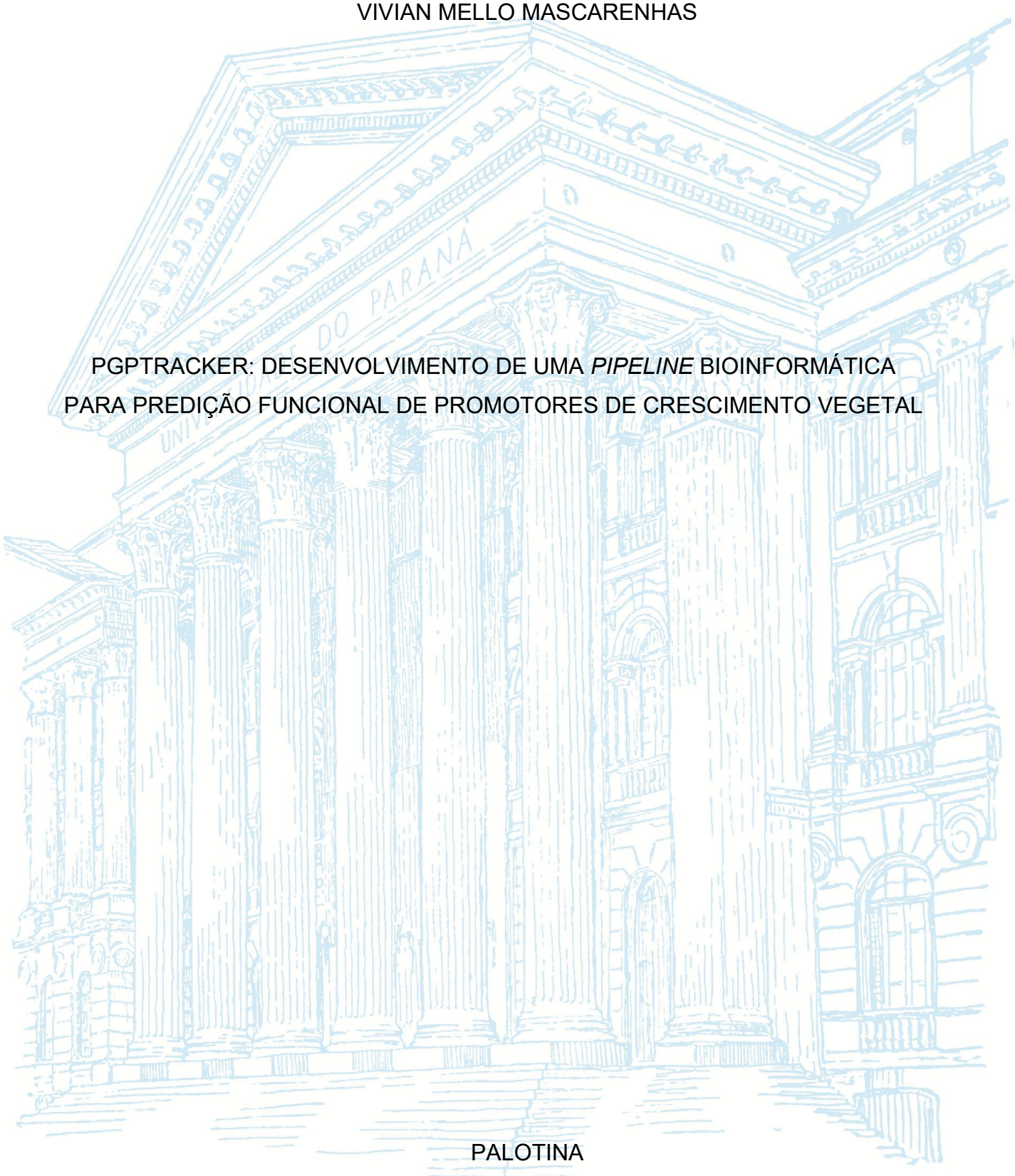
UNIVERSIDADE FEDERAL DO PARANÁ

VIVIAN MELLO MASCARENHAS



PGPTRACKER: DESENVOLVIMENTO DE UMA *PIPELINE* BIOINFORMÁTICA PARA PREDIÇÃO FUNCIONAL DE PROMOTORES DE CRESCIMENTO VEGETAL

PALOTINA

2025

Vivian Mello Mascarenhas

PGPTRACKER: DESENVOLVIMENTO DE UMA *PIPELINE* BIOINFORMÁTICA PARA PREDIÇÃO FUNCIONAL DE PROMOTORES DE CRESCIMENTO VEGETAL

Trabalho de conclusão de curso apresentado ao curso de Engenharia de Bioprocessos e Biotecnologia, Setor de Palotina, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Bacharel em Engenharia de Bioprocessos e Biotecnologia.

