predominantly sourced from *Vanilla planifolia* and *Vanilla* × *tahitensis*, commercial vanilla cultivation is concentrated in non-native regions such as Madagascar, which supplies over 80% of the world's output, often under monoculture systems that exacerbate environmental degradation and economic instability for smallholder farmers (Martin et al., 2020; Watteyn et al., 2020).

Crop wild relatives (CWRs) of vanilla represent an invaluable genetic repository, encompassing undomesticated species that harbor traits for enhanced resilience, such as drought tolerance, disease resistance, and diverse aromatic profiles, which are essential for breeding programs aimed at sustainable agriculture (Castañeda-Álvarez et al., 2016; Khoury et al., 2021). In the Neotropics, particularly South America, vanilla CWRs like *V. chamissonis*, *V. cribbiana*, and *V. odorata* exhibit chemical diversity, including elevated levels of vanillin and related compounds, surpassing those in commercial cultivars and offering potential for flavor diversification and bioprospecting (da Silva Oliveira et al., 2025; Pansarin, 2025a). These relatives, often restricted to fragmented habitats in biomes such as the Atlantic Forest and Cerrado, face extinction risks from deforestation and overharvesting, with fewer than 10% of *Vanilla* species assessed by the IUCN, many classified as endangered or critically endangered (Flanagan et al., 2022; Goettsch et al., 2021). The integration of CWRs into cultivation systems, as demonstrated in agroforestry trials in Costa Rica, can mitigate these threats by promoting in situ conservation while providing alternative income sources for local communities, thus aligning with global strategies for agrobiodiversity preservation (Watteyn et al., 2022; Karremans et al., 2025). Furthermore, studies on Mesoamerican CWRs highlight their role in maintaining evolutionary lineages that could bolster climate-resilient vanilla production, emphasizing the need for ex situ collections to safeguard genetic variability (Chambers et al., 2021; Serna-Sánchez et al., 2025).

Species distribution modeling (SDM) emerges as a pivotal tool for the valorization and conservation of vanilla CWRs, leveraging statistical correlations between species occurrences and environmental variables to forecast habitat suitability and guide resource management (Guisan et al., 2017; Watteyn et al., 2020). For example, by incorporating bioclimatic data, SDMs have revealed potential shifts in vanilla distributions under climate change scenarios, such as reduced suitability in traditional Mexican growing areas for *V. planifolia*, projecting up to 17% habitat loss by 2070 under high-emission pathways (Armenta-Montero et al., 2022). In conservation contexts, SDM facilitates the identification of biodiversity hotspots and priority zones for protected areas, as seen in analyses of Neotropical orchids where models predict declines in pollinator overlap with vanilla species, threatening reproductive success (Ramírez-Mosqueda et al., 2025; Pansarin, 2025b). Ensemble SDMs, which aggregate multiple algorithms for improved accuracy, have been applied to map suitable cultivation sites for vanilla CWRs in Costa Rica, integrating land-use data to propose joint sparing-sharing strategies that balance intensive farming with habitat preservation (Watteyn et al., 2020; Nadal et al., 2025). Such models also support sustainable agroforestry initiatives, where vanilla integration enhances ecosystem services like biodiversity maintenance and carbon sequestration, particularly in regions with high endemism (Martin et al., 2020; Maurya et al., 2025).

Previous studies revealed a Mesoamerican bias, with comprehensive revisions of Mexican and Central American species documenting 15 taxa and their phylogenetic relationships via ITS nuclear ribosomal DNA, highlighting ecological dependencies on specific pollinators and habitats (Soto Arenas and Dressler, 2010; Karremans et al., 2020). In contrast, South American distributions, encompassing over 40 species with centers of diversity in Brazil, remain underexplored, often limited to taxonomic reappraisals that clarify species boundaries but lack integrative modeling for conservation planning (Pansarin, 2025c; Oliveira et al., 2024). Recent advances, such as metabolomic profiling of Brazilian species like *V. bahiana*, *V. chamissonis* and *V. cribbiana*, underscore their aromatic potential, yet SDM applications in this region are nascent, overlooking synergies that could potentially highlight vulnerability to threats (e.g. deforestation) or how current protection measures, such as protected areas, covers their distribution (Franco et al., 2024; Silva et al., 2023). Improvements could involve multi-species ensemble models incorporating fine-scale variables like soil type and pollinator distributions, as demonstrated in broader orchid studies, to better predict richness hotspots and inform national conservation strategies (Guisan et al., 2017; Pansarin and Menezes, 2023). Our study advances the field by applying SDM to Brazilian vanilla CWRs, mapping ecological niches, assessing overlap with protected areas, and evaluating sustainable use potential, thereby bridging gaps in Neotropical biodiversity research and supporting global vanilla supply chain resilience (Karremans et al., 2025; Flanagan et al., 2022).

We address these conservation challenges by applying species distribution modeling to native Brazilian *Vanilla* species with high aromatic potential. Specifically, we (i) model the distribution of economically important taxa, (ii) assess environmental factors shaping habitat

suitability, (iii) construct species richness maps highlighting areas of overlapping diversity, and (iv) identify priority zones for the conservation of *Vanilla* species in Brazil of commercial interest due to their aromatic descriptors. By combining ecological modeling with the emerging metabolomic and sensory data on Brazilian *Vanilla*, we aim to bridge biodiversity conservation with strategies for sustainable cultivation and value chain diversification.

Material and Methods

Species data. We compiled occurrence records for *Vanilla cribbiana*, *V. chamissonis*, *V. bahiana* and *V. phaeantha* from GBIF, SpeciesLink, Reflora, and SIBBR platforms with time span from 1970 to 2025. Records were read, species names standardized, and rows with missing coordinates removed prior to quality control. To reduce spatial sampling bias and autocorrelation, we applied a 10-km thinning distance using *spThin* package v4.3.3, since distances approximating predictor resolution reduces spatial autocorrelation and sampling bias without excessive data loss (Boria et al., 2014; Aiello-Lammens et al., 2015). Occurrence records compiled and used for modeling are publicly available at https://github.com/joanapaulaso/mde_vanilla (da Silva Oliveira, 2025).

Accessible area (M) and study extent. To approximate the species' M (sensu BAM framework), we constructed a convex hull around thinned presences and buffered it by 50 km, a distance commonly applied to approximate accessible area (M) for plant species with regional distributions in the absence of strong dispersal barriers, following the BAM framework (Barve et al., 2011; Peterson et al., 2011). The buffer was intersected with a "South–Central–North America" land polygon to remove ocean and out-of-scope areas and served as the spatial mask for predictors and background sampling.

Environmental predictors. Climatic variables (average temperature, precipitation, solar radiation, wind speed, water vapor pressure, and 19 bioclimatic layers BIO1–BIO19) were obtained from WorldClim v2.1 averaged for the period 1970-2000 (www.worldclim.org). Edaphic variables (available water capacity, cation exchange capacity, clay, coarse fragments, nitrogen, organic carbon, pH, sand, and silt) were sourced from SoilGrids v2.0 (www.soilgrids.org). NDVI layers (mean 2003–2007, cumulative, minimum, and seasonal) were obtained from the Silvis Lab platform (Spatial Analysis For Conservation and Sustainability, http://silvis.forest.wisc.edu/data/dhis). Topographic variables (elevation, slope, aspect, topographic position index, vector roughness measure, composite topographic index, and Heat Load Index) and cloud cover variables (January mean, June mean, and intra-annual variability) were obtained from EarthEnv (https://www.earthenv.org/topography and https://www.earthenv.org/cloud). Aridity index was obtained from the CGIAR-CSI Global Aridity database (https://cgiarcsi.community/data/global-aridity-and-pet-database). All layers were then clipped by the trimmed M buffer and resampled by bilinear interpolation to a common resolution of 30 arc-seconds (~1 km) prior to stacking into a multi-layer

SpatRaster for modeling. To avoid correlation and collinearity among variables, we computed a Pearson correlation matrix and Variance Inflation Factor (VIF), removing those exceeding an absolute pairwise correlation of 0.7 ($|r| \geq 0.7$) VIF > 5.

Modeling framework. We fit correlative species distribution models using BIOMOD2 v4.2-6-2 (Guéguen et al., 2025). Pseudo-absences (PAs) were generated within M with 5 PA replicates and species-specific numbers of pseudo-absences (*V. bahiana*: 7,000; *V. chamissonis*: 5,500; *V. cribbiana*: 7,000; *V. phaeantha*: 10,000; *V. pompona*: 10,000), the Surface Range Envelope (SRE) strategy. Models were calibrated with six algorithms: Generalized Linear Models (GLM), Generalized Additive Models, Gradient Boosting Machines (GBM), Random Forest (RF), MaxEnt and XGBOOST.

Cross-validation (CV) followed a repeated random holdout using five replicates with 70% of data for training and 30% for testing, stratified by prevalence. Variable importance was computed during fitting via permutation procedure (Guisan et al. 2017). Following the initial round with non-correlated predictors, we ran a reduced predictor set for the final modeling round based on the most important variables (same algorithms and CV).

Model evaluation, selection, and ensembles. Models were evaluated with the True Skill Statistic (TSS) at each CV replicate. Individual algorithms meeting a TSS ≥ 0.7 were retained for ensembling using (i) simple mean (EMmean) and (ii) committee averaging (EMca), stratified by PA set and replicate. This ensemble strategy is adapted from prior SDM applications in *Vanilla* conservation that combine algorithmic diversity with threshold-based selection and averaging to stabilize predictions (Watteyn et al., 2020).

Projections and thresholds. Final ensembles were projected over M using the reduced predictor set. We requested binary outputs using the TSS-maximizing threshold. Binary maps reflect pixels predicted suitable/unsuitable under the ensemble using TSS as the decision criterion.

Conservation areas designation. Continuous maps of predicted habitat suitability for each of the five target species were generated using ensemble species distribution models (SDMs) and binarized to presence–absence maps by applying thresholds of TSS ≥ 0.7 and suitability $\geq 80\%$, ensuring high model performance and minimizing commission errors. These binary maps were then summed across species to create species richness maps, highlighting areas of potential species overlap and co-occurrence. The richness maps were overlaid and compared with the vector layer of protected areas coverage from the Polígono CNUC 2025_08 shapefiles (active Conservation Units registered in the Brazilian National Register of Conservation Units as of August 2025; cataloged and updated on 03/09/2025, sourced from the "Portal de Dados Abertos") (Watteyn et al., 2020). High-richness hotspots (areas suitable for ≥ 4 simultaneous species) were designated as priority areas for conservation and sustainable cultivation planning. This approach follows recommendations that TSS

functions as a prevalence-independent accuracy metric appropriate for binarizing SDMs (Allouche et al., 2006). It also aligns with evidence demonstrating the value of ensemble consensus in strengthening conservation prioritization outputs (Meller et al., 2014). Furthermore, it accounts for the documented sensitivity of richness hotspot identification and protection assessments to input resolution and classification error, necessitating the use of thresholded, high-accuracy suitability maps (McKerrow et al., 2018).

Results

1. Model performance and validation

Our models exhibited strong predictive performance across the five *Vanilla* species, with TSS by mean ensembles (EMmean) consistently achieving the highest validation metrics in all cases (AUC ROC range: 0.93–0.98; TSS range: 0.73–0.87). The species *V. phaeantha* showed the best overall fit (EMmean validation AUC ROC = 0.98, TSS = 0.87) with high sensitivity (ROC: 99.29%, TSS: 99.19%) and specificity (ROC: 92.43%, TSS: 92.41%), followed closely by *V. bahiana* (EMmean AUC ROC = 0.96, TSS = 0.84). In contrast, *V. pompona* displayed the lowest performance (EMmean AUC ROC = 0.93, TSS = 0.73), though still indicating good discrimination. Individual algorithms varied, with RF frequently attaining perfect or near-perfect sensitivity and specificity at optimized thresholds but markedly lower validation TSS (0.06–0.70), suggesting potential overfitting; MaxEnt and GBM provided balanced results across species. All ensembles outperformed single models in validation AUC ROC and TSS (Table 1).

Table 1. Performance metrics for individual algorithms and the EMmean ensemble across five *Vanilla* species (*V. cribbiana*, *V. chamissonis*, *V. pompona*, *V. bahiana*, *V. phaeantha*). Metrics include ROC- and TSS-optimized cutoffs, sensitivity (%), specificity (%), and validation AUC ROC (0–1) and TSS (0–1). Validation metrics represent overall model performance on the hold-out validation dataset under ROC and TSS optimization criteria, respectively.

Model	Metric	Eval.	*V. cribbiana*	*V. chamissonis*	*V. pompona*	*V. bahiana*	*V. phaeantha*
GAM	Cutoff	ROC	502.69	411.90	495.76	331.76	327.68
		TSS	502.71	412.81	496.97	331.10	324.10
	Sensitivity	ROC	93.67	83.07	74.99	92.82	93.91
		TSS	94.38	82.64	74.75	92.73	94.56
	Specificity	ROC	88.80	80.79	77.84	91.16	78.61

		TSS	86.15	80.84	77.55	91.04	75.49
	Validation	ROC	0.89	0.88	0.80	0.93	0.91
		TSS	0.63	0.58	0.48	0.76	0.64
GBM	Cutoff	ROC	548.24	484.97	404.74	459.05	425.29
		TSS	544.06	477.16	404.06	454.26	420.94
	Sensitivity	ROC	98.31	87.44	97.27	97.56	98.14
		TSS	100.00	86.84	96.88	97.58	96.70
	Specificity	ROC	96.65	89.15	89.44	94.47	92.03
		TSS	93.44	88.97	89.46	94.27	92.16
	Validation	ROC	0.84	0.90	0.90	0.95	0.97
		TSS	0.45	0.62	0.66	0.76	0.74
GLM	Cutoff	ROC	502.18	383.56	514.27	397.29	450.29
		TSS	497.29	380.84	513.10	398.68	437.52
	Sensitivity	ROC	90.94	93.07	78.61	97.03	99.10
		TSS	91.40	92.89	78.30	96.93	99.39
	Specificity	ROC	89.00	81.64	81.23	92.06	92.05
		TSS	85.93	81.50	81.18	92.06	88.94
	Validation	ROC	0.77	0.90	0.85	0.94	0.97
		TSS	0.51	0.62	0.54	0.76	0.81
MaxEnt	Cutoff	ROC	315.52	187.29	277.73	206.10	105.40
		TSS	310.19	178.61	274.23	204.13	108.16
	Sensitivity	ROC	94.78	88.30	82.30	95.98	96.55
		TSS	94.78	88.54	82.30	95.92	95.89
	Specificity	ROC	94.53	87.68	90.95	92.63	91.19
		TSS	94.17	87.03	90.66	92.40	91.39
	Validation	ROC	0.88	0.92	0.88	0.95	0.97
		TSS	0.63	0.67	0.65	0.77	0.81

RF	Cutoff	ROC	389.23	426.29	291.90	450.58	405.65
		TSS	388.52	425.87	290.45	450.63	405.35
	Sensitivity	ROC	100.00	100.00	100.00	100.00	100.00
		TSS	100.00	100.00	100.00	100.00	100.00
	Specificity	ROC	100.00	100.00	99.99	100.00	100.00
		TSS	100.00	100.00	99.99	100.00	100.00
	Validation	ROC	0.89	0.92	0.90	0.97	0.94
		TSS	0.06	0.24	0.17	0.70	0.11
XGBOOST	Cutoff	ROC	762.02	621.50	459.40	672.89	731.55
		TSS	761.37	618.65	456.61	670.73	730.60
	Sensitivity	ROC	98.16	98.82	88.87	99.96	99.67
		TSS	98.16	98.74	88.88	99.96	100.00
	Specificity	ROC	99.76	98.03	79.87	99.60	98.30
		TSS	99.76	98.03	79.84	99.57	95.58
	Validation	ROC	0.77	0.86	0.83	0.94	0.88
		TSS	0.15	0.45	0.53	0.68	0.56
EMmean	Cutoff	ROC	450.03	239.00	420.67	362.07	347.71
		TSS	444.67	235.40	425.00	358.37	346.54
	Sensitivity	ROC	96.00	86.67	97.76	98.95	99.29
		TSS	96.00	86.67	97.14	99.03	99.19
	Specificity	ROC	92.09	88.12	88.36	93.92	92.43
		TSS	91.92	87.80	88.65	93.76	92.41
	Validation	ROC	0.93	0.94	0.93	0.96	0.98
		TSS	0.83	0.76	0.73	0.84	0.87

2. Environmental influence on *Vanilla* species

The modeled *Vanilla* species showed clear ecological differentiation driven by integrated effects of thermal buffering, moisture regime, light availability, and exposure, rather than by

individual climatic variables in isolation (Figure 1 and Table 2). *V. cribbiana* was predicted to occur in areas with low diurnal thermal amplitude and intermediate isothermality, indicating preference for semi-buffered, humid forest mosaics rather than highly stable or strongly exposed environments. At the same time, its suitability increased markedly with wettest-month precipitation (from ~0.02 at ~70 mm to ~0.43 at ~875 mm) and mean surface shortwave radiation (>0.50 above ~21,000 kJ m^{-2} day^{-1}), while declining with increasing wind speed, suggesting ecological affinity for humid, sheltered but high-radiance habitats such as forest edges, canopy openings, or early-successional forests.