Orientador: Prof. Dr. Marco Antônio Bacellar Barreiros

PALOTINA

2025

UNIVERSIDADE FEDERAL DO PARANÁ
DEPARTAMENTO DE BIOCIÊNCIAS
Rua Pioneiro, 2153, - - Bairro Jardim Dallas,
Palotina/PR, CEP 85953-128
Telefone: 3360-5000 - https://ufpr.br/

ATA DE REUNIÃO

Aos **5 dias do mês de dezembro do ano de dois mil e vinte e cinco, às 13:30 horas** , na Sala **21 do** Bloco Didático **04**, Universidade Federal do Paraná, Setor Palotina, realizou-se a Defesa Pública e Oral do Trabalho de Conclusão de Curso intitulado " **PGPTRACKER: DESENVOLVIMENTO DE UMA PIPELINE BIOINFORMÁTICA PARA PREDIÇÃO FUNCIONAL DE PROMOTORES DE CRESCIMENTO VEGETAL** " apresentado **pelo(a)** discente **VIVIAN MELLO MASCARENHAS, orientado pelo Prof. Dr. Marco Antônio Bacellar Barreiros**, como um dos requisitos obrigatórios para conclusão do curso de graduação em Engenharia de Bioprocessos e Biotecnologia. Iniciados os trabalhos, o orientador e Presidente da Banca concedeu a palavra ao discente, para exposição do seu trabalho. A seguir, foi concedida a palavra em ordem sucessiva aos membros da Banca de Exame, os quais passaram a arguir o discente. Ultimada a defesa, que se desenvolveu nos termos normativos, a Banca de Exame, em sessão secreta, passou aos trabalhos de julgamento, tendo atribuído ao discente as seguintes notas: **Prof. Julio Cezar da Silva Ferreira, nota: 100 (cem)** , **Prof(a). Dr(a). Luciana Grange, nota: 100 (cem)**, e **Prof. Dr. Marco Antônio Bacellar Barreiros, nota: 100 (100)** . A nota final do discente, após a média aritmética dos três membros da banca de exame, foi **100 (cem)**. As considerações e sugestões feitas pela Banca de Exame deverão ser atendidas pelo discente sob acompanhamento de seu orientador. Nada mais havendo a tratar foi lavrada a presente ata, que, lida e aprovada, vai por todos assinada eletronicamente.

Documento assinado eletronicamente por **MARCO ANTONIO BACELLAR BARREIROS**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 13/12/2025, às 14:11, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **LUCIANA GRANGE**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 16/12/2025, às 15:48, conforme art. 1º, III, "b", da Lei 11.419/2006.

Documento assinado eletronicamente por **JULIO CEZAR DA SILVA FERREIRA**, **PROFESSOR DO MAGISTERIO SUPERIOR**, em 23/12/2025, às 23:23, conforme art. 1º, III, "b", da Lei 11.419/2006.

A autenticidade do documento pode ser conferida aqui informando o código verificador **8470416** e o código CRC **7E1E2781**.

RESUMO

O microbioma do solo desempenha um papel crucial na agricultura sustentável, oferecendo serviços ecossistêmicos vitais por meio de características de promoção de crescimento vegetal. No entanto, a complexidade de traduzir dados taxonômicos de 16S rRNA em insights funcionais permanece um desafio, frequentemente limitado por ferramentas computacionais que exigem infraestrutura de alto desempenho e pela dificuldade de correlacionar esses dois tipos de dados. Assim, este trabalho apresenta o desenvolvimento do PGPTracker (Plant Growth-Promoter Tracker), uma interface de linha de comando bioinformática projetada para conectar a taxonomia microbiana à funcionalidade potencial. A ferramenta implementa um *pipeline* em dois estágios: (1) Processamento das sequências do usuário, no qual se utilizam envoltórios de algoritmos estabelecidos do QIIME2 e PICRUSt2 para classificação taxonômica e predição dos KEGG Orthologs (KOs). Em seguida, esses dados são relacionados à PLaBAse. Os resultados desse estágio compreendem duas tabelas de abundância: uma que indica quais PGPTs estão presentes em cada amostra e outra que relaciona quais táxons são responsáveis pela produção de cada PGPT. (2) Análise, em que primeiro se aplica a normalização *Centered Log-Ratio* (CLR), seguida por testes estatísticos de diversidade funcional, testes de hipótese (por exemplo, Kruskal–Wallis), abordagens de aprendizado de máquina (Random Forest, Boruta) e geração de visualizações integradas. Por fim, uma interface gráfica permite que o usuário explore visualmente como seus dados se relacionam com os PGPTs. A ferramenta foi validada utilizando dados do Earth Microbiome Project (EMP), demonstrando capacidade de processar grandes volumes de dados em hardware de especificações moderadas (64GB RAM, 8 vCPUs). O PGPTracker oferece uma solução acessível e robusta para pesquisadores que desejam correlacionar a composição microbiana com promotores de crescimento vegetal, incluindo análise estratificada que atribui contribuições funcionais a táxons específicos.

Palavras-chave: Microbioma do solo; PGP; Inferência funcional; 16S rRNA; Earth Microbiome Project; PICRUSt2; PLaBAse.

ABSTRACT

The soil microbiome plays a crucial role in sustainable agriculture, providing vital ecosystem services through plant growth-promoting traits. However, the complexity of translating 16S rRNA taxonomic data into functional insights remains a challenge, often limited by computational tools that require high-performance infrastructure and by the difficulty of correlating these two types of data. This work presents the development of PGPTracker (Plant Growth-Promoter Tracker), a bioinformatics command-line interface designed to connect microbial taxonomy to potential functionality. The tool implements a two-stage pipeline: (1) Processing of user sequences, using wrappers around established QIIME2 and PICRUSt2 algorithms for taxonomic classification and prediction of KEGG Orthologs (KOs), which are then mapped to PLaBAse. The final outputs of this stage are two abundance tables: one indicating which PGPTs are present in each sample, and another showing which taxa are responsible for producing each PGPT. (2) Analysis, in which the Centered Log-Ratio (CLR) normalization is applied, followed by statistical tests of functional diversity, hypothesis testing (for example, Kruskal–Wallis), machine-learning approaches (Random Forest, Boruta), and integrated visualizations. Finally, a graphical interface allows the user to visually explore how their data relates to PGPTs. The tool was validated using data from the Earth Microbiome Project (EMP), demonstrating the ability to process large datasets on moderate hardware (64 GB RAM, 8 vCPUs). PGPTracker provides an accessible and robust solution for researchers aiming to correlate microbial community composition with plant growth-promoting traits, including stratified analysis that assigns functional contributions to specific taxa.