V. chamissonis occupied an intermediate ecological position strongly shaped by coastal exposure. It showed optimal suitability under moderate diurnal thermal range (~8.8–9.0 °C), low isothermality (~46–48), intermediate wet-season rainfall (~200–240 mm), and moderate, persistent wind regimes (~1.5 m s^{-1}), reflecting adaptation to maritime climates with continuous airflow and high atmospheric humidity. Its positive response to diffuse light and increasing cloud fraction during the dry season further indicates ecological affinity for coastal vegetation where direct irradiance is frequently attenuated but overall light availability remains high.

In contrast, *V. phaeantha* displayed weak structuring along thermal axes and low overall suitability responses, indicating that temperature stability is not a primary constraint for the species. Instead, its niche was characterized by tolerance to moderate seasonality and a marked negative association with high wet-season precipitation, with suitability declining from ~0.20 at ~50–80 mm to <0.01 above ~300 mm. This pattern suggests specialization for comparatively drier or seasonally open forest environments, where rooting microclimate and soil drainage likely override macroclimatic thermal controls.

V. bahiana was primarily structured by exposure and light regime rather than moisture. It showed increasing suitability with wind speed (from ~0.02 at ~1.0 m s^{-1} to ~0.54 at ~3.1 m s^{-1}), strong negative association with cloud cover during the dry season, and declining suitability under increasing temperature seasonality, indicating preference for highly exposed, sun-dominated coastal plateaus or open formations with stable thermal conditions and intense irradiance. Wet-season precipitation exerted little influence on its distribution, suggesting hydric tolerance within coastal systems.

V. pompona emerged as the broadest ecological generalist among the species. It tolerated a wide range of diurnal thermal amplitudes, including both buffered (~5–9 °C) and moderately variable (~16–18 °C) regimes, and exhibited a pronounced unimodal response to radiation, with highest suitability under intermediate light levels (~14,000–18,500 kJ m^{-2} day^{-1}). This pattern is consistent with occupation of partially shaded understory or agroforestry environments, where structural buffering, water storage, and vine architecture mitigate both thermal and hydric variability.

Collectively, these results indicate that habitat differentiation among *Vanilla* species is structured primarily along a gradient of microclimatic buffering versus exposure, integrating

moisture availability, light regime, and wind stress, rather than by single climatic variables. This integrated ecological filtering explains the coexistence of closely related species across contrasting forest interiors, semi-open mosaics, and highly exposed coastal environments.

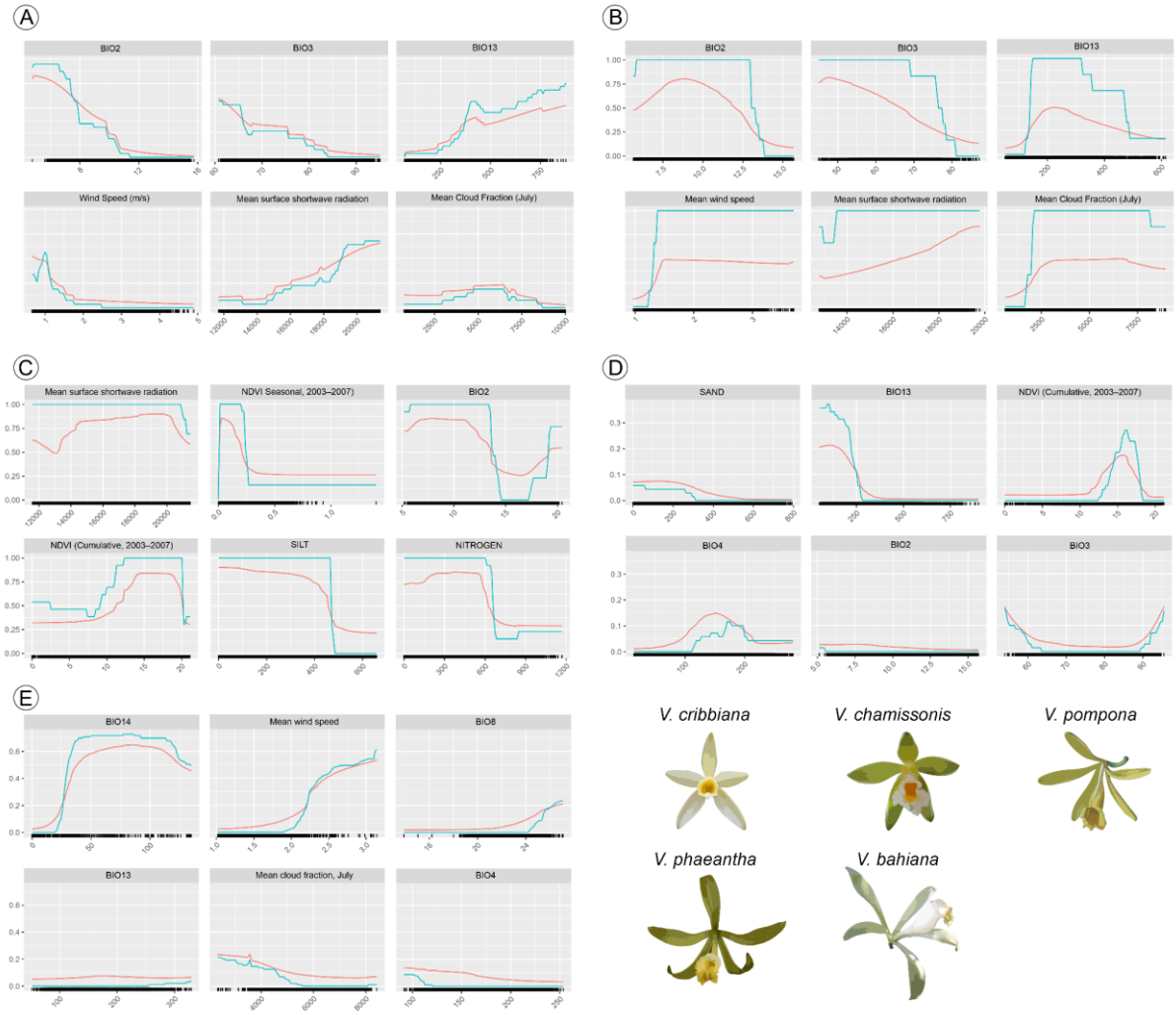


Figure 1. Response curves of key environmental predictors for five *Vanilla* species: (A) *V. cribbiana*, (B) *V. chamissonis*, (C) *V. pompona*, (D) *V. phaeantha*, and (E) *V. bahiana*. Panels display selected bioclimatic (BIO2–BIO19), edaphic (e.g., soil pH, sand, silt, nitrogen), atmospheric (e.g., mean wind speed, mean cloud fraction in July), remote sensing (e.g., mean surface shortwave radiation, NDVI seasonality and cumulative NDVI for 2003–2007), and topographic (elevation) variables. Blue lines represent the committee averaging ensemble models and red lines represent the ensemble models by mean, both weighted by TSS. Illustrations of each species are shown in the bottom-right corner of the panels.

Table 2. Mean (± SD) permutation importance of environmental and climatic predictors across four *Vanilla* species models. BIO2 = Mean diurnal temperature range (°C); BIO3 = Isothermality; BIO4 = Temperature seasonality; BIO5 = Maximum temperature of warmest month; BIO8 = Mean temperature of wettest quarter; BIO13 = Precipitation of wettest month; BIO14 = Precipitation of driest month; BIO18 = Precipitation of warmest quarter; BIO19 = Precipitation of coldest quarter; NDVI = Normalized Difference Vegetation Index; HLI = Heat Load Index; CEC = Cation Exchange Capacity; SAND, CLAY = soil texture fractions; *m* = meters. Values represent the mean ± standard deviation of predictor importance aggregated across ensemble model replicates for each species (*V. cribbiana*, *V. chamissonis*, *V. bahiana*, *V. phaeantha*).

Predictor	*V. cribbiana*	*V. chamissonis*	*V. bahiana*	*V. phaeantha*	*V. pompona*
BIO2	0.3 ± 0.2	0.4 ± 0.2	—	0.1 ± 0.1	0.3 ± 0.1
BIO3	0.2 ± 0.2	0.2 ± 0.1	—	0.1 ± 0.1	—
BIO4	0.3 ± 0.2	—	0.0 ± 0.0	0.3 ± 0.2	—
BIO5	—	—	0.0 ± 0.0	—	—
BIO8	—	—	0.1 ± 0.1	—	—
BIO13	0.3 ± 0.2	0.2 ± 0.1	0.0 ± 0.0	0.5 ± 0.2	—
BIO14	—	—	0.5 ± 0.2	—	—
BIO18	0.3 ± 0.2	—	—	—	—
BIO19	—	0.2 ± 0.1	—	—	0.1 ± 0.1
Mean wind speed	0.2 ± 0.2	0.1 ± 0.1	0.2 ± 0.1	—	—
Mean surface shortwave radiation	0.2 ± 0.2	0.1 ± 0.1	—	—	0.3 ± 0.1
NDVI (Seasonality, 2003–2007)	0.2 ± 0.2	—	—	—	0.5 ± 0.1
NDVI (Cumulative, 2003–2007)	—	—	—	0.4 ± 0.2	0.2 ± 0.1

Mean cloud fraction, January	—	—	0.0 ± 0.1	—	—
Mean cloud fraction, July	0.1 ± 0.1	0.1 ± 0.1	0.0 ± 0.0	—	—
Soil pH	—	0.3 ± 0.1	—	—	—
Cation Exchange Capacity (Soil CEC)	—	—	0.0 ± 0.0	—	—
Clay	—	0.1 ± 0.1	—	—	—
SAND	—	—	—	0.5 ± 0.2	—
Elevation (m)	0.2 ± 0.1	0.0 ± 0.1	—	—	—
Heat Load Index (HLI)	—	0.1 ± 0.0	—	—	—
SILT	—	—	—	—	0.2 ± 0.1
NITROGEN	—	—	—	—	0.2 ± 0.1
Intra-annual Variability in Cloud Fraction	—	—	—	—	0.1 ± 0.1
SOC	—	—	—	—	0.1 ± 0.1

3. Spatial patterns of suitable habitat and uncertainty measurements (per species)

Among the five modeled Vanilla species, *V. pompona* exhibits the largest predicted suitable habitat at 546,575 km², spanning diverse biomes from northeastern Atlantic Forest remnants to extensive Amazonian forests, Caatinga shrublands, and Cerrado savannas, with the highest protection coverage at 31.53% through a mix of environmental protection areas, forests, and parks emphasizing multi-ecosystem adaptability in humid to semi-arid tropical environments (Figure 2); in contrast, *V. cribbiana* has the smallest habitat at 14,541 km², primarily concentrated in humid tropical Atlantic Forest with minor Amazonian extensions, achieving 26.57% protection mainly via environmental protection areas and wildlife refuges, reflecting a more restricted coastal focus. *V. phaeantha* ranks second in size at 314,855 km² but has the lowest protection at only 6.21%, distributed across central Cerrado savannas, transitional Atlantic Forest zones, and seasonally dry tropical forests, highlighting vulnerability in biodiverse but fragmented inland ecosystems (Figure 2). *V. chamissonis* covers 73,960 km² with 30.55% protection, centered in southeastern and southern Atlantic Forest with Cerrado transitions, suited to subtropical humid forests and urban-proximate areas, while *V. bahiana* spans 41,408 km² with 22.35% protection, focused on northeastern

and southeastern coastal Atlantic Forest including restinga and mangroves, indicating a strong affinity for marine-influenced humid sandy environments (Figure 2). Overall, protection levels vary from over 30% in broadly distributed species like *V. pompona* and *V. chamissonis* to under 10% in Cerrado-dominant *V. phaeantha*, underscoring conservation priorities for species with lower coverage despite larger habitats, with all sharing ties to Atlantic Forest but differing in biome breadth and environmental preferences from coastal humidity to inland seasonality.

Uncertainty analysis through the committee averaging models shows that species with extensive distributions across multiple biomes exhibit greater model disagreement in transitional and peripheral zones, reflecting potential environmental variability or data gaps, whereas those with more confined ranges in coastal or humid tropical environments demonstrate stronger consensus on habitat suitability and unsuitability. This analysis underscores the robustness of ensemble predictions in core areas while identifying higher predictive variability in expansive or fragmented ecosystems, aiding in prioritizing conservation efforts where model agreement is weakest.