Keywords: Soil microbiome; PGP; Functional inference; 16s rRNA; Earth Microbiome Project; PICRUSt2; PLaBAse.

LISTA DE SIGLAS

ACC	1-aminociclopropano-1-carboxilato
ASV	Amplicon Sequence Variants
BIOM	Biological Observation Matrix
CLI	Command Line Interface
CLR	Centered Log-Ratio
COG	Clusters of Orthologous Groups
DADA2	Divisive Amplicon Denoising Algorithm 2
DEBLUR	Divisive Error-correction and Abundance Learning Using Random forests
DNA	Ácido Desoxirribonucleico
E.C	Enzyme Commission
EMP	Earth Microbiome Project
FDR	False Discovery Rate
GAPPA	Genesis and Placement Phylogenetic Analysis
GB	Gigabyte
GTDB	Genome Taxonomy Database
GUI	Graphical User Interface
HSP	Hidden State Prediction
IAA	Indole-3-Acetic Acid (Ácido Indol-3-Acético)
IMG	Integrated Microbial Genomes
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Ortholog
LASSO	Least Absolute Shrinkage and Selection Operator
NSTI	Nearest Sequenced Taxon Index
OTU	Operational Taxonomic Units
PCA	Principal Component Analysis
PERMANOVA	Permutational Multivariate Analysis of Variance
PGP	Plant Growth-Promoting
PGPT	Plant Growth-Promoting Traits
PICRUSt2	Phylogenetic Investigation of Communities by Reconstruction of Unobserved States 2

PLaBAse	Plant-Associated Bacteria Database
PyPI	Python Package Index
QIIME 2	Quantitative Insights Into Microbial Ecology 2
RAM	Random Access Memory
RNA	Ácido Ribonucleico
rRNA	RNA ribossomal
SEPP	SATé-enabled Phylogenetic Placement
t-SNE	t-Distributed Stochastic Neighbor Embedding
vCPU	Virtual Central Processing Unit
VM	Virtual Machine

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTO E PROBLEMA

O desenvolvimento de métodos avançados e baratos de sequenciamento genético vem permitindo nas últimas décadas a construção massiva de bancos de dados de DNA, bem como estudos de correlações taxonômicas e funcionais (Escobar-Zepeda et al., 2015). Esta revolução metodológica possibilitou a identificação de Amplicon Sequence Variants (ASVs), que oferecem resolução taxonômica superior aos Operational Taxonomic Units (OTUs) tradicionais, permitindo a detecção de diferenças de nucleotídeo único e fornecendo uma representação mais precisa da diversidade microbiana real (Joos et al., 2020).

Dessa forma, estudos de microbioma e agricultura especializada estão cada vez mais em alta, dado que agora é possível responder perguntas cada vez mais específicas, o que torna o plantio mais eficiente e produtivo a longo prazo. Com isso em mente, a metagenômica busca entender o microbioma do solo, que representa um dos ecossistemas mais complexos e funcionalmente diversos do planeta, abrigando comunidades microbianas que desempenham papéis fundamentais nos ciclos biogeoquímicos globais e na sustentabilidade dos sistemas agrícolas (Hartmann et al., 2015). Essa vasta biodiversidade composta principalmente por fungos, bactérias e arquéias é a força motriz por trás de funções ecossistêmicas essenciais, como a ciclagem de nutrientes e a promoção da saúde vegetal.

Neste contexto, os microrganismos dotados de características de promoção de crescimento de plantas (em inglês: PGPTs - Plant Growth-Promoting Traits) representam um componente de particular interesse científico. No entanto, a compreensão desses organismos é complexa devido à natureza interconectada das comunidades microbianas do solo, onde múltiplas interações simbióticas ocorrem simultaneamente em redes funcionais interdependentes (Adomako et al., 2022).

Os PGPs abrangem mecanismos diretos, como a fixação de nitrogênio, a solubilização de fosfato e a produção de fitormônios, e mecanismos indiretos, como o controle de patógenos e a indução de resistência na planta hospedeira. A identificação e aplicação desses microrganismos representa uma estratégia promissora para otimizar a produção agrícola e projeção dos efeitos que aquele solo terá no crescimento das plantas, afinal, estudos demonstram que bactérias

promotoras de crescimento vegetal produzem uma diversidade considerável de biomoléculas ativas, incluindo auxinas como o ácido indol-3-acético (IAA), sideróforos para quelação de ferro, enzimas ACC deaminase para regulação do estresse etilênico, e compostos envolvidos na solubilização de nutrientes essenciais como fosfato e potássio (Glick et al., 2012).

Sendo assim, o desenvolvimento de modelos preditivos capazes de integrar dados sobre composição microbiana, perfis metabólicos e funcionalidades específicas possibilitaria a otimização racional de sistemas agrícolas para objetivos diversos, incluindo a redução do uso de fertilizantes sintéticos através da seleção dirigida de comunidades microbianas benéficas (Shayanthan et al., 2022). Contudo, um desafio técnico fundamental impede o pleno aproveitamento desse potencial, embora o sequenciamento do gene 16S rRNA nos permita identificar a composição taxonômica das comunidades microbianas, existe uma lacuna entre o perfil taxonômico e a capacidade funcional real da comunidade, um desafio conhecido como a "caixa preta microbiana" do solo (Macrae, 2000; Ostos et al., 2024).