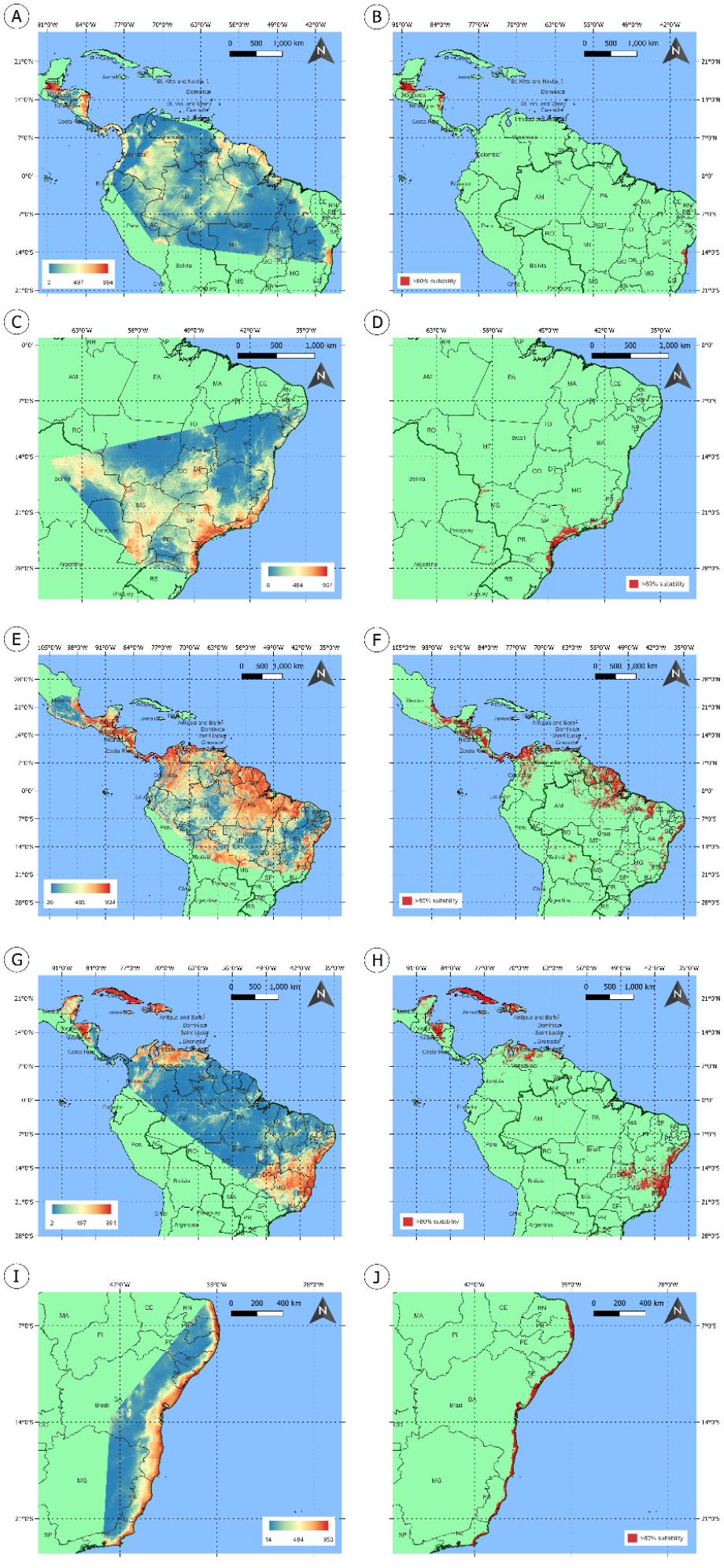

A
B
C
D
E
F
G
H
I
J
0 500 1,000 km
0 200 400 km
>80% suitability

Figure 2. TSS-weighted prediction and ensemble habitat suitability for Vanilla spp. across Central and South America. On the left side are the corresponding continuous, TSS-weighted ensemble suitability surface. Rightside: TSS-binarized predicted habitat suitability for *Vanilla* spp. Pixels in red are classified as suitable, with > 80% suitability, after applying the TSS threshold = 0.7 to the ensemble prediction; The map is restricted to the accessible area (*M*) used for modeling. Resolution ≈ 30″ (~1 km at the equator); datum WGS84. (A) and (B) *V. cribbiana*; (C) and (D) *V. chamissonis*; (E) and (F) *V. pompona*; (G) and (H) *V. phaeantha;* (I) and (J) *V. bahiana*.

4. Conservation-relevant synthesis, cross-species congruence and richness hotspots

The geospatial analysis identified suitability hotspots for five *Vanilla* species (*V. cribbiana*, *V. chamissonis*, *V. pompona*, *V. bahiana*, and *V. phaeantha*) based on ensemble species distribution models, with suitability thresholds defined by a True Skill Statistic (TSS) ≥ 0.7 and predicted presence probabilities ≥ 80% (Figure 3). These hotspots were delineated by overlaying binary suitability maps, where richness levels represent the number of species for which an area is simultaneously suitable: richness 1 indicates suitability for one species, up to richness 4 for four species co-occurring. No areas exhibited suitability for all five species, as no spatial overlap was detected between *V. cribbiana* and *V. chamissonis* (Figure 3). The total hotspot areas decreased with increasing richness, ranging from 1,930,689.72 km² for richness 1 to 988.26 km² for richness 4 (Figure 3). Overlap with Brazilian protected areas (PAs), encompassing a total dissolved area of 1,570,601.22 km², revealed varying degrees of coverage for these hotspots. Overall, approximately 20.4% of richness 4 hotspots (201.51 km² overlap) and 19.1% of richness 3 hotspots (2,645.62 km² overlap) were covered by PAs, indicating moderate protection for higher-diversity areas.

Hotspots suitable for four *Vanilla* species simultaneously (richness = 4), totaling 988.26 km², showed moderate protection, with approximately 20.4% of their area (201.51 km²) overlapping Brazilian protected areas (PAs). This coverage was strongly dominated by Sustainable Use categories, which accounted for 16.81% of the hotspot area (166.13 km²), whereas Strict Protection PAs represented only 3.63% (35.88 km²). Among management types, Environmental Protection Areas and Extractive Reserves were the most frequent categories contributing to protection, while strictly restrictive units such as Parks and Biological Reserves played a minor role (Supplementary material S3), and several categories showed no overlap at all. Overall, these patterns indicate that multi-species *Vanilla* hotspots are predominantly encompassed by less restrictive, sustainable-use regimes rather than fully protected areas, underscoring the relevance of flexible conservation strategies for these regions and suggesting that expansion or reallocation of stricter protection categories could improve long-term conservation effectiveness.

Areas simultaneously suitable for four *Vanilla* species formed well-defined coastal hotspots in northeastern Brazil, totaling 955.37 km², with a strong concentration along the Atlantic Forest domain (Fig. 7). These hotspots were overwhelmingly dominated by Bahia, which

accounted for 91.8% of the total area, while Espírito Santo contributed a smaller fraction (8.1%) and Sergipe only marginally (0.1%), indicating a pronounced regional clustering rather than a diffuse coastal distribution (Fig. 7). Two species combinations structured this pattern: hotspots including *V. cribbiana* represented the majority of the area (89.8%) and were restricted to southern Bahia, consistent with the species' ecological association with humid, semi-open Atlantic Forest mosaics, whereas hotspots including *V. chamissonis* comprised a smaller fraction (10.2%) and were concentrated mainly in Espírito Santo, with limited extensions into adjacent coastal areas (Fig. 7). This spatial segregation suggests that shared suitability among species is shaped by broad-scale biogeographic and ecological filters rather than fine-scale administrative boundaries. Priority areas for conservation therefore align with these coastal Atlantic Forest hotspots (Fig. 7), which were delineated using high-confidence ensemble predictions (TSS ≥ 0.7; predicted presence ≥ 80%), reinforcing their relevance for multi-species persistence under ongoing pressures from habitat fragmentation and climate change.

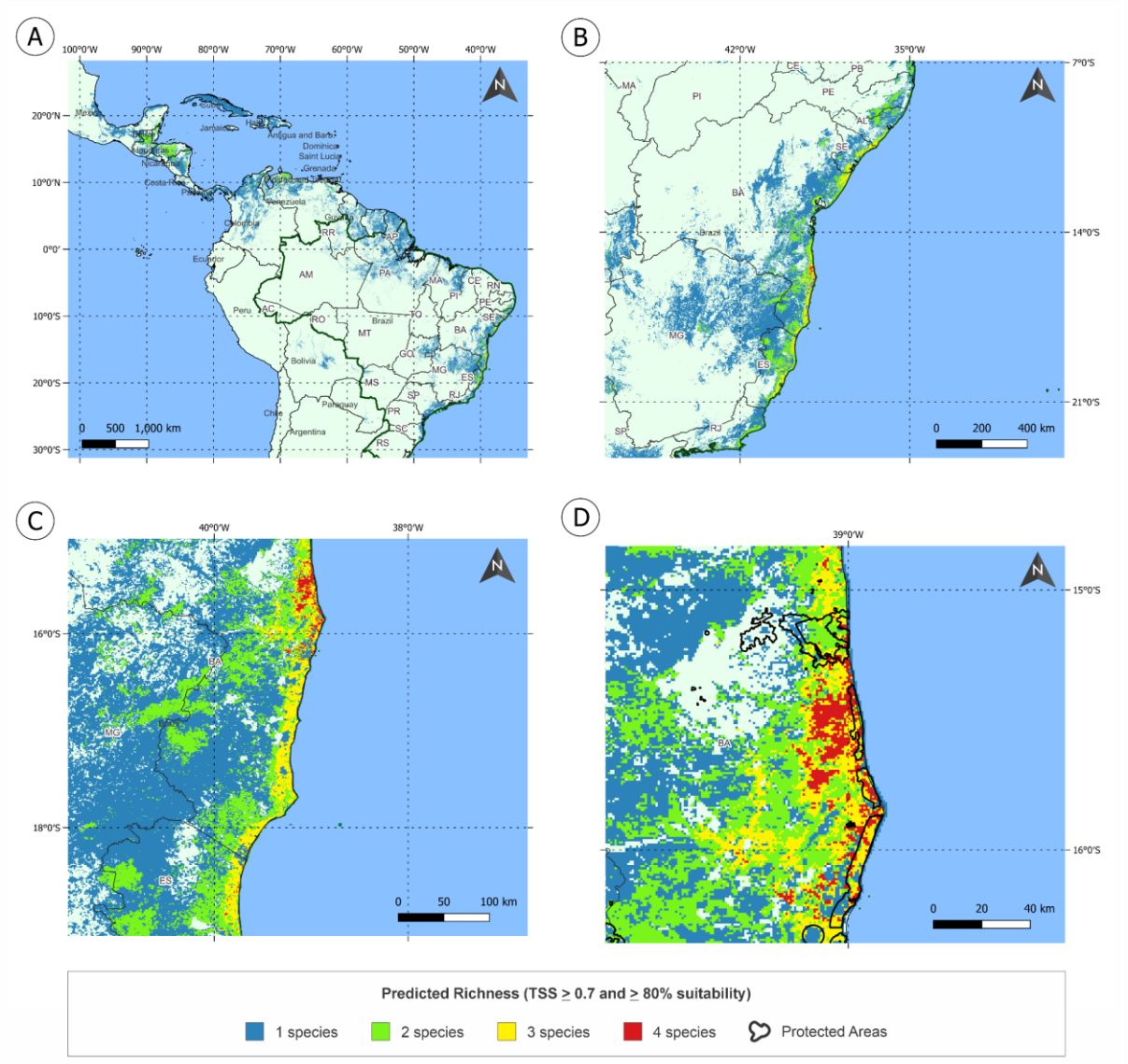


Figure 3. Potential richness maps for *Vanilla* species (*V. cribbiana*, *V. chamissonis*, *V. pompona, V. bahiana*, and *V. phaeantha*), based on TSS ≥ 0.7 and ensemble model suitability of ≥ 80%. A) Full extent respective to the sum of all five models. B) Zoom into the coastal region of Brazil, where hotspots for ≥ 2 species were mostly detected. C) Zoom into the Southern Bahia and Northern Espírito Santo, where hotspots for ≥ 3 species were mostly detected. D) Widest hotspot area, simultaneously suitable for 4 species (*V. cribbiana*, *V. pompona, V. bahiana*, and *V. phaeantha*) in Southern Bahia, and the overlap of this hotspot with PAs.

Discussion

Our ensemble models demonstrated high predictive accuracy, indicating that the selected set of environmental predictors captured key ecological constraints governing the distribution of *Vanilla* species. The superior performance of the EMmean relative to individual algorithms reflects the capacity of ensemble approaches to reduce algorithm-specific biases and uncertainty, producing more stable estimates of species–environment relationships (Araújo & New, 2007; Thuiller et al., 2009). This robustness allows ecological interpretation of response curves and spatial patterns with a high degree of confidence.

The modeled response curves revealed pronounced niche differentiation among species along integrated axes of thermal variability, moisture availability, light regime, and microclimatic exposure. Temperature seasonality (BIO4) emerged as a major structuring factor, but species responses diverged markedly. *V. bahiana* showed declining suitability under increasing seasonality, consistent with adaptation to thermally stable coastal systems, whereas *V. cribbiana* exhibited increased suitability under strong seasonal oscillations, reflecting specialization for inland or semi-deciduous forest mosaics where phenology is tightly linked to annual climatic cycles. *V. phaeantha* occupied an intermediate position, with a unimodal response suggesting tolerance to moderate seasonality but reduced performance at both extremes. These contrasting patterns indicate that BIO4 acts not as a universal constraint, but as a species-specific ecological filter reflecting possibly distinct evolutionary and physiological strategies.

Hydric variables further reinforced niche partitioning, but again with divergent responses driven by ecological context rather than methodological noise. The strong positive association of *V. cribbiana* with wettest-month precipitation indicates reliance on persistently humid environments, whereas the unimodal response of *V. chamissonis* suggests sensitivity to both drought stress and prolonged saturation typical of highly exposed coastal habitats. In contrast, *V. phaeantha* exhibited declining suitability with increasing wet-season precipitation, pointing to adaptation to better-drained or seasonally drier environments. This pattern helps explain discrepancies with studies reporting precipitation-driven expansion for *V. phaeantha*, as those analyses relied primarily on annual or seasonal precipitation metrics

(BIO12, BIO14, BIO15) and did not explicitly capture peak wet-season effects or soil–atmosphere interactions (Watteyn et al., 2025).

A key contribution of our study lies in demonstrating that exposure versus buffering constitutes a dominant ecological axis structuring *Vanilla* habitats, an effect that becomes apparent only when atmospheric and radiative variables are incorporated. Declining suitability of *V. cribbiana* under increasing wind speed reflects dependence on sheltered understory or forest-edge environments, whereas the positive response of *V. bahiana* indicates tolerance—or even ecological advantage—under highly exposed coastal and plateau conditions. *V. chamissonis* showed an optimum under moderate, persistent airflow, consistent with maritime environments characterized by continuous ventilation and high humidity. Similarly, responses to surface shortwave radiation and dry-season cloud fraction differentiated species occupying high-radiance environments, filtered-light understories, or diffuse-light coastal systems. These mechanisms are largely absent from previous SDMs focused primarily on standard bioclimatic variables, explaining why exposure-related niche differentiation has remained underexplored despite its ecological relevance for climbing orchids with strong microclimatic dependencies (Zotz, 2013; Fay, 2018).

Differences between our results and previous studies therefore arise not from contradictory ecological signals, but from differences in predictor selection and analytical resolution. While prior SDMs corroborate the importance of thermal and hydric gradients for *Vanilla* species (Watteyn et al., 2020; Watteyn et al., 2025), they often rely on permutation importance metrics and future climate projections, without resolving the functional shape of species responses or identifying precise ecological thresholds. By contrast, our use of quantitative response curves reveals how and where suitability increases, peaks, or collapses along environmental gradients, allowing mechanistic interpretation of niche segregation rather than solely correlative inference.

Spatial patterns of suitability and protection further reflect these ecological distinctions. The broad predicted extent of *V. pompona* across multiple Brazilian biomes aligns with its generalist strategy and tolerance of variable light and moisture regimes, consistent with reports from lowland rainforests and drier Neotropical regions (Watteyn et al., 2020). In contrast, the restricted suitable area and low protection of *V. phaeantha* are congruent with its endangered status and limited area of occupancy, despite contrasting projections of future expansion in studies incorporating climate scenarios and alternative taxonomic treatments (Crop Trust, 2021; Watteyn et al., 2025). Intermediate patterns observed for *V. chamissonis*, *V. bahiana*, and *V. cribbiana* reflect species-specific trade-offs between ecological specialization, geographic restriction, and exposure tolerance, with *V. cribbiana* emerging as the most spatially constrained and conservation-dependent species, for which prior SDM-based assessments were lacking (Crop Trust, 2021).

Patterns of model uncertainty reinforce these interpretations. Higher ensemble disagreement in transitional and peripheral zones, particularly for widespread species such as *V. pompona* and *V. phaeantha*, reflects genuine ecological ambiguity at biome interfaces rather than methodological instability. While previous studies attribute uncertainty primarily to sampling sparsity or extrapolation limits, our explicit spatial quantification of disagreement highlights zones where environmental filtering is weaker or multiple niche dimensions overlap (Watteyn et al., 2020; Watteyn et al., 2025).

One limitation of the present study is the absence of future climate scenario projections. This is particularly relevant given the cultivation timeline of vanilla: newly planted individuals typically require three to five years to produce their first fruits, and commercial harvesting cycles can extend across decades of continuous cultivation (Havkin-Frenkel & Belanger, 2018; Bory et al., 2008). This means that any cultivation or conservation decision informed by the current suitability maps will be implemented and evaluated under climate conditions that may already differ from the 1970–2000 baseline used here. Under high-emission scenarios, tropical regions where vanilla CWRs are native, including the Atlantic Forest and Cerrado biomes, are projected to experience shifts in precipitation seasonality, increased temperature extremes, and contraction of climatically stable refugia within the coming decades (Pecl et al., 2017; Bellard et al., 2012). SDM projections under future climate scenarios (e.g., SSP2-4.5 and SSP5-8.5 from CMIP6) would allow growers, conservationists, and policymakers to assess whether areas currently identified as suitable will remain so across the cultivation horizon of a newly planted individual, and to anticipate range shifts that could affect both in situ conservation strategies and the viability of agroforestry investments (Thuiller et al., 2019; Watteyn et al., 2020). Incorporating such projections represents a direct and tractable next step for this modeling framework, and one with concrete socioeconomic relevance given the long-term commitments involved in vanilla cultivation.

Overall, our results demonstrate that differences between this study and existing literature are largely driven by the explicit incorporation of exposure-related variables and by a focus on present-day, species-specific functional responses rather than future projections or aggregate importance scores. By resolving how individual *Vanilla* species respond to interacting climatic and microclimatic gradients, this work provides a mechanistic ecological framework that complements previous SDMs and strengthens their application in conservation planning, sustainable use, and the valorization of *Vanilla* crop wild relatives (Watteyn et al., 2020; Armenta-Montero et al., 2022; Flanagan et al., 2022; Watteyn et al., 2022; de Oliveira et al., 2022).

Notes on current taxonomic debate

Our study treated *V. bahiana* and *V. phaeantha* as distinct taxa for distribution modeling, despite the ongoing taxonomic debate proposing their synonymy based on morphological

overlap and biogeographical proximity, allowing us to explore both congruences and differences in their ecological niches rather than adopting a unified approach (Karremans et al. 2020; Barberena et al. 2021). This decision aligns with unresolved consensus in recent literature, where some analyses support synonymizing *V. bahiana* under the priority name *V. phaeantha* due to indistinguishable floral traits and broader habitat preferences extending from xerophytic caatinga to semideciduous forests, challenging prior assumptions of narrow endemism (Karremans et al. 2020; Pansarin 2022). However, our SDM results corroborate aspects of this view by revealing shared bioclimatic preferences, such as sensitivity to dry-quarter precipitation, which could explain historical connectivity across their ranges, yet they also highlight disagreements through distinct predicted hotspots—*V. bahiana* concentrated in Brazil's Northeast Bahia, contrasting *V. phaeantha*'s wider Neotropical spread including Amazonian extensions—suggesting potential ecological divergence or cryptic speciation not captured by morphology alone (de Oliveira et al. 2022; Barberena et al. 2021). By modeling them separately, we provide nuanced insights into conservation priorities, such as unprotected Brazilian fragments for *V. bahiana* facing greater fragmentation risks, which would be obscured under synonymy and underscore the value of integrative approaches combining SDM with molecular data to resolve such debates (Karremans et al. 2020; Pansarin 2022).

Conclusions

These results highlight that the conservation prospects of Brazil's native *Vanilla* species depend on spatially explicit strategies that account for species-specific ecological constraints and current patterns of habitat fragmentation. Several species modeled here—particularly *V. cribbiana*—are restricted to fragmented refugia with limited formal protection, while multispecies richness hotspots are largely confined to the coastal Atlantic Forest, a biome that has already lost more than 80% of its original cover. At the same time, these landscapes encompass agroforestry matrices, long-standing traditional land-use systems, and community-based production contexts that overlap with areas of high ecological suitability. By delineating high-certainty core habitats, zones of overlapping suitability, and protection gaps, our study provides a spatial framework to inform conservation planning, including the prioritization of protected area expansion where richness is highest, targeted ex situ strategies for narrowly distributed species, and the identification of regions where sustainable agroforestry initiatives may be ecologically viable. In a global vanilla market increasingly characterized by genetic uniformity and environmentally intensive production systems, the conservation and valorization of Brazil's wild *Vanilla* species represent an opportunity to align biodiversity protection with resilient, diversified, and socially grounded supply-chain models.

References

Aiello-Lammens, M. E., Boria, R. A., Radosavljevic, A., Vilela, B., & Anderson, R. P. (2015). spThin: An R package for spatial thinning of species occurrence records for use in ecological niche models. *Ecography*, 38(5), 541–545. https://doi.org/10.1111/ecog.01132

Amatulli, G., Domisch, S., Tuanmu, M. N., Parr, B., Tordoni, A., Cosentino, L., & Jetz, W. (2018). A suite of global, cross-scale topographic variables for environmental and biodiversity modeling. *Scientific Data*, 5, 180040. https://doi.org/10.1038/sdata.2018.40

Allouche, O., Tsoar, A., Kadmon, R., 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). Journal of Applied Ecology 43, 1223–1232. https://doi.org/10.1111/j.1365-2664.2006.01214.x

Armenta-Montero, S., Menchaca-García, R., Pérez-Silva, A., & Velázquez-Rosas, N. (2022). Changes in the Potential Distribution of Vanilla planifolia Andrews under Different Climate Change Projections in Mexico. Sustainability, 14(5), 2881. https://doi.org/10.3390/su14052881

Azofeifa-Bolaños, J. B., Gigant, L. R., Nicolás-García, M., Pignal, M., Tavares-González, F. B., Hágsater, E., Salazar-Chávez, G. A., Reyes-López, D., Archila-Morales, F. L., García-García, J. A., Da Silva, D., Allibert, A., Solano-Campos, F., Rodríguez-Jimenes, G. D. C., Paniagua-Vásquez, A., Besse, P., Pérez-Silva, A., & Grisoni, M. (2017). A new vanilla species from Costa Rica closely related to V. planifolia (Orchidaceae). European Journal of Taxonomy, 284. https://doi.org/10.5852/ejt.2017.284

Barberena, F. F. V. A., Hermoso, E. L., & Oliveira, M. A. J. de. (2021). Distribución espacial de Vanilla bahiana (Orchidaceae) en dos fitofisonomías de restinga. ¿El patrón espacial varía? Collectanea Botanica, 40, e001–e001. https://doi.org/10.3989/collectbot.2021.v40.001

Barreda-Castillo, J. M., Pansarin, E. R., Monribot-Villanueva, J. L., Guerrero-Analco, J. A., & Menchaca-García, R. A. (2024). Allogamy, metaxenia and hybrids in orchids, what do we know about it? Botanical Sciences, 102(3), 646–670. https://doi.org/10.17129/botsci.3470

Barragán-Ocaña, A., Silva-Borjas, P., & Cecilio-Ayala, E. (2024). Vanilla production in the world and Mexico: Market value and technology. Social Sciences & Humanities Open, 10, 101076. https://doi.org/10.1016/j.ssaho.2024.101076

Barve, N., Barve, V., Jiménez-Valverde, A., Lira-Noriega, A., Maher, S. P., Peterson, A. T., Soberón, J., & Villalobos, F. (2011). The crucial role of the accessible area in ecological niche modeling and species distribution modeling. *Ecological Modelling*, 222(11), 1810–1819. https://doi.org/10.1016/j.ecolmodel.2011.02.011

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology Letters*, 15(4), 365–377. https://doi.org/10.1111/j.1461-0248.2012.01764.x

Boria, R. A., Olson, L. E., Goodman, S. M., & Anderson, R. P. (2014). Spatial filtering to reduce sampling bias can improve the performance of ecological niche models. *Ecological Modelling*, 275, 73–77. https://doi.org/10.1016/j.ecolmodel.2013.12.012

Bory, S., Grisoni, M., Duval, M. F., & Besse, P. (2008). Biodiversity and preservation of vanilla: Present state of knowledge. *Genetic Resources and Crop Evolution*, 55(4), 551–571. https://doi.org/10.1007/s10722-007-9260-3

Chambers, A., Cibrián-Jaramillo, A., Karremans, A. P., Moreno Martinez, D., Hernandez-Hernandez, J., Brym, M., Resende, M. F. R., Moloney, R., Sierra, S. N., Hasing, T., Alomia, Y. A., & Hu, Y. (2021). Genotyping-By-Sequencing diversity analysis of international Vanilla collections uncovers hidden diversity and enables plant improvement. Plant Science, 311, 111019. https://doi.org/10.1016/j.plantsci.2021.111019

Crop Trust. (2021). *Crop wild relatives*. https://www.croptrust.org/what-we-do/projects/crop-wild-relatives/. Accessed: March 28, 2026.

da Silva Oliveira, J. P. (2025). *Occurrence data for vanilla species distribution modeling* [Data set]. GitHub. https://github.com/joanapaulaso/mde_vanilla

da Silva Oliveira, J. P., Garrett, R., Koblitz, M. G. B., & Macedo, A. F. (2025). Integrated targeted and untargeted metabolomics profiling of *Vanilla* species from the Atlantic Forest: Unveiling the bioeconomic potential of *Vanilla cribbiana*. *Food Chemistry*, *464*(Pt 1), 141650. https://doi.org/10.1016/j.foodchem.2024.141650

de Lima, J. F., de Oliveira, D. C., Kuster, V. C., & Moreira, A. S. F. P. (2024). Aerial and terrestrial root habits influence the composition of the cell walls of Vanilla phaeantha (Orchidaceae). Protoplasma. https://doi.org/10.1007/s00709-024-01980-9

De Oliveira, R. T., Da Silva Oliveira, J. P., Da Silva, A. L. M., Carrão Dantas, E. K., Koblitz, M. G. B., Bello, M. L., Felzenszwalb, I., Araújo-Lima, C. F., & Macedo, A. F. (2024). Vanilla from Brazilian Atlantic Forest: In vitro and in silico toxicity assessment and high-resolution metabolomic analysis of Vanilla spp. ethanolic extracts. Food Chemistry, 456, 139948. https://doi.org/10.1016/j.foodchem.2024.139948

de Oliveira, R.T., da Silva Oliveira, J.P., & Macedo, A.F. (2022). Vanilla beyond Vanilla planifolia and Vanilla × tahitensis: Taxonomy and historical notes, reproductive biology, and metabolites. Plants, 11(3311). https://doi.org/10.3390/plants11233311

Divakaran, M., & Fathima Rafieah, N. T. (2021). Genetic Resources of The Universal Flavor, Vanilla. In R. Shabir Ahmad (Ed.), Herbs and Spices—New Processing Technologies. IntechOpen. https://doi.org/10.5772/intechopen.99043

Ellestad, P. (2023). Anything but Bland: Uncovering the Hidden Diversity and Genomic Origin of the Threatened Vanilla Spice [Doctor of Philosophy in Ecology, Evolution, and Behavior, Boise State University]. https://doi.org/10.18122/td.2102.boisestate

Estopinan, J., Servajean, M., Bonnet, P., Munoz, F., & Joly, A. (2022). Deep Species Distribution Modeling From Sentinel-2 Image Time-Series: A Global Scale Analysis on the Orchid Family. Frontiers in Plant Science, 13, 839327. https://doi.org/10.3389/fpls.2022.839327

Favre, F., Jourda, C., Grisoni, M., Piet, Q., Rivallan, R., Dijoux, J.-B., Hascoat, J., Lepers-Andrzejewski, S., Besse, P., & Charron, C. (2022). A genome-wide assessment of the genetic diversity, evolution and relationships with allied species of the clonally propagated crop Vanilla planifolia Jacks. Ex Andrews. Genetic Resources and Crop Evolution, 69(6), 2125–2139. https://doi.org/10.1007/s10722-022-01362-1

Fay, M. F. (2018). Orchid conservation: How can we meet the challenges in the twenty-first century? *Botanical Studies*, 59(1), 16. https://doi.org/10.1186/s40529-018-0232-z

Ferreira, A. W. C., De Oliveira, M. S., Silva, E. O., Campos, D. S., Pansarin, E. R., & Guarçoni, E. A. E. (2017). Vanilla bahiana Hoehne and Vanilla pompona Schiede (Orchidaceae, Vanilloideae): Two new records from Maranhão state, Brazil. Check List, 13(6), 1131–1137. https://doi.org/10.15560/13.6.1131

Flanagan, N. S., Navia-Samboni, A., González-Pérez, E. N., & Mendieta-Matallana, H. (2022). Distribution and conservation of vanilla crop wild relatives: The value of local community engagement for biodiversity research. Neotropical Biology and Conservation, 17(3), 205–227. https://doi.org/10.3897/neotropical.17.e86792

Franco, J. L., Ayres, E. M. M., de Oliveira, D., Martins, I. B. A., Macedo, A. F., Deliza, R., & Koblitz, M. G. B. (2024). Exploring the potential of the Vanilla species from the Brazilian Atlantic Forest: Sensory description and consumer acceptance. Journal of Sensory Studies, 39(1), e12896. https://doi.org/10.1111/joss.12896

Goettsch, B., Aranguren-Riaño, N., Aguilar-Lome, J., et al. (2021). Hundreds of cactus species threatened with extinction following new IUCN assessment. *Nature Plants*, 7(3), 259–265. https://doi.org/10.1038/s41477-021-00842-5

Guéguen, M., Blancheteau, H., Thuiller, W., 2025. biomod2: Ensemble Platform for Species Distribution Modeling. URL https://biomodhub.github.io/biomod2/ (accessed 11.1.25)

Guisan, A., Thuiller, W., & Zimmermann, N. E. (2017). Habitat Suitability and Distribution Models: With Applications in R (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781139028271

Hending, D., Andrianiaina, A., Rakotomalala, Z., & Cotton, S. (2023). Conservation value of vanilla agroecosystems for vertebrate diversity in north-east Madagascar. Oryx, 57(1), 118–128. https://doi.org/10.1017/S0030605321001265

Hu, Y., Resende, M. F. R., Bombarely, A., Brym, M., Bassil, E., & Chambers, A. H. (2019). Genomics-based diversity analysis of Vanilla species using a Vanilla planifolia draft genome and Genotyping-By-Sequencing. Scientific Reports, 9(1), 3416. https://doi.org/10.1038/s41598-019-40144-1

Karremans, A. P., Chinchilla, I. F., Rojas-Alvarado, G., Cedeño-Fonseca, M., Damián, A., & Léotard, G. (2020). A reappraisal of Neotropical Vanilla. With a note on taxonomic inflation and the importance of alpha taxonomy in biological studies. Lankesteriana. https://doi.org/10.15517/lank.v20i3.45203