Esta lacuna metodológica manifesta-se em múltiplas dimensões. Primeiramente, a desconexão entre nomenclatura genômica, por exemplo, os genes ortólogos do KEGG (KOs), e os fenótipos ecológicos exige que pesquisadores realizem manualmente o mapeamento entre centenas de genes e suas funções biológicas específicas. Em segundo lugar, o sequenciamento shotgun permanece caro e de difícil acesso quando comparado ao sequenciamento de 16S rRNA da região v4 de bactérias, este, no entanto não fornece informação direta sobre o conteúdo gênico funcional da comunidade (Matchado et al., 2024), assim ferramentas de inferência funcional baseadas em filogenia, como o PICRUSt2, ajudam a fazer a ligação entre as sequências 16S e os KOs (Douglas et al., 2020), mas permanecem distantes da interpretação biológica direta no contexto de interações planta-microrganismo.

Além disso, a simples presença de um gênero bacteriano não garante que ele possua ou expresse os genes para uma função PGPT específica, já que diferentes linhagens bacterianas dentro do mesmo gênero podem apresentar capacidades funcionais distintas, enquanto microrganismos taxonomicamente divergentes podem compartilhar essas capacidades através de vias metabólicas homólogas (Chen et al., 2022).

Resolver esses desafios possibilitaria a seleção racional de microrganismos candidatos para desenvolvimento de inoculantes, a otimização de práticas de manejo do solo baseadas em critérios funcionais, e o avanço da compreensão fundamental das interações microbioma-planta (Patz et al., 2024). Assim, a problemática central desse trabalho reside em conectar ASVs aos PGPTs, a partir da classificação taxonômica advinda do sequenciamento da região v4 do 16S rRNA, ligando-a aos ortólogos do KEGG e estes aos PGPTs através da PLaBAse, uma base de dados disponibilizada pela universidade de Tübingen em colaboração com o Leibniz Institute of Vegetable and Ornamental Crops, que correlaciona diretamente os KOs aos PGPTs em 5 níveis de especificidade diferentes (Patz et al., 2021), e então realizar uma série de testes estatísticos voltados para compreensão de dados tão esparsos e diversos como são os da microbiota do solo.

Diante desse cenário, o objetivo deste trabalho foi desenvolver o PGPTracker, uma *pipeline* bioinformática capaz de (i) integrar dados de 16S rRNA com predição funcional baseada em PICRUSt2 e PLaBAse; (ii) gerar tabelas estratificadas Táxon × PGPT × amostra; e (iii) aplicar análises estatísticas e de aprendizado de máquina para identificar padrões funcionais associados a metadados ambientais.

2 DESENVOLVIMENTO

2.1 REFERENCIAL TEÓRICO

2.1.1 Revolução metodológica no sequenciamento de DNA

A evolução tecnológica dos métodos de sequenciamento gerou transformações estruturais na capacidade de análise de comunidades microbianas, estabelecendo os fundamentos teóricos para estudos metagenômicos contemporâneos. Segundo Escobar-Zepeda et al. (2015), essa revolução metodológica criou as condições necessárias para investigações em escala populacional que fundamentam abordagens como a utilizada neste estudo.

A redução nos custos de sequenciamento e o aumento do *throughput* possibilitaram diretamente a construção de bases de dados genéticos em larga escala. Repositórios globais como GenBank e o European Nucleotide Archive exemplificam essa expansão massiva, transformando a pesquisa genética ao suportar novas formas de análise comparativa, evolutiva e funcional (Leinonen et al., 2010). Estas bases de dados agora cobrem uma ampla gama de taxonomias e geografias, desde comunidades microbianas até grandes cortes populacionais, fornecendo a infraestrutura necessária para estudos metagenômicos em escala global como o Earth Microbiome Project (Thompson et al., 2017).

2.1.2 Avanços na resolução taxonômica: de OTUs para ASVs.

A escolha metodológica entre OTUs e ASVs possui implicações teóricas fundamentais para a precisão de predições funcionais subsequentes. Enquanto o agrupamento tradicional de OTUs aplica um threshold arbitrário de similaridade, os ASVs preservam variações biologicamente significativas que podem correlacionar com diferenças funcionais específicas (Joos et al., 2020).

Estudos comparativos demonstram consistentemente que métodos baseados em ASVs fornecem maior resolução taxonômica do que o agrupamento tradicional de OTUs em diversos habitats microbianos. Métodos como DADA2 conseguem resolver diferenças de nucleotídeo único e distinguir de forma confiável variantes biológicas verdadeiras de erros de sequenciamento, enquanto o

agrupamento de OTUs em um limiar de 97% pode mascarar diversidade em escala fina e potencialmente perder informações ecologicamente relevantes (Callahan et al., 2017).

A seleção do algoritmo DEBLUR para este estudo, em detrimento de alternativas como DADA2, fundamenta-se em considerações de escalabilidade computacional demonstradas por Amir et al. (2017). Para *datasets* de grande escala como o EMP, o DEBLUR oferece eficiência de processamento mantendo acurácia comparável na correção de erros, justificando sua adoção em análises que requerem processamento de milhares de amostras simultaneamente.

2.1.3 Complexidade funcional do microbioma do solo e PLaBAse

A base teórica para estudos de PGPTs fundamenta-se no conceito de redundância funcional em comunidades microbianas complexas, assim, a diversidade funcional em solos opera através de redes de interações que podem ser quantificadas através de métricas de co-ocorrência e associação, proporcionando o fundamento metodológico para análises integradas de perfis funcionais (Hartmann et al., 2015).