Karremans, A. P., Pupulin, F., & Damián-Parizaca, A. (2025). A revised infrageneric classification of Vanilla (Orchidaceae) based on phylogenetic, morphological and biogeographical evidence. Phytotaxa, 715(3), 207–228. https://doi.org/10.11646/phytotaxa.715.3.2

Karremans, A. P., Watteyn, C., Scaccabarozzi, D., Pérez-Escobar, O. A., & Bogarín, D. (2023). Evolution of Seed Dispersal Modes in the Orchidaceae: Has the Vanilla Mystery Been Solved? Horticulturae, 9(12), Article 12. https://doi.org/10.3390/horticulturae9121270

Khoury, C. K., Carver, D., Greene, S. L., et al. (2021). Crop wild relatives on the move: Climatically suitable areas will contract and shift due to climate change. *Global Change Biology*, 26(3), 1613–1624. https://doi.org/10.1111/gcb.14973

Li, J., Demesyeux, L., Brym, M., & Chambers, A. H. (2020). Development of species-specific molecular markers in Vanilla for seedling selection of hybrids. Molecular Biology Reports, 47(3), 1905–1920. https://doi.org/10.1007/s11033-020-05287-9

Maceda, A., Delgado-Alvarado, A., Salazar-Rojas, V. M., & Herrera-Cabrera, B. E. (2023). Vanilla planifolia Andrews (Orchidaceae): Labellum Variation and Potential Distribution in Hidalgo, Mexico. Diversity, 15(5), 678. https://doi.org/10.3390/d15050678

Martin, D. A., Andriafanomezantsoa, R., Dröge, S., Osen, K., Rakotomalala, E., Wurz, A., Andrianarimisa, A., & Kreft, H. (2021). Bird diversity and endemism along a land-use gradient in Madagascar: The conservation value of vanilla agroforests. Biotropica, 53(1), 179–190. https://doi.org/10.1111/btp.12859

Maurya, P., Kumar, V., Jain, S., Saroj, N., Maurya, S., Mishra, D., & Prasad, K. (2025). Crop Wild Relatives (CWRs) and Their Potential Role in Improvement of Horticultural Crops. In H. S. Jatav, V. D. Raiput, & T. Minkina (Eds.), Ecologically Mediated Development (Vol. 41, pp. 51–72). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-2413-3_3

McKerrow, A.J., Tarr, N.M., Rubino, M.J., Williams, S.G., 2018. Patterns of species richness hotspots and estimates of their protection are sensitive to spatial resolution. Diversity and Distributions 24, 1464–1477. https://doi.org/10.1111/ddi.12779

Meller, L., Cabeza, M., Pironon, S., Barbet-Massin, M., Maiorano, L., Georges, D., Thuiller, W., 2014. Ensemble distribution models in conservation prioritization: from consensus predictions to consensus reserve networks. Diversity and Distributions 20, 309–321. https://doi.org/10.1111/ddi.12162

Miller, J. (2010). Species Distribution Modeling. Geography Compass, 4(6), 490–509. https://doi.org/10.1111/j.1749-8198.2010.00351.x

Nadal, M. C., Teixeira, G. D. N., Coronado, J. F. O., Ochiutto, A. A., Cunha Neto, A. R. D., & Reis, M. V. D. (2025). Potential of Brazil's native vanilla species: A study on distribution and research gaps. Ornamental Horticulture, 31, e312874. https://doi.org/10.1590/2447-536x.v31.e312874

Pansarin, E. R. (2021). Unravelling the enigma of seed dispersal in Vanilla. Plant Biology, 23(6), 974–980. https://doi.org/10.1111/plb.13331

Pansarin, E. R. (2022). Vanilla flowers: Much more than food-deception. Botanical Journal of the Linnean Society, 198(1), 57–73. https://doi.org/10.1093/botlinnean/boab046

Pansarin, E. R. (2025a). Epiphyte Vanilla relies on birds as long-distance seed dispersers. Plant Biology, plb.70086. https://doi.org/10.1111/plb.70086

Pansarin, E. R. (2025b). Systematics of the Vanilla chamissonis clade (Orchidaceae): A study based on integrative taxonomy. Plant Ecology and Evolution, 158(2), 260–278. https://doi.org/10.5091/plecevo.154789

Pansarin, E. R. (2025c). Vanilla × robusta (Orchidaceae: Vanilloideae), the first natural Vanilla hybrid for South America. Nordic Journal of Botany, 2025(7), e04743. https://doi.org/10.1111/njb.04743

Pansarin, E. R., & Ferreira, A. W. C. (2021). Evolutionary disruption in the pollination system of Vanilla (Orchidaceae). Plant Biology, 23(1), 157–167. https://doi.org/10.1111/plb.13356

Pansarin, E. R., & Menezes, E. da L. F. (2023). A new remarkable Vanilla Mill. (Orchidaceae) species endemic to the Espinhaço Range, Brazil: Its phylogenetic position and evolutionary relationships among Neotropical congeners. PhytoKeys, 227, 151. https://doi.org/10.3897/phytokeys.227.101963

Pecl, G. T., Araújo, M. B., Bell, J. D., Blanchard, J., Bonebrake, T. C., Chen, I. C., … Williams, S. E. (2017). Biodiversity redistribution under climate change: Impacts on ecosystems and human well-being. *Science*, 355(6332), eaai9214. https://doi.org/10.1126/science.aai9214

Pérez-Silva, A., Peña-Mojica, E., Ortega-Galeana, A., López-Cruz, J. I., Ledesma-Escobar, C. A., Rivera-Rivera, M., & Paz-Gamboa, E. (2025). Maya Vanilla (Vanilla cribbiana Soto Arenas): A New Species in Commerce. Plants, 14(3), 300. https://doi.org/10.3390/plants14030300

Peterson, A. T., & Soberón, J. (2012). Species Distribution Modeling and Ecological Niche Modeling: Getting the Concepts Right. Natureza & Conservação, 10(2), 102–107. https://doi.org/10.4322/natcon.2012.019

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological Niches and Geographic Distributions*. Princeton University Press.

Ramampiandra, E. C., Scheidegger, A., Wydler, J., & Schuwirth, N. (2023). A comparison of machine learning and statistical species distribution models: Quantifying overfitting supports model interpretation. Ecological Modelling, 481, 110353.

Silva, F. N. D., Vieira, R. F., Bizzo, H. R., et al. (2023). Chemical characterization and sensory potential of Brazilian vanilla species. *Pesquisa Agropecuária Brasileira*, 58, e03308. https://doi.org/10.1590/s1678-3921.pab2023.v58.03308

Serna-Sánchez, M. A., Karremans, A. P., Bogarín, D., & Fuchs, E. J. (n.d.). OPEN Contrasting clonal and population genetic structure in two endangered Costa Rican Vanilla species of commercial interest. Scientific Reports.

Soto Arenas, M. A., & Cribb, P. (2009). A new infrageneric classification and synopsis of the genus Vanilla Plum. ex mill. (Orchidaceae: Vanillinae). Lankesteriana: International Journal on Orchidology. https://doi.org/10.15517/lank.v0i0.12071

Trabucco, A., & Zomer, R. (2019). *Global aridity index and potential evapotranspiration climate database v2*. CGIAR Consortium for Spatial Information. https://doi.org/10.6084/m9.figshare.7504448

Thuiller, W., Guéguen, M., Renaud, J., Karger, D. N., & Zimmermann, N. E. (2019). Uncertainty in ensembles of global biodiversity scenarios. *Nature Communications*, 10, 1446. https://doi.org/10.1038/s41467-019-09519-w

Velázquez-Rosas, N., Sinaca Colin, S., Vázquez-Domínguez, G., Velasco-Murguía, A., Silva Rivera, E., Ruiz-Guerra, B., Friedrich, F. L., Cortés Galindo, R., Armenta-Montero, S., & Martínez-Mota, R. (2025). Importance of Traditional Vanilla Cultivation in the Conservation of Plant Diversity in Tropical Forests in Northern Veracruz, Mexico. Sustainability, 17(6), 2598. https://doi.org/10.3390/su17062598

Villanueva-Viramontes, S., Hernández-Apolinar, M., Fernández-Concha, G. C., Dorantes-Euán, A., Dzib, G. R., & Martínez Castillo, J. (2017). Wild Vanilla planifolia and its relatives in the Mexican Yucatan Peninsula: Systematic analyses with ISSR and ITS. Botanical Sciences, 95(2), 169–187. https://doi.org/10.17129/botsci.668

Wang, X.-M., Tang, Y., Peng, X.-F., Wang, J., Zhang, S.-Q., Feng, Y., & Peng, P.-H. (2024). Identifying priorities under highly heterogeneous environments through species distribution models to facilitate orchid conservation. Biodiversity and Conservation, 33(2), 647–665. https://doi.org/10.1007/s10531-023-02764-y

Watteyn, C., Dejonghe, O., Van Hoyweghen, K., Azofeifa Bolaños, J. B., Karremans, A. P., Vranken, L., Reubens, B., Muys, B., & Maertens, M. (2022). Exploring farmer preferences towards innovations in the vanilla supply chain. Journal of Cleaner Production, 330, 129831. https://doi.org/10.1016/j.jclepro.2021.129831

Watteyn, C., Fremout, T., Karremans, A. P., Huarcaya, R. P., Azofeifa Bolaños, J. B., Reubens, B., & Muys, B. (2020). Vanilla distribution modeling for conservation and sustainable cultivation in a joint land sparing/sharing concept. Ecosphere, 11(3), e03056. https://doi.org/10.1002/ecs2.3056

Watteyn, C., Fremout, T., Karremans, A. P., Van Meerbeek, K., Janssens, S. B., De Backer, S., Lipińska, M. M., & Muys, B. (2025). Wild Vanilla and pollinators at risk of spatial mismatch in a changing climate. Frontiers in Plant Science, 16, 1585540. https://doi.org/10.3389/fpls.2025.1585540

Zotz, G. (2013). The systematic distribution of vascular epiphytes – a critical update. *Botanical Journal of the Linnean Society*, 171(3), 453–481. https://doi.org/10.1111/boj.12010

2. Proposta de Divulgação Científica - ReSourceae (https://resourceae.com/)

Sobre a criação da marca. A identidade visual da ReSourceae (Figura 2) é tecnicamente consistente com o conceito do aplicativo e revela um desenho simbólico, no qual forma, cor e tipografia operam como camadas semióticas complementares. A marca comunica permanência, rigor científico e continuidade, atributos coerentes com uma plataforma orientada à ciência aberta, reutilização de dados e tomada de decisão informada.



Figura 2. Logomarca da plataforma online ReSourceae.

Do ponto de vista simbólico-formal, o elemento central combina o símbolo do infinito com uma folha estilizada. O infinito funciona como metáfora direta da reutilização contínua de dados, da circularidade do conhecimento e da ideia de que dados científicos não se esgotam em um único uso, mas ganham valor à medida que são reinterpretados e entregues em novos contextos. A folha introduz a dimensão biológica, ancorando o infinito em um sistema vivo, e ecológico. O cruzamento suave das curvas sugere fluxo, trânsito de informação e interoperabilidade entre camadas de dados (química, ecológica, socioeconômica), um reflexo direto do pipeline de bioprospecção que a plataforma propõe. A escolha por linhas contínuas, orgânicas e de espessura uniforme reforça a noção de estabilidade metodológica e transparência. Não há ângulos agudos nem contrastes bruscos, o que dialoga com a ideia de mediação entre ciência e sociedade: a plataforma conduz o usuário por interpretações baseadas em evidência. No que se refere à paleta cromática, os tons terrosos, verdes suaves e bege funcionam como um sistema cromático sem ruído semântico: o verde remete à biodiversidade, regeneração e sistemas vivos. O bege e os tons argilosos evocam território, solo, cartografia e materialidade, importante para uma plataforma que trabalha explicitamente com mapas de adequabilidade, *hotspots* e espacialização de dados. O cinza-esverdeado escuro atua como elemento de ancoragem visual, transmitindo sobriedade, confiabilidade e maturidade institucional.

A tipografia serifada escolhida para o logotipo introduz uma camada acadêmica e histórica. Serifas estão associadas à tradição do conhecimento científico, à literatura clássica e à perenidade dos registros, o que dialoga diretamente com a proposta de sustentabilidade informacional. A logo funciona bem tanto em fundos claros quanto em aplicações monocromáticas, o que é essencial para relatórios técnicos, artigos científicos, apresentações institucionais e possíveis desdobramentos editoriais do projeto (como o futuro "atlas de bioprospecção"). Em síntese, a marca ReSourceae traduz visualmente o eixo central do aplicativo: dados como recursos vivos, conhecimento como fluxo contínuo e ciência como ponte entre biodiversidade, território e inovação sustentável.

Métodos de construção da plataforma online. A ReSourceae (https://resourceae.com/) é uma aplicação web estática com mapas interativos (Leaflet) e renderização de GeoTIFFs diretamente no navegador. A aplicação organiza mapas por espécie e por riqueza, com suporte a PT/EN via dicionário interno.

Principais escolhas:

- Frontend estático (HTML/CSS/JS) para facilitar publicação em qualquer servidor simples.
- Mapas no navegador usando Leaflet + GeoRaster (GeoTIFF) + shpjs (*shapefile*).
- Sem *backend*: tudo roda localmente ou em hospedagem estática.

A plataforma ReSourceae foi desenvolvida com base em tecnologias *web* consolidadas, priorizando simplicidade de implantação, interoperabilidade e acessibilidade direta no navegador. A estrutura do aplicativo é construída em HTML5, responsável pela organização semântica do conteúdo, enquanto o CSS é empregado para o controle do estilo visual e a adaptação responsiva da interface a diferentes tamanhos de tela e dispositivos. A camada de interatividade e lógica do sistema é implementada em JavaScript, que desempenha múltiplas funções centrais no funcionamento da plataforma, incluindo a inicialização e o controle dos mapas interativos por meio da biblioteca Leaflet, a renderização de dados *raster* diretamente no navegador utilizando a combinação das bibliotecas geotiff.js, georaster e georaster-layer-for-leaflet, a leitura de arquivos vetoriais no formato *shapefile* de forma client-side com o auxílio da biblioteca shpjs, a implementação do mecanismo de internacionalização (PT/EN) por meio de dicionários locais e o gerenciamento de indicadores visuais de carregamento para melhorar a experiência do usuário durante a manipulação de dados espaciais.

As dependências externas necessárias ao funcionamento dos mapas são carregadas diretamente via redes de distribuição de conteúdo (CDN), integradas aos arquivos HTML responsáveis pela visualização cartográfica. Entre essas dependências estão a biblioteca Leaflet, utilizada para a construção dos mapas base, o conjunto formado por geotiff.js, georaster e georaster-layer-for-leaflet, que possibilita a leitura e a visualização de arquivos GeoTIFF no próprio navegador, e a biblioteca shpjs, empregada para a leitura de *shapefiles* compactados. Essa estratégia elimina a necessidade de processos de build ou empacotamento de código, simplificando a manutenção e a replicabilidade do sistema, embora implique a dependência de acesso à rede para o carregamento das bibliotecas no ambiente do usuário.

A preparação dos dados espaciais segue um fluxo automatizado por meio de um script responsável por organizar e compactar os arquivos necessários à aplicação. Nesse processo, os arquivos GeoTIFF, gerados individualmente para cada espécie modelada, são copiados para o diretório data/rasters/, enquanto o shapefile correspondente às áreas protegidas é compactado e armazenado como extensão .zip. De forma análoga, o shapefile de riqueza de espécies, resultante da união dos modelos, é compactado e disponibilizado no caminho data/richness/*.zip. Essa organização padronizada dos dados facilita a leitura direta pelo navegador, assegura a rastreabilidade das fontes espaciais e contribui para a sustentabilidade informacional e a reutilização dos dados no contexto da plataforma.

Integração multidisciplinar, sistemas complexos e materialização do arcabouço conceitual na plataforma ReSourceae. Ao longo desta tese, foi sustentado que a valorização efetiva da biodiversidade e o avanço da bioprospecção contemporânea dependem da superação de abordagens fragmentadas, exigindo a integração de múltiplas disciplinas, escalas analíticas e tipos de dados em estruturas coerentes de interpretação (Morin, 2005; Kitano, 2002). A biodiversidade, entendida como um sistema complexo adaptativo, não pode ser adequadamente compreendida apenas a partir de níveis moleculares isolados, tampouco exclusivamente por descrições ecológicas ou econômicas desconectadas. Nesse contexto, a integração entre metabolômica, bioinformática, modelagem ecológica e análise territorial emerge como condição necessária para transformar dados em conhecimento acionável no âmbito da bioeconomia e da conservação (Selvarajoo, 2013; Misra et al., 2019).

A literatura discutida na introdução desta tese aponta que sistemas biológicos e socioambientais apresentam propriedades emergentes que não são inferíveis a partir da análise isolada de seus componentes, mas apenas a partir da articulação entre níveis

hierárquicos distintos, conectados por fluxos de informação (Kitano, 2002; Morin, 2005). Essa perspectiva é particularmente relevante para a bioprospecção, na qual compostos químicos, organismos produtores, ambientes ecológicos e contextos socioeconômicos formam um sistema interdependente. Abordagens baseadas exclusivamente na descoberta molecular, sem contextualização ecológica e territorial, tendem a gerar conhecimento pouco transferível para estratégias sustentáveis de uso e conservação (Wolfender et al., 2019; Johnson et al., 2016).

Nesse sentido, os capítulos desta tese foram concebidos como partes interconectadas de um mesmo sistema analítico. A revisão metodológica em metabolômica estabeleceu as bases conceituais para a anotação molecular e a generalização por classes químicas, evidenciando a importância da reutilização estruturada de dados e da transparência metodológica (Sumner et al., 2007; Wolfender et al., 2019). Os estudos metabolômicos aplicados às espécies nativas de *Vanilla* geraram dados químicos primários de alto rendimento, posteriormente reinterpretáveis por meio de fluxos *in silico* de predição de bioatividade, farmacocinética baseada em fisiologia e *docking* molecular, ampliando o potencial informacional desses conjuntos de dados (Misra et al., 2019; Ferreira et al., 2021). Por fim, a integração com modelos de distribuição de espécies permitiu situar esses metabólitos e seus potenciais funcionais no espaço ambiental, conectando diversidade química, adequabilidade ecológica e conservação (Elith & Leathwick, 2009; Peterson et al., 2011).

O desfecho natural desse percurso conceitual e metodológico é a materialização dessas integrações no objeto-síntese da tese: a plataforma ReSourceae. A ReSourceae operacionaliza, em nível aplicado, os princípios de ciência aberta, sustentabilidade informacional e pensamento sistêmico discutidos ao longo do trabalho, ao tratar dados biológicos como recursos reutilizáveis e dinâmicos, cuja recombinação gera novos significados e aplicações. Ao integrar mapas de adequabilidade ambiental, *hotspots* de biodiversidade, informações sobre proteção legal e indicadores de valor bioeconômico, a plataforma traduz dados científicos complexos em camadas interpretativas acessíveis a pesquisadores, gestores públicos e tomadores de decisão, reduzindo a distância entre produção de conhecimento e aplicação prática (Morin, 2005; Bugge et al., 2016).

Sob a ótica da teoria de sistemas complexos, a ReSourceae não atua como um simples repositório de dados, mas como um sistema integrador que favorece a emergência de novas leituras a partir da interação entre dados biotecnológicos, ecológicos e territoriais (Kitano,

2002; Selvarajoo, 2013). Essa abordagem reforça a ideia de que a inovação bioeconômica e a conservação da biodiversidade não são processos antagônicos, mas dimensões complementares de um mesmo sistema socioecológico, cuja compreensão exige ferramentas capazes de articular múltiplas escalas e disciplinas (Bugge et al., 2016; Morin, 2005).

Dessa forma, a plataforma ReSourceae fecha o ciclo argumentativo proposto na introdução da tese, ao demonstrar que a integração de dados multidisciplinares pode gerar, com sucesso, informações essenciais para tornar o pipeline de bioprospecção mais transparente, acessível e orientado à sustentabilidade. Como perspectivas futuras, a expansão da plataforma para incluir outras espécies e grupos taxonômicos, em colaboração com diferentes pesquisadores, consolida a ReSourceae como um atlas de bioprospecção orientado por dados, capaz de apoiar políticas públicas, estratégias de conservação e iniciativas de inovação baseadas na biodiversidade de forma sistêmica e informada.

3. Conclusões Gerais

Esta tese mostrou que a integração de abordagens e dados das ciências naturais e computacionais, incluindo metabolômica de alta resolução, modelagem bioinformática (predição de bioatividade, farmacocinética e *docking* molecular) e modelos de distribuição de espécies, constitui um caminho fértil e robusto para a bioprospecção orientada por dados. Ao alinhar as descobertas de cada capítulo, observou-se que o conhecimento químico-funcional de metabólitos especializados pode ser contextualizado no espaço ambiental e socioeconômico e que essa sinergia gera informação essencial para orientar estratégias de conservação, uso sustentável e inovação bioeconômica. O eixo estruturante do trabalho mostrou que dados de disciplinas distintas, quando integrados por fluxos analíticos transparentes, permitem transformar evidências dispersas em produtos de valor concreto para a sociedade.

A materialização mais tangível desse esforço foi a criação da plataforma ReSourceae (https://resourceae.com/), que trata dados biológicos e biogeográficos como recursos vivos, enfatizando que a reutilização responsável, fora do nicho acadêmico, agrega valor científico, ecológico e socioeconômico. Na plataforma, mapas de adequabilidade ambiental, *hotspots* de biodiversidade e resumos interpretativos do potencial bioeconômico são disponibilizados de forma a traduzir o conhecimento científico para gestores públicos, produtores rurais, comunidades locais, formuladores de políticas e investidores, conectando valor químico, potencial bioativo e viabilidade ecológica. Ao realizar tal conexão, a ReSourceae demonstra que dados complexos podem ser apresentados de forma acessível, fortalecendo a comunicação entre ciência e tomada de decisão e contribuindo para a sustentabilidade informacional e a eficiência epistemológica.

Como perspectivas futuras, planeja-se ampliar a divulgação da plataforma e expandir o conjunto de dados incorporados, envolvendo outras espécies de interesse econômico e ecológico em colaboração com diferentes grupos de pesquisa, de modo que a ReSourceae evolua para um verdadeiro atlas de bioprospecção. Essa expansão permitirá comparar padrões bioquímicos, ecológicos e geográficos entre táxons e regiões, enriquecendo a interface entre diversidade química e diversidade biológica e proporcionando bases ainda mais sólidas para decisões sobre conservação e desenvolvimento sustentável.

4. Referências Bibliográficas

Araújo, M. B., & Peterson, A. T. (2012). Uses and misuses of bioclimatic envelope modeling. *Ecology*, 93(7), 1527–1539. https://doi.org/10.1890/11-1930.1

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., & Supuran, C. T. (2021). Natural products in drug discovery: Advances and opportunities. *Nature Reviews Drug Discovery*, 20(3), 200–216. https://doi.org/10.1038/s41573-020-00114-z

Baker, Aryn/Sahabevava. 2018. *Vanilla Is Nearly as Expensive as Silver. That Spells Trouble for Madagascar*. TIME Magazine. https://time.com/5308143/vanilla-price-climate-change-madagascar/. acesso em 23 de dezembro de 2025.

Bellard, C., Bertelsmeier, C., Leadley, P., Thuiller, W., & Courchamp, F. (2012). Impacts of climate change on the future of biodiversity. *Ecology Letters*, 15(4), 365–377. https://doi.org/10.1111/j.1461-0248.2011.01736.x

Bezerra, L. A. V., Garcez, D. S., da Silva, M. C., Gurgel-Lourenço, R. C., Pinto, L. M., Valentim, G. A., Batista, L. B., Avelar, V. S., Siqueira, W. L. X., Loiola, S. C., Rossi, S., Freitas, J. G. de, & Sánchez-Botero, J. I. (2025). BRValuation: a systematic comparison of ecosystem services across Brazilian biomes and ecosystems. *Environmental Development, 56*, 101279. https://doi.org/10.1016/j.envdev.2025.101279

Blaženović, I., Kind, T., Ji, J., & Fiehn, O. (2018). Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites*, 8(2), 31. https://doi.org/10.3390/metabo8020031

Bugge, M. M., Hansen, T., & Klitkou, A. (2016). What is the bioeconomy? A review of the literature. *Sustainability*, 8(7), 691. https://doi.org/10.3390/su8070691

CAPES. (2020). *Documento de área: Biodiversidade*. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

Castaneda-Ovando, A., Pacheco-Hernández, M. L., Páez-Hernández, M. E., Rodríguez, J. A., & Galán-Vidal, C. A. (2009). Chemical studies of anthocyanins: A review. *Food Chemistry*, 113(4), 859–871. https://doi.org/10.1016/j.foodchem.2008.09.001

Silva, F. N. da, Vieira, R. F., Bizzo, H. R., Gama, P. E., Brumano, C. N., Vidigal, M. C. T. R., Fernandes Neto, A. A., Crepalde, L. T., & Minim, V. P. R. (2023). Chemical

characterization and sensory potential of Brazilian vanilla species. *Pesquisa Agropecuária Brasileira, 58*, e03308. https://doi.org/10.1590/S1678-3921.pab2023.v58.03308

Dettmer, K., Aronov, P. A., & Hammock, B. D. (2007). Mass spectrometry-based metabolomics. *Mass Spectrometry Reviews*, 26(1), 51–78. https://doi.org/10.1002/mas.20108

Díaz, S., Settele, J., Brondízio, E. S., Ngo, H. T., Guèze, M., Agard, J., … Zayas, C. N. (2019). Pervasive human-driven decline of life on Earth points to the need for transformative change. *Science*, 366(6471), eaax3100. https://doi.org/10.1126/science.aax3100

Elith, J., & Leathwick, J. R. (2009). Species distribution models: Ecological explanation and prediction across space and time. *Annual Review of Ecology, Evolution, and Systematics*, 40, 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159

Fadda, L. A., Lasa-Covarrubias, R., Osorio-Olvera, L., Murúa, M. G., & Lira-Noriega, A. (2026). Applications of ecological niche and species distribution models in agricultural, livestock, and forestry systems: A comprehensive review. *Agricultural Systems, 231*, 104542. https://doi.org/10.1016/j.agsy.2025.104542

Ferreira, L. G., dos Santos, R. N., Oliva, G., & Andricopulo, A. D. (2015). Molecular *docking* and structure-based drug design strategies. *Molecules*, 20(7), 13384–13421. https://doi.org/10.3390/molecules200713384

Fiehn, O. (2002). Metabolomics—The link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1–2), 155–171. https://doi.org/10.1023/A:1013713905833

Guisan, A., Tingley, R., Baumgartner, J. B., Naujokaitis-Lewis, I., Sutcliffe, P. R., Tulloch, A. I. T., … Buckley, Y. M. (2013). Predicting species distributions for conservation decisions. *Ecology Letters*, 16(12), 1424–1435. https://doi.org/10.1111/ele.12189

Hammami, A. M. & Elliott, C. 2025. *Sourcing Vanilla Closer to Home: Market Potential for US Production*. https://www.choicesmagazine.org/choices-magazine/submitted-articles/sourcing-vanilla-closer-to-home-market-potential-for-us-production. acesso em 2 de janeiro de 2026.

Harvey, A. L., Edrada-Ebel, R., & Quinn, R. J. (2015). The re-emergence of natural products for drug discovery in the genomics era. *Nature Reviews Drug Discovery*, 14(2), 111–129. https://doi.org/10.1038/nrd4510

Hood, L., Lovejoy, J. C., & Price, N. D. (2012). Integrating big data and actionable health coaching to optimize wellness. *BMC Medicine*, 10, 30. https://doi.org/10.1186/1741-7015-10-30

Hughes, J. P., Rees, S., Kalindjian, S. B., & Philpott, K. L. (2011). Principles of early drug discovery. *British Journal of Pharmacology*, 162(6), 1239–1249. https://doi.org/10.1111/j.1476-5381.2010.01127.x

Ideker, T., Galitski, T., & Hood, L. (2001). A new approach to decoding life: Systems biology. *Annual Review of Genomics and Human Genetics*, 2, 343–372. https://doi.org/10.1146/annurev.genom.2.1.343

IPBES (2019): Global assessment report on biodiversity and ecosystem services of the Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. E. S. Brondizio, J. Settele, S. Díaz, and H. T. Ngo (editors). IPBES secretariat, Bonn, Germany. 1148 pages. https://doi.org/10.5281/zenodo.3831673

Johnson, C. H., Ivanisevic, J., & Siuzdak, G. (2020). Metabolomics: Beyond biomarkers and towards mechanisms. *Nature Reviews Molecular Cell Biology*, 17(7), 451–459. https://doi.org/10.1038/nrm.2016.25

Joly, C. A., Metzger, J. P., & Tabarelli, M. (2014). Experiences from the Brazilian Atlantic Forest: Ecological findings and conservation initiatives. *New Phytologist*, 204(3), 459–473. https://doi.org/10.1111/nph.12989

Joly, C. A., Scarano, F. R., Bustamante, M., Gadda, T., Metzger, J. P., Seixas, C. S., … Young, C. E. F. (2018). Brazilian assessment on biodiversity and ecosystem services: Summary for policy makers. *Biota Neotropica*, 18(4), e20180589. https://doi.org/10.1590/1676-0611-BN-2019-0865

Kell, D. B., & Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays*, 26(1), 99–105. https://doi.org/10.1002/bies.10385

Kirchmair, J., Göller, A. H., Lang, D., Kunze, J., Testa, B., Wilson, I. D., … Schneider, G. (2015). Predicting drug metabolism: Experiment and/or computation? *Nature Reviews Drug Discovery*, 14(6), 387–404. https://doi.org/10.1038/nrd4581

Kitano, H. (2002). Systems biology: A brief overview. *Science*, 295(5560), 1662–1664. https://doi.org/10.1126/science.1069492

Leonelli, S. (2019). Data governance is key to interpretation: Reconceptualizing data in data science. *Harvard Data Science Review*, 1(1). https://doi.org/10.1162/99608f92.17405bb6

Leonelli, S., Spichtinger, D., & Prainsack, B. (2015). Sticks and carrots: Encouraging open science at its source. *Geo*, 2(1), 12–16. https://doi.org/10.1002/geo2.2

Misra, B. B. (2021). New software tools, databases, and resources in metabolomics: Updates from 2020. *Metabolomics*, 17(5), 49. https://doi.org/10.1007/s11306-021-01796-1

Misra, B. B., Langefeld, C. D., Olivier, M., & Cox, L. A. (2019). Integrated omics: Tools, advances and future approaches. *Journal of Molecular Endocrinology*, 62(1), R21–R45. https://doi.org/10.1530/JME-18-0055