Além disso, a categorização hierárquica de PGPTs proposta por Glick et al. (2012) estabelece o *framework* conceitual para agregação de funções em análises quantitativas. Esta hierarquia permite a transição de análises gene-específicas para análises funcionais integradas, justificando a abordagem de agregação de genes ortólogos KEGG em categorias funcionais mais amplas.

Dessa forma, a base de dados PLaBAse emergiu como um recurso abrangente para anotação de PGPTs, contendo aproximadamente 6.900 PGPTs associados a quase 7 milhões de sequências proteicas (Patz et al., 2021).

2.1.4 QIIME2 e classificação taxonômica com Greengenes2

QIIME2 (Quantitative Insights Into Microbial Ecology 2) representa a segunda geração de uma das plataformas mais amplamente utilizadas para análise de dados de microbioma, desenvolvida para processar dados de sequenciamento de amplicons de forma modular, reprodutível e escalável (Bolyen et al., 2019). Assim, a plataforma opera através de plugins modulares que executam tarefas específicas

como controle de qualidade, agrupamento de sequências, classificação taxonômica e análises de diversidade, permitindo construção de workflows customizados que atendem necessidades experimentais variadas.

Neste trabalho, o QIIME2 foi utilizado para implementar a conversão de formatos de arquivos característicos de dados metagenômicos e para classificação taxonômica das ASVs. A classificação taxonômica no QIIME2 é realizada através do *plugin* 'feature-classifier', que implementa o algoritmo 'classify-sklearn' de aprendizado de máquina para atribuir taxonomia a sequências representativas de ASVs ou OTUs.

Atualmente, o QIIME2 suporta de forma oficial o Greengenes2, ele contém uma variedade de classificadores treinados em diferentes regiões de RNA (McDonald et al., 2024). Os dados são retirados principalmente da GTDB (Genome Taxonomy Database), uma iniciativa que busca padronizar e classificar o máximo de genomas, mesmo aqueles sem nomes ou publicações oficiais (Parks et al., 2018). O classificador utiliza o método Naive Bayes, onde os dados são treinados em bases de dados específicas (no caso, a GTDB), que aprende padrões de k-mers característicos de diferentes táxons e calcula probabilidades posteriores de pertencimento taxonômico para cada sequência *query* (Bokulich et al., 2018).

Dessa forma, o Greengenes2 disponibiliza um classificador pré-treinado na região v4 do 16S rRNA otimizada para uso direto no QIIME2.

2.1.5 Inferência funcional a partir de dados filogenéticos usando PICRUSt2

O princípio teórico subjacente à predição funcional filogenética baseia-se na conservação evolutiva de características metabólicas entre organismos relacionados. Contudo, como observado por Breitkreuz et al. (2021), esta premissa apresenta limitações quando aplicada a comunidades edáficas, onde transferência horizontal de genes e pressões seletivas locais podem resultar em discrepâncias entre relacionamento filogenético e capacidade funcional. Esse fenômeno é descrito como 'a caixa preta microbiana', sendo um desafio de fronteira da metagenômica atualmente (Nobu et al., 2015; Ostos et al., 2024).

O PICRUSt2 busca solucionar em partes esse dilema. A ferramenta usa posicionamento filogenético de ASVs em uma árvore de referência contendo sequências de 20.000 genes 16S rRNA de comprimento total de genomas

bacterianos e arqueais no banco de dados Integrated Microbial Genomes (IMG). Esta abordagem possibilita predição de números de cópias de genes funcionais e abundâncias de vias com base em relacionamentos evolutivos (Douglas et al., 2020).

O fluxo de trabalho do PICRUSt2 consiste em três etapas principais: (1) posicionamento filogenético usando SEPP (SATé-enabled Phylogenetic Placement), (2) predição de estado oculto (Hidden Markov Models) para inferir abundâncias de genes ortólogos KEGG, e (3) inferência de abundância de vias através de mapeamentos estruturados. O Nearest Sequenced Taxon Index (NSTI) serve como uma métrica de qualidade, com valores <2.0 geralmente considerados aceitáveis para predições confiáveis (Douglas et al., 2020).

2.1.6 Validação e limitações dos métodos de predição funcional

De acordo com o próprio artigo do Douglas (2020) criador do PICRUSt2, a confiança da ferramenta na predição de KOs (conjuntos de genes de funções específicas) a partir de dados de sequenciamento de 16S rRNA é relativamente alta, refletida pelo valor de correlação de Spearman que varia de uma média de 0,79 a 0,88, dependendo do conjunto de dados analisado. Essas correlações indicam uma forte concordância entre as abundâncias preditas e as observadas via metagenômica *shotgun*, sugerindo um bom nível de confiança nas predições do PICRUSt2 para esses dados.

Entretanto, o estudo também destaca que a precisão das predições de funções, incluindo os KOs, é moderada e apresenta limitações, especialmente na detecção de funções específicas em ambientes pouco representados por genomas referenciados. Ainda assim, o método mostrou desempenho superior a outras abordagens, especialmente em ambientes de microbiomas não associados a humanos.

2.1.7 Análise de comunidades microbianas e dados composicionais

Os dados de abundância relativa derivados de sequenciamento 16S rRNA apresentam características composicionais que requerem tratamento estatístico especializado. A natureza composicional implica que a abundância de um

componente influencia artificialmente a abundância relativa dos demais componentes, criando correlações espúrias (Aitchison, 1982).

Assim, a transformação *Centered Log-Ratio* (CLR) emerge como abordagem padrão para linearizar relações composicionais e atender pressupostos de normalidade requeridos por análises estatísticas multivariadas. Ela expressa cada abundância como logaritmo de sua razão pela média geométrica de todas as abundâncias, o que transforma os dados do simplex para o espaço euclidiano enquanto preserva as relações de proporcionalidade entre componentes e remove a restrição de soma constante (Zhang et al., 2024).

2.1.8 Earth microbiome project e padronização metodológica

O Earth Microbiome Project (EMP) representa iniciativa global de caracterização sistemática da diversidade microbiana, tendo gerado dados padronizados de sequenciamento 16S rRNA para mais de 200.000 amostras de biomas diversos. O protocolo utiliza primers universais 515F-806R para amplificação da região V4 do gene 16S rRNA, com metadados ambientais abrangentes que permitem análises integradas de fatores ambientais e composição funcional (Thompson et al., 2017).

2.1.9 Polars e processamento eficiente de grandes conjuntos de dados

Polars é uma biblioteca moderna de processamento de DataFrames desenvolvida em Rust e projetada para análise de dados de larga escala em ambiente de máquina única. Ao contrário de bibliotecas tradicionais como Pandas, que são construídas sobre NumPy e operam predominantemente em modo *single-threaded*, Polars foi arquitetado desde sua concepção para explorar paralelismo massivo através de execução *multi-core* segura e garantida pelas características da linguagem Rust. A biblioteca utiliza o formato de memória Apache Arrow como *backend* nativo, padrão emergente para análises colunares *in-memory* que elimina custos de serialização/desserialização entre etapas de *pipelines* de dados, reduzindo consumo de memória RAM em 50-80% comparado ao Pandas e acelerando operações comuns em fatores de 5 a 100 vezes (Vink et al., 2025)

Uma característica distintiva fundamental do Polars é o suporte nativo a execução lazy (preguiçosa), onde operações são registradas em um plano de consulta otimizado ao invés de serem executadas imediatamente. O otimizador de consultas analisa automaticamente o conjunto de operações, identificando oportunidades para reordenação de execuções, eliminação de cálculos redundantes, e aplicação de predicados de filtragem antes de operações custosas como junções ou agregações. Esta abordagem contrasta radicalmente com a execução *eager* (imediata) obrigatória do Pandas, onde transformações são sempre executadas sequencialmente na ordem escrita pelo usuário, sem possibilidade de otimização global e automática (Vink et al., 2025).

No contexto do PGPTracker, a adoção de Polars como motor de processamento de dados é justificada pela necessidade de manipular tabelas estratificadas taxonomicamente de dimensões massivas, onde *datasets* típicos envolvem dezenas de milhares de ASVs multiplicados por múltiplos níveis taxonômicos e dezenas de amostras, gerando matrizes que facilmente excedem centenas de milhares de linhas. Dessa forma, a capacidade do Polars de executar transformações CLR, pivotagens de formato e agregações tabulares enquanto explora todos os núcleos de CPU disponíveis e otimiza o consumo de memória, permite reduzir o pico de uso de RAM, simultaneamente acelerando o processamento de minutos para segundos e viabilizando a execução da ferramenta em estações de trabalho mais convencionais e acessíveis, ao invés de servidores de alto desempenho.

2.2 METODOLOGIA

2.2.1 Obtenção e processamento inicial dos dados

Para suportar o processamento de dados metagenômicos em larga escala, foi estabelecida uma infraestrutura baseada em computação em nuvem utilizando o Google Cloud Platform; uma máquina virtual foi configurada com sistema operacional Linux Ubuntu, com 8 núcleos de vCPU e 64 GB de memória RAM. Esta configuração representa o ambiente recomendado para processamento de *datasets* de tamanho moderado, em torno de 70 amostras com aproximadamente 15.000

ASVs, embora o PGPTracker possa operar em máquinas com especificações inferiores dependendo do tamanho do *dataset.*

2.2.2 Conjunto de dados e critérios de seleção

O estudo utilizou dados do Earth Microbiome Project (EMP), obtidos do repositório Zenodo, compreendendo tabelas de abundância de ASVs processadas através do *pipeline* DEBLUR com truncamento a 90 pares de bases, sequências representativas e metadados ambientais abrangentes. O tratamento inicial foi realizado utilizando QIIME2 versão 2025.10.

A seleção de amostras foi realizada aplicando-se critérios de garantia para a homogeneidade ambiental e completude de metadados. Primeiramente, aplicou-se o filtro '*empo_3='Soil (non-saline)*' sobre os metadados, selecionando exclusivamente solos não-salinos, depois aplicou-se um filtro para retenção apenas de amostras com valores preenchidos para pH e temperatura, a fim de garantir conformidade nos testes analíticos subsequentes. Para demonstração específica das capacidades do PGPTracker, um subconjunto final foi selecionado contendo aproximadamente 12.257 ASVs distribuídas através de 66 amostras representativas de diferentes biomas e gradientes ambientais.

2.2.3 Arquitetura do PGPTracker

O PGPTracker foi desenvolvido como ferramenta bioinformática integrada implementada em Python 3.13, estruturada como pacote distribuível via PyPI. A arquitetura do sistema organiza-se em dois estágios principais de processamento. (i) Estágio 1, voltado à inferência funcional e ao mapeamento das ASVs em KOs e PGPTs; e (ii) Estágio 2, dedicado às análises estatísticas multivariadas e ao aprendizado de máquina. Adicionalmente, uma interface gráfica baseada em Streamlit oferece capacidades de exploração interativa dos resultados.