Mittermeier, R. A., Turner, W. R., Larsen, F. W., Brooks, T. M., & Gascon, C. (2011). Global biodiversity conservation: The critical role of *hotspots*. In F. E. Zachos & J. C. Habel (Eds.), *Biodiversity hotspots* (pp. 3–22). Springer. https://doi.org/10.1007/978-3-642-20992-5_1

Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L. O. B., & Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, 37(1), 49–56. https://doi.org/10.3233/ISU-170824

Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., & Kent, J. (2000). Biodiversity *hotspots* for conservation priorities. *Nature*, 403(6772), 853–858. https://doi.org/10.1038/35002501

Nadal, M. C., Teixeira, G. das N., Oyarzún Coronado, J. F., Ochiutto, A. A., Cunha Neto, A. R. da, & Reis, M. V. dos. (2025). Potential of Brazil's native vanilla species: A study on distribution and research gaps. *Ornamental Horticulture, 31*, e312874. https://doi.org/10.1590/2447-536X.v31.e312874

Newman, D. J., & Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products*, 83(3), 770–803. https://doi.org/10.1021/acs.jnatprod.9b01285

Nicholson, J. K., Holmes, E., Kinross, J. M., Darzi, A. W., Takats, Z., & Lindon, J. C. (2012). Metabolic phenotyping in clinical and surgical environments. *Nature*, 491(7424), 384–392. https://doi.org/10.1038/nature11708

Nicholson, J. K., Lindon, J. C., & Holmes, E. (1999). 'Metabonomics': Understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29(11), 1181–1189. https://doi.org/10.1080/004982599238047

OECD. (2018). *Meeting policy challenges for a sustainable bioeconomy*. OECD Publishing. https://doi.org/10.1787/9789264292345-en

Oliveira, J. P. da S., Garrett, R., Koblitz, M. G. B., & Macedo, A. F. (2025). Integrated targeted and untargeted metabolomics profiling of *Vanilla* species from the Atlantic Forest: Unveiling the bioeconomic potential of *Vanilla cribbiana*. *Food Chemistry, 464*(Part 1), 141650. https://doi.org/10.1016/j.foodchem.2024.141650

Oliveira, J. P. da S., Oliveira, R. T., Guedes, A. L., Oliveira, M. da C., & Macedo, A. F. (2022). Metabolomic studies of anthocyanins in fruits by means of a liquid chromatography coupled to mass spectrometry workflow. *Current Plant Biology*, 32, 100260. https://doi.org/10.1016/j.cpb.2022.100260

Pansarin, E. R., et al. (2023). *Vanilla* (Orchidaceae) [taxonomic contribution including Brazilian endemism]. *PhytoKeys*.

Pascual, U., Adams, W. M., Díaz, S., Lele, S., Mace, G. M., & Turnhout, E. (2021). Biodiversity and the challenge of pluralism. *Nature Sustainability*, 4, 567–572. https://doi.org/10.1038/s41893-021-00694-7

Patti, G. J., Yanes, O., & Siuzdak, G. (2012). Innovation: Metabolomics: The apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology*, 13(4), 263–269. https://doi.org/10.1038/nrm3314

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., & Araújo, M. B. (2011). *Ecological niches and geographic distributions*. Princeton University Press.

Sansone, S.-A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. L., & Thurston, M. (2019). FAIRsharing as a community approach to standards, repositories

and policies. *Nature Biotechnology*, 37(4), 358–367. https://doi.org/10.1038/s41587-019-0080-8

Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D., & McLean, J. A. (2016). Untargeted metabolomics strategies—Challenges and emerging directions. *Journal of the American Society for Mass Spectrometry*, 27(12), 1897–1905. https://doi.org/10.1007/s13361-016-1469-y

Schymanski, E. L., Jeon, J., Gulde, R., Fenner, K., Ruff, M., Singer, H. P., & Hollender, J. (2014). Identifying small molecules via high resolution mass spectrometry: Communicating confidence. *Environmental Science & Technology*, 48(4), 2097–2098. https://doi.org/10.1021/es5002105

Selvarajoo, K. (2021). The need for integrated systems biology approaches for biotechnological applications. *Biotechnology Notes*, *2*, 39–43. https://doi.org/10.1016/j.biotno.2021.08.002

Steffen, W., Richardson, K., Rockström, J., Cornell, S. E., Fetzer, I., Bennett, E. M., … Sörlin, S. (2015). Planetary boundaries: Guiding human development on a changing planet. *Science*, 347(6223), 1259855. https://doi.org/10.1126/science.1259855

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., … Robinson, G. E. (2015). Big data: Astronomical or genomical? *PLoS Biology*, 13(7), e1002195. https://doi.org/10.1371/journal.pbio.1002195

Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., … Viant, M. R. (2007). Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3(3), 211–221. https://doi.org/10.1007/s11306-007-0082-2

Sutherland, W. J., Dicks, L. V., Ockendon, N., & Smith, R. K. (2018). *What works in conservation 2018*. Open Book Publishers. https://doi.org/10.11647/OBP.0131

Viant, M. R., Kurland, I. J., Jones, M. R., & Dunn, W. B. (2017). How close are we to complete annotation of metabolomes? *Current Opinion in Chemical Biology*, 36, 64–69. https://doi.org/10.1016/j.cbpa.2017.01.001

Wink, M. (2003). Evolution of secondary metabolites from an ecological and molecular phylogenetic perspective. *Phytochemistry*, 64(1), 3–19. https://doi.org/10.1016/S0031-9422(03)00300-5

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18

Wolfender, J.-L., Glauser, G., Boccard, J., & Rudaz, S. (2009). MS-based plant metabolomic approaches for biomarker discovery. *Natural Product Communications*, *4*(10), 1417–1430. https://doi.org/10.1177/1934578X0900401019

Wolfender, J.-L., Nuzillard, J.-M., van der Hooft, J. J. J., Renault, J.-H., & Bertrand, S. (2019). Accelerating metabolite identification in natural product research: Toward an ideal combination of liquid chromatography–high-resolution tandem mass spectrometry and NMR profiling, *in silico* databases, and chemometrics. *Analytical Chemistry*, 91(1), 704–742. https://doi.org/10.1021/acs.analchem.8b05112

Zhao, S., Zhang, Z., Gao, C., Dong, Y., Jing, Z., Du, L., & Hou, X. (2025). MaxEnt-based predictions of suitable potential distribution of *Leymus secalinus* under current and future climate change. *Plants, 14*(2), 293. https://doi.org/10.3390/plants14020293

5. Apêndice

5.2. Script do LigTMap (ligtmap.py) refatorado

```
from __future__ import division

import csv
import os
import sys

if os.environ.get("LIGTMAP_DEBUG_SIM", "").lower() in ("1", "true", "yes"):
    print("DEBUG_SIM mode enabled")


def extract_smiles_from_line(line):
    stripped = line.strip()
    if not stripped:
        raise ValueError("Empty SMILES line")

    # Preferred: tab-separated CID, Name, SMILES
    if "\t" in stripped:
        parts = [field.strip().strip('"').strip("'") for field in stripped.split("\t")]
        if len(parts) >= 3:
            cid, name = parts[0], parts[1]
            smi = parts[-1]
            return cid, name, smi

    if "\t" in stripped:
        fields = [field.strip().strip('"').strip("'") for field in stripped.split("\t") if
field.strip()]
        if fields:
            return None, None, fields[-1]

    if "," in stripped:
        fields = next(csv.reader([stripped], skipinitialspace=True))
        cleaned = [field.strip().strip('"').strip("'") for field in fields if field.strip()]
        if cleaned:
            return None, None, cleaned[-1]

    parts = stripped.split()
    if parts:
        return None, None, parts[-1].strip('"').strip("'")

    raise ValueError("Could not parse SMILES from line: {}".format(line))


def read_first_smiles(smile_path):
    with open(smile_path, "r") as handle:
        for raw_line in handle:
            if raw_line.strip():
                return extract_smiles_from_line(raw_line)
    raise ValueError("No SMILES found in {}".format(smile_path))


def main_part():
    smile_inf_name = sys.argv[1]
    input_num = sys.argv[2]
    tanifing_str = os.environ.get("LIGTMAP_TANIMOTO_CUTOFF", "0.85")
    try:
        tanifingcut = float(tanifing_str)
    except Exception:
        tanifingcut = 0.4
    select_db = sys.argv[3]
    total_db = int(sys.argv[4])
    rootpa = sys.argv[5]
    output_path = "Output/Input_" + input_num + "/" + select_db
    summary_path = "Output/Input_" + input_num
```

```
    cid, molname, smile = read_first_smiles(smile_inf_name)
    if cid is None:
        cid = os.path.splitext(os.path.basename(smile_inf_name))[0]
    if molname is None:
        molname = ""
    inp = smile
    output_list = []
    all_candidates = []
    fallback_top_n = int(os.environ.get("LIGTMAP_FALLBACK_TOPN", "20"))
    debug_errors = []

    output_root = "Output"
    os.makedirs(output_root, exist_ok=True)
    status_file_path = os.path.join(output_root, "status.txt")
    status_f = None
    try:
        status_f = open(status_file_path, "a")
        import rdkit  # noqa: F401
        import openbabel  # noqa: F401
        import pybel  # noqa: F401
        import oddt
        from rdkit import Chem
        from rdkit.Chem.EState import Fingerprinter
        from rdkit.Chem import Descriptors
        from rdkit.Chem.rdmolops import RDKFingerprint
        import pandas as pd
        import numpy as np
        from sklearn.preprocessing import StandardScaler

        # Prefer native pychem if importable; otherwise provide a minimal rdkit-backed shim.
        try:
            import pychem  # type: ignore
            from pychem.pychem import Chem as pychem_Chem  # type: ignore
            from pychem import fingerprint as pychem_fingerprint  # type: ignore
            PYCH_AVAILABLE = True
        except Exception:
            PYCH_AVAILABLE = False

            class _ShimChem:
                @staticmethod
                def MolFromSmiles(smi):
                    return Chem.MolFromSmiles(smi)

            class _ShimFingerprint:
                @staticmethod
                def CalculateMorganFingerprint(mol, radius=2):
                    fp = Chem.AllChem.GetMorganFingerprint(mol, radius)
                    return (None, None, fp)

                @staticmethod
                def CalculateMACCSFingerprint(mol):
                    from rdkit.Chem import MACCSkeys

                    fp = MACCSkeys.GenMACCSKeys(mol)
                    return (None, None, fp)

                @staticmethod
                def CalculateDaylightFingerprint(mol):
                    fp = Chem.RDKFingerprint(mol)
                    return (None, None, fp)

                @staticmethod
                def CalculateSimilarity(fp1, fp2, metric="Tanimoto"):
                    # Use appropriate similarity depending on fingerprint types
                    if isinstance(fp1, rdkit.DataStructs.cDataStructs.UIntSparseIntVect) or
isinstance(
                        fp2, rdkit.DataStructs.cDataStructs.UIntSparseIntVect
```

```
                    ):
                        return rdkit.DataStructs.TanimotoSimilarity(fp1, fp2)
                    return Chem.DataStructs.FingerprintSimilarity(fp1, fp2)

            pychem_Chem = _ShimChem
            pychem_fingerprint = _ShimFingerprint

        from rdkit import Chem, DataStructs, RDConfig
        from rdkit.Chem import AllChem
        from rdkit.Chem import ChemicalFeatures
        from rdkit.Chem.Pharm2D import Gobbi_Pharm2D, Generate
        from rdkit.Chem.rdMolDescriptors import GetHashedAtomPairFingerprintAsBitVect,
GetHashedTopologicalTorsionFingerprintAsBitVect
        from rdkit.Chem.AtomPairs.Sheridan import GetBPFingerprint
        from rdkit.Chem.EState.Fingerprinter import FingerprintMol
        from rdkit.Avalon.pyAvalonTools import GetAvalonFP  # GetAvalonCountFP  #int vector
version
        from rdkit.Chem.AllChem import GetMorganFingerprintAsBitVect, GetErGFingerprint
        from rdkit.DataStructs.cDataStructs import ConvertToNumpyArray
        import rdkit.DataStructs.cDataStructs
        from sklearn.metrics import r2_score
        from numpy.random import randn
        from numpy.random import seed
        from scipy.stats import pearsonr
        import csv
        import subprocess
        import pickle
        import copy

        from six.moves import zip_longest
        from itertools import chain
        from collections import OrderedDict
        import numpy as np
        from scipy.sparse import csr_matrix, isspmatrix_csr
        import oddt
        from oddt.utils import is_openbabel_molecule
        from oddt.interactions import (
            pi_stacking,
            hbond_acceptor_donor,
            hbonds,
            salt_bridge_plus_minus,
            hydrophobic_contacts,
            acceptor_metal,
            close_contacts,
        )
        from oddt import toolkit
        from oddt import shape
        from oddt import fingerprints
        from rdkit.Chem import Draw
        from oddt import interactions

        __all__ = [
            "InteractionFingerprint",
            "SimpleInteractionFingerprint",
            "SPLIF",
            "similarity_SPLIF",
            "ECFP",
            "PLEC",
            "dice",
            "tanimoto",
            "close_contacts",
            "hbond_acceptor_donor",
            "hbonds",
            "halogenbond_acceptor_halogen",
            "halogenbonds",
            "pi_stacking",
            "salt_bridge_plus_minus",
            "salt_bridges",
```

```
            "hydrophobic_contacts",
            "pi_cation",
            "acceptor_metal",
            "pi_metal",
        ]

        print("Step 1: Ligand Similarity Search")
        # PART1: Similarity <--> Tanifing + Tanipharm Score

        try:
            # Generate the fingerprint
            # factory = Gobbi_Pharm2D.factory
            mol1 = pychem_Chem.MolFromSmiles(inp)
            AllChem.EmbedMolecule(mol1, useRandomCoords=True)
            # Precompute fingerprints with RDKit for compatibility
            FP_inp_Morgan = AllChem.GetMorganFingerprint(mol1, 2)
            from rdkit.Chem import MACCSkeys

            FP_inp_MACCSF = MACCSkeys.GenMACCSKeys(mol1)
            FP_inp_Daylight = RDKFingerprint(mol1)
        except Exception:
            with open(summary_path + "/smile_err.dat", "w"):
                pass
        if os.environ.get("LIGTMAP_DEBUG_SIM", "").lower() in ("1", "true", "yes"):
            print("DEBUG_SIM fingerprints ready")
        # --- Similarity search using RDKit fingerprints ---
        id_ligname = {}
        id_proname = {}
        index_processed = 0
        index_path = os.path.join(rootpa, "Index", select_db + ".csv")
        with open(index_path, "r") as index_file:
            label_line = index_file.readline()
            for y in index_file:
                index_processed += 1
                y = y.rstrip()
                parts = y.split(";")
                if len(parts) < 5:
                    continue
                data_pdb, data_smile, affinity = parts[0], parts[1], parts[2]
                id_proname[data_pdb] = parts[3]
                id_ligname[data_pdb] = parts[4]
                # Load stored fingerprints; skip if missing
                morgan_path = os.path.join(rootpa, "Index/fing/Morgan", select_db, data_pdb +
".bin")
                maccs_path = os.path.join(rootpa, "Index/fing/MACCSF/New", select_db, data_pdb
+ ".dat")
                if not (os.path.isfile(morgan_path) and os.path.isfile(maccs_path)):
                    debug_errors.append("MISSING_FP:%s" % data_pdb)
                    continue

                with open(morgan_path, "rb") as data_morgan_file:
                    data_morgan_bin = data_morgan_file.read()
                    data_morgan = DataStructs.UIntSparseIntVect(data_morgan_bin)

                with open(maccs_path, "r", encoding="utf-8", errors="ignore") as
data_maccsf_file:
                    data_maccsf_bin = data_maccsf_file.readline().rstrip()
                    data_maccsf = DataStructs.CreateFromBitString(data_maccsf_bin)

                Tanifing_Morgan = DataStructs.TanimotoSimilarity(FP_inp_Morgan, data_morgan)
                Tanifing_MACCSF = DataStructs.FingerprintSimilarity(FP_inp_MACCSF,
data_maccsf)

                if total_db > 1:
                    with open(os.path.join(rootpa, "Index/fing/Daylight", select_db, data_pdb
+ ".dat"), "r", encoding="utf-8", errors="ignore") as data_daylight_file:
                        data_daylight_bin = data_daylight_file.readline().rstrip()
                        data_daylight = DataStructs.CreateFromBitString(data_daylight_bin)
```

```
                    Tanifing_Daylight = DataStructs.FingerprintSimilarity(FP_inp_Daylight,
data_daylight)
                    sum_score = (Tanifing_Morgan + Tanifing_MACCSF + Tanifing_Daylight) / 3
                else:
                    sum_score = (Tanifing_Morgan + Tanifing_MACCSF) / 2

                candidate_row = [
                    inp,
                    data_smile,
                    data_pdb,
                    affinity,
                    str(Tanifing_Morgan),
                    str(Tanifing_MACCSF),
                    str(sum_score),
                ]
                all_candidates.append((sum_score, candidate_row))
                if sum_score >= tanifingcut:
                    output_list.append(candidate_row)

        # PART1 COMPLETED
        print("DEBUG summary raw counts: candidates=%d above_cutoff=%d errors=%d" %
(len(all_candidates), len(output_list), len(debug_errors)))
        if all_candidates:
            all_candidates_sorted = sorted(all_candidates, key=lambda s: s[0], reverse=True)
            top_score = all_candidates_sorted[0][0]
            top_ids = [row[2] for _, row in all_candidates_sorted[:5]]
        else:
            all_candidates_sorted = []
            top_score = "n/a"
            top_ids = []

        print(
            "Similarity summary: candidates=%d above_cutoff=%d cutoff=%s top_score=%s
top_ids=%s processed=%d errors=%d"
            % (len(all_candidates), len(output_list), tanifingcut, top_score,
",".join(top_ids), index_processed, len(debug_errors))
        )
        print("DEBUG_SIM after summary print")
        # Persist debug info per target for inspection
        os.makedirs(output_path, exist_ok=True)
        with open(os.path.join(output_path, "similarity_debug.txt"), "w") as dbg:
            dbg.write(
                "candidates=%d above_cutoff=%d cutoff=%s top_score=%s top_ids=%s\n"
                % (len(all_candidates), len(output_list), tanifingcut, top_score,
",".join(top_ids))
            )
        if debug_errors:
            print("Encountered %d errors in similarity step, see error.txt" %
len(debug_errors))
            with open(os.path.join(output_path, "error.txt"), "a") as error_file:
                for err in debug_errors[:5]:
                    error_file.write("EXC:%s\n" % err)

        # Always cap to top-N to avoid huge docking runs and ensure we keep highest scores
        if all_candidates and fallback_top_n > 0:
            all_candidates.sort(key=lambda s: s[0], reverse=True)
            output_list = [row for _, row in all_candidates[:fallback_top_n]]

        if not output_list:
            # No ligands passed similarity cutoff; write placeholder outputs and exit
gracefully.
            os.makedirs(output_path, exist_ok=True)
            out_csv = os.path.join(output_path, "output.csv")
            with open(out_csv, "w") as out_file:

out_file.write("Input;Data;PDB;Affinity;Tani_Morgan;Tani_MACCSF;LigandScore;ILDScore\n")
            status_f.write(input_num + ":" + select_db + ":NoHits\n")
            status_f.close()
```

```
            print(
                "No candidates passed similarity cutoff; skipping docking/prediction. "
                "all_candidates=%d, cutoff=%s, top_score=%s"
                % (
                    len(all_candidates),
                    tanifingcut,
                    all_candidates[0][0] if all_candidates else "n/a",
                )
            )
            return

        print("Step 2: Docking")
        # PART2: Docking
        docking_ok = False
        try:
            # Prepare the input ligand
            command = ["obabel", "-:" + smile, "-opdb", "-O", output_path + "/input.pdb", "--
gen3d"]
            try:
                subprocess.check_output(command, stderr=subprocess.STDOUT)
            except Exception:
                # fallback to RDKit 3D generation
                try:
                    mol = Chem.AddHs(Chem.MolFromSmiles(smile))
                    AllChem.EmbedMolecule(mol, useRandomCoords=True)
                    AllChem.UFFOptimizeMolecule(mol)
                    Chem.MolToPDBFile(mol, output_path + "/input.pdb")
                except Exception as rd_exc:
                    raise rd_exc

            # prepare ligand for docking
            command = [
                os.environ.get("MGLTools") + "/bin/pythonsh",
                os.environ.get("MGLTools") +
"/MGLToolsPckgs/AutoDockTools/Utilities24/prepare_ligand4.py",
                "-l",
                output_path + "/input.pdb",
                "-o",
                output_path + "/input.pdbqt",
                "-A",
                "hydrogens",
                "-U",
                "nphs_lps_waters",
            ]
            subprocess.check_output(command, stderr=subprocess.STDOUT)

            # Get the PDBID
            pid_list = []
            with open(output_path + "/file_list", "w") as out_file:
                for ele in output_list:
                    out_file.write(ele[2] + "\n")
                    pid_list.append(ele[2])

            # Docking
            command = [rootpa + "docking_files/dock.sh", "pso", output_path, "1", "1", rootpa]
            subprocess.check_call(command, stderr=subprocess.STDOUT)
            docking_ok = True
        except Exception as dock_exc:
            docking_ok = False
            os.makedirs(output_path, exist_ok=True)
            with open(os.path.join(output_path, "docking_error.txt"), "a") as df:
                df.write(str(dock_exc) + "\n")
            print("Docking failed, skipping docking-dependent steps: {}".format(dock_exc))

        # Write the docking score to the output.csv
        if docking_ok:
            with open(output_path + "/DOCK_LOG/score_1.dat", "r") as score_file:
                score_list = score_file.readlines()
```

```
            with open(output_path + "/output.csv", "w") as out_file:

out_file.write("CID;Name;Input;Data;PDB;Affinity;Tani_Morgan;Tani_MACCSF;LigandScore;ILDScore\
n")
                i = 0
                for line in output_list:
                    out_line = ";".join([cid, molname] + line)
                    idscore = score_list[i].rstrip()
                    if idscore != "":
                        out_line = out_line + ";" + idscore + "\n"
                    else:
                        out_line = out_line + ";n.a.\n"
                    out_file.write(out_line)
                    i = i + 1
        else:
            # docking failed: write output with zero dock scores
            with open(output_path + "/output.csv", "w") as out_file:

out_file.write("CID;Name;Input;Data;PDB;Affinity;Tani_Morgan;Tani_MACCSF;LigandScore;ILDScore\
n")
                for line in output_list:
                    out_line = ";".join([cid, molname] + line)
                    out_line = out_line + ";0.0\n"
                    out_file.write(out_line)

        # Create complex file if docking succeeded
        if docking_ok:
            os.makedirs(output_path + "/Complex", exist_ok=True)
            for pdbid in pid_list:
                ligand_filename = output_path + "/DOCK_LOG/" + pdbid + "/" + pdbid +
"_ligand_1.pdb"
                protein_filename = rootpa + "docking_files/protein/" + pdbid + "_protein.pdb"
                filenames = [protein_filename, ligand_filename]
                output_filename = output_path + "/Complex/complex_" + pdbid + ".pdb"
                with open(output_filename, "w") as outfile:
                    for fname in filenames:
                        if os.path.isfile(fname):
                            with open(fname, "r") as infile:
                                outfile.write(infile.read())

        print("Step 3: Activity Prediction")
        # PART3: Activity prediction
        # Read the data, affinity prediction
        csv_name = output_path + "/output.csv"
        data = pd.read_csv(csv_name, sep=";", header=[0], encoding="latin1")

        # Add some new columns
        data["Mol"] = data["Input"].apply(Chem.MolFromSmiles)
        num_mols = len(data)

        def MorganFingerprint(mol):
            return FingerprintMol(mol)[0]

        # Scale X to unit variance and zero mean
        data["Fingerprint"] = data["Mol"].apply(MorganFingerprint)

        X = np.array(list(data["Fingerprint"]))

        st = StandardScaler()
        X = np.array(list(data["Fingerprint"]))
        Test = X

        def ExplicitBitVect_to_NumpyArray(bitvector):
            bitstring = bitvector.ToBitString()
            intmap = map(int, bitstring)
            return np.array(list(intmap))
```

```
        class fingerprint:
            def __init__(self, fp_fun, name):
                self.fp_fun = fp_fun
                self.name = name
                self.x = []

            def apply_fp(self, mols):
                for mol in mols:
                    fp = self.fp_fun(mol)
                    if isinstance(fp, tuple):
                        fp = np.array(list(fp[0]))
                    if isinstance(fp, rdkit.DataStructs.cDataStructs.ExplicitBitVect):
                        fp = ExplicitBitVect_to_NumpyArray(fp)
                    if isinstance(fp, rdkit.DataStructs.cDataStructs.IntSparseIntVect):
                        fp = np.array(list(fp))

                    self.x += [fp]

                    if str(type(self.x[0])) != "<class 'numpy.ndarray'>":
                        print("WARNING: type for ", self.name, "is ", type(self.x[0]))

        # Load the model
        with open(rootpa + "Model/" + select_db + ".sav", "rb") as f:
            try:
                Model = pickle.load(f, encoding="latin1")
            except TypeError:
                Model = pickle.load(f)

        predictions = Model.predict(Test)

        # PART3 COMPLETED

        if not globals().get("docking_ok", True):
            status_f.write(input_num + ":" + select_db + ":NoDock\n")
            status_f.close()
            print("Skipping binding similarity because docking failed.")
            return

        countline = -1
        with open(output_path + "/output.csv", "r") as countlinefile:
            for line in countlinefile:
                countline += 1

        print("Step 4: Binding Similarity Search (Total " + str(countline) + " target pdb,
please wait...)")
        # PART4: Similarity <--> Binding fingerprint

        def tanimoto(a, b, sparse=False):
            if sparse:
                a = np.unique(a)
                b = np.unique(b)
                a_b = float(len(np.intersect1d(a, b, assume_unique=True)))
                denominator = len(a) + len(b) - a_b
                if denominator > 0:
                    return a_b / denominator
            else:
                a = a.astype(bool)
                b = b.astype(bool)
                a_b = (a & b).sum().astype(float)
                denominator = a.sum() + b.sum() - a_b
                if denominator > 0:
                    return a_b / denominator
            return 0.0

        IFP_filename = summary_path + "/IFP_result.csv"
        with open(output_path + "/output.csv", "r") as first_result_table:
            ignore_label = first_result_table.readline()
            for line in first_result_table:
```

```
                line_list = line.rstrip().split(";")
                if len(line_list) < 10:
                    continue
                # output.csv header:
CID;Name;Input;Data;PDB;Affinity;Tani_Morgan;Tani_MACCSF;LigandScore;ILDScore
                r_cid = line_list[0]
                r_name = line_list[1]
                r_smiles = line_list[2] if len(line_list) > 2 else ""
                r_pdbid = line_list[4]
                r_ligandscore = line_list[8]
                r_dockscore = line_list[9]

                bindscore = None
                if os.path.isfile(rootpa + "Index/IFP/" + select_db + "/" + r_pdbid + ".bin")
and os.path.isfile(
                    output_path + "/DOCK_LOG/" + r_pdbid + "/" + r_pdbid + "_ligand_1.pdb"
                ):
                    crystal_IFP = np.fromfile(rootpa + "Index/IFP/" + select_db + "/" +
r_pdbid + ".bin", dtype=np.uint8)
                    bind_protein = next(oddt.toolkit.readfile("pdb", rootpa +
"docking_files/protein/" + r_pdbid + "_protein.pdb"))
                    bind_protein.protein = True
                    bind_ligand = next(oddt.toolkit.readfile("pdb", output_path + "/DOCK_LOG/"
+ r_pdbid + "/" + r_pdbid + "_ligand_1.pdb"))
                    IFP = fingerprints.InteractionFingerprint(bind_ligand, bind_protein)
                    if crystal_IFP.shape == IFP.shape:
                        bindscore = tanimoto(crystal_IFP, IFP)

                if bindscore is not None:
                    ligtmapscore = 0.7 * float(r_ligandscore) + 0.3 * bindscore
                else:
                    ligtmapscore = float(r_ligandscore)

                # Write the binding fingerprint score to the IFP_result.csv (sort)
                if not os.path.isfile(IFP_filename):
                    with open(IFP_filename, "w") as IFP_file:
                        IFP_file.write(

"CID;Name;SMILES;PDB;Class;TargetName;LigandName;LigandSimilarityScore;BindingSimilarityScore;
LigTMapScore;PredictedAffinity;DockingScore\n"
                        )
                        IFP_file.write(
                            r_cid
                            + ";"
                            + r_name
                            + ";"
                            + r_smiles
                            + ";"
                            + r_pdbid
                            + ";"
                            + select_db
                            + ";"
                            + id_proname[r_pdbid]
                            + ";"
                            + id_ligname[r_pdbid]
                            + ";"
                            + str(round(float(r_ligandscore), 6))
                            + ";"
                            + (str(round(bindscore, 6)) if bindscore is not None else "n.a.")
                            + ";"
                            + str(round(ligtmapscore, 6))
                            + ";"
                            + str(round(predictions[0], 6))
                            + ";"
                            + str(round(float(r_dockscore), 6))
                            + "\n"
                        )
                else:
```

```
                    tmp_list = []
                    with open(IFP_filename, "r") as tmp_file:
                        label_line = tmp_file.readline()
                        for line in tmp_file.readlines():
                            line = line.rstrip()
                            line = line.split(";")
                            tmp_list.append(line)
                        tmp_list.append(
                            [
                                r_cid,
                                r_name,
                                r_smiles,
                                r_pdbid,
                                select_db,
                                id_proname[r_pdbid],
                                id_ligname[r_pdbid],
                                str(round(float(r_ligandscore), 6)),
                                str(round(bindscore, 6)) if bindscore is not None else "n.a.",
                                str(round(ligtmapscore, 6)),
                                str(round(predictions[0], 6)),
                                str(round(float(r_dockscore), 6)),
                            ]
                        )
                    tmp_list.sort(key=lambda s: float(s[9]), reverse=True)
                    with open(IFP_filename, "w") as IFP_file:
                        IFP_file.write(label_line)
                        for ele in tmp_list:
                            write_str = ";".join(ele)
                            IFP_file.write(write_str + "\n")
                # missing IFP handled above; nothing to append here
        # PART4 COMPLETED

        status_f.write(input_num + ":" + select_db + ":Complete\n")
        status_f.close()
        print("DONE")
    except Exception as e:
        status_f = status_f or open(status_file_path, "a")
        status_f.write(input_num + ":" + select_db + ":Fail\n")
        status_f.close()
        print("NO RESULT: {}".format(e))


main_part()
```