A implementação utiliza a biblioteca Polars para manipulação eficiente de dados, substituindo pandas em todos os módulos autorais devido a melhor performance em operações de agregação sobre *datasets* maiores e esparsos. As bibliotecas Python utilizadas incluem scikit-bio para cálculos de diversidade

ecológica, scikit-learn para aprendizado de máquina, matplotlib e seaborn para visualizações estáticas, e Plotly para gráficos interativos na GUI.

Além disso, o PGPTracker implementa módulos de ferramentas externas (QIIME2 e PICRUSt2) para auxiliar em tarefas mais complexas de tratamento de dados de microbioma.

2.2.4 Conversão de formato e mapeamento de ortólogos do KEGG

O primeiro passo do estágio 1 é verificar o formato de arquivos de *input* do usuário, o PGPTracker identifica se são arquivos nativos do QIIME2 com formato '.qza' (QIIME Zipped Artifact), se positivo, transforma a tabela de sequências do usuário em um arquivo FASTA e a tabela de abundâncias em um arquivo BIOM (Biological Observation Matrix), para isso se utiliza de um envoltório (que aqui chamaremos de *wrapper*) que importa a função de conversão nativa do QIIME2.

É importante salientar que o usuário deve já ter tratado suas sequências FASTQ com algum algoritmo de demultiplexação, controle de qualidade, remoção de ruído e que tenha gerado a tabela de feições/abundâncias juntamente com a de sequências representativas para então poder utilizar o PGPTracker corretamente.

Se caso não for necessária a etapa de conversão dos arquivos QIIME, o estágio 1 começará pela importação dos *wrappers* do PICRUSt2, eles orquestram os algoritmos estabelecidos do PICRUSt2 Standalone para posicionamento filogenético e predição funcional, mantendo a validade metodológica das abordagens validadas por Douglas et al. (2020). Os scripts 'place_seqs.py' e 'hsp_prediction.py' executam os binários especializados SEPP e scripts R que utilizam o pacote castor para cálculos filogenéticos.

O módulo *wrapper* place_seqs.py invoca o binário run_sepp.py do PICRUSt2 para posicionamento filogenético das ASVs através do algoritmo SEPP, que gera uma árvore filogenética de referência em formato 'jplace', depois converte para formato 'tre' via GAPPA, para fins de compatibilidade. Por conseguinte, o hsp_prediction.py gerencia a execução de scripts R customizados advindos do próprio PICRUSt2, seus papéis são de predizer estados ocultos (Hidden State Prediction) para inferir abundâncias de genes ortólogos KEGG, fazem isso através de máxima parcimônia e do cálculo de NSTI, o que mantém a confiabilidade alta.

Esse passo do estágio 1 gera duas saídas importantes, primeiro a ‘KO_predicted.tsv.gz’, uma tabela estratificada que contém as previsões das abundâncias relativas dos KOs para cada ASV, e a segunda é a ‘marker_nsti_predicted.tsv’, um arquivo que fornece o indicador NSTI das previsões relacionadas às abundâncias de marcadores filogenéticos.

2.2.5 Normalização de abundâncias por comprimento genômico

O módulo ‘gen_ko_abun.py’ implementa normalização de abundâncias seguindo a mesma lógica estabelecida do PICRUSt2, porém com otimizações algorítmicas para redução substancial de tempo de execução. Desse modo, o algoritmo ajusta contagens brutas de ASVs por comprimento de genoma previsto e número de cópias do gene 16S rRNA, isso corrige vieses introduzidos por diferenças na eficiência de amplificação e no tamanho genômico entre diferentes táxons.

A implementação utiliza operações vetorizadas do Polars para realizar a normalização, o que elimina a necessidade de loops explícitos comuns em códigos baseados em Pandas, exigindo menos poder de processamento e executando a tarefa em menos tempo. Assim, o módulo trabalha com três entradas principais: A tabela de abundâncias em formato BIOM, a tabela comprimida com as predições do gene marcador 16S e a tabela comprimida com as predições de KOs. Como saída, o processo gera dois arquivos essenciais, a tabela ‘seqtab_norm.tsv’, que contém as abundâncias normalizadas das ASVs, e o arquivo ‘pred_metagenome_unstrat.tsv’, que reúne as abundâncias funcionais não estratificadas dos KOs por amostra.

2.2.6 Classificação taxonômica via QIIME2

A classificação taxonômica das sequências representativas foi realizada pela integração da ferramenta QIIME 2, utilizando o *plugin* ‘q2-feature-classifier’. Assim, o envoltório aplica a classificação usando o método ‘Classify-sklearn’ em cima da tabela de abundâncias já normalizada gerada pelo módulo anterior, isso garante que as atribuições taxonômicas sejam aplicadas às mesmas ASVs cujas abundâncias foram corrigidas por comprimento genômico.

O classificador ‘backbone.v4.nb.fna’ utilizado foi lançado em setembro de 2024 e corresponde à região V4 do gene 16S rRNA, ele foi construído com Naive Bayes e disponibilizado oficialmente pelo QIIME2 em parceria com o Greengenes2.

2.2.7 Integração de taxonomia com abundâncias normalizadas

A próxima etapa do estágio 1 reside no módulo ‘merge_tax_abun.py’, também programado em Polars, ele integra as atribuições taxonômicas geradas pela etapa anterior com a tabela de abundâncias normalizadas. Para isso, o algoritmo valida e compara a coluna de ASVs entre as duas tabelas, anexa colunas de hierarquia taxonômica completa (Reino, Filo, Classe, Ordem, Família, Gênero, Espécie) à tabela normalizada, e exporta a matriz integrada ‘norm_wt_feature_table.tsv’, que servirá como entrada crítica para as etapas subsequentes.

Esta integração é essencial pois estabelece a ponte entre identidade taxonômica e potencial funcional ao nível de ASV individual, permitindo que as agregações subsequentes atribuam contribuições funcionais a níveis taxonômicos específicos. O módulo implementa verificações de integridade incluindo detecção de ASVs não-classificadas e validação de completude da hierarquia taxonômica.

2.2.8 Mapeamento funcional de KOs para PGPTs

Para realizar o mapeamento de KOs para PGPTs, o método empregado foi o de correlacionar a tabela ‘KO_predicted.tsv.gz’ gerada pelo PICRUSt2 (ligação entre ASVs e KOs), à base de dados PLaBAse, que contém a ligação entre KOs (presentes na coluna ‘PGPT_ID’ da Figura 1) e PGPTs em 5 níveis diferentes de especificidade, que podem ser escolhidos pelo usuário. Dessa forma, a união de tabelas se faz possível e permite identificar quais ASVs produzem quais PGPTs, além de suas abundâncias relativas. Abaixo, a Figura 1 demonstra como a PLaBAse se parece:

Figura 1 – amostra da base de dados PLaBAse e classificações funcionais

S.NO	ID	PGPT_ID	BACTERIA	PATH_ COUNT	Lv1	Lv2	Lv3	Lv4	Lv5
25	PGPT00 00105	PGPT0000105-vnfD-K22896	Azospirillum brasilense Sp245	1	DIRECT_ EFFECTS	BIO-FERTILIZATION	NITROGEN_A CQUISITION	N-AQUISITION-ATMOSHPHERIC_NITROGEN_FIXATION	N-FIX-NITROGENASE_BIOSYNTHESIS
27	PGPT00 00110	PGPT0000110-vnfE-K22903	Azotobacter vinelandii DJ	1	DIRECT_ EFFECTS	BIO-FERTILIZATION	NITROGEN_A CQUISITION	N-AQUISITION-ATMOSHPHERIC_NITROGEN_FIXATION	N-FIX-NITROGENASE_BIOSYNTHESIS
55	PGPT00 00225	PGPT0000225-hupU-K23548	Rhizobium leguminosarum Norway,Bradyrhizobium diazoefficiens USDA 110	1	DIRECT_ EFFECTS	BIO-FERTILIZATION	NITROGEN_A CQUISITION	N-AQUISITION-ATMOSHPHERIC_NITROGEN_FIXATION	N-FIX-HYDROGENASE_BIOSYNTHESIS
91	PGPT00 00405	PGPT0000405-nasS-K22067	Pseudomonas sp. UW4	1	DIRECT_ EFFECTS	BIO-FERTILIZATION	NITROGEN_A CQUISITION	N-AQUISITION-DENITRIFICATION|NITRATE_USAGE	DENITRIFICATION-NITRATE_REDUCTION
219	PGPT00 00900	PGPT0000900-tgnR-K23466	Pseudomonas fluorescens F113	2	DIRECT_ EFFECTS	BIO-FERTILIZATION	NITROGEN_A CQUISITION	N-AQUISITION-TRIGONELLINE_USAGE	N-AQUISITION-TRIGONELLINE_METABOLISM

Fonte: Adaptado de PATZ, S. et al. (2021).

Durante a execução do trabalho, dois métodos de mapeamento mostraram-se necessários, nomeados como não estratificado e estratificado. Eles permitem a geração de duas tabelas, a primeira possui as amostras do usuário nas colunas e os PGPTs encontrados nas linhas e é em formato amplo (*wide*) tradicional, ela responde qual o potencial total de um PGPT dentro de uma amostra. A segunda é em formato longo (*long*) e possui quatro colunas: Taxon, PGPT, amostra e abundância, indicando de forma estratificada, qual Taxon produz qual PGPT dentro de cada amostra e qual sua abundância.

Além disso, a ferramenta permite a personalização do nível de especificidade dos PGPTs para ambos os formatos de tabela. No caso da análise estratificada, é possível também definir o nível taxonômico de agrupamento. Entretanto, deve-se considerar que níveis taxonômicos mais finos geram tabelas significativamente maiores, elevando o custo computacional da análise.

2.2.9 Estratificação

A construção da tabela estratificada foi implementada através de um algoritmo personalizado desenvolvido com a biblioteca Polars, estruturado para operar sob demanda (*lazy evaluation*) e processamento em fluxo (*streaming*). Dessa forma, o processo de vinculação taxonômica-funcional ocorre através de três etapas de agregação sequenciais.

A primeira função chamada ‘aggregate_by_tax_level_sample’ agrupa a tabela de abundâncias normalizada via o parâmetro ‘tax-level’, selecionado pelo usuário, enquanto também agrupa por amostra, somando as abundâncias de todas as ASVs que compartilham a mesma atribuição taxonômica, o que gera uma matriz Taxon x Amostra. A operação utiliza *groupby* do Polars com agregação *sum*, evitando materialização de *DataFrames* intermediários desnecessários.

A segunda etapa realiza a mesma operação da primeira, mas para a tabela de KOs, que traduz a quantidade de KO codificada para cada táxon e gera uma matriz Taxon x KO. A implementação utiliza operações de junção seguidas de *groupby* em modo *lazy*, permitindo que Polars otimize o plano de execução antes de materializar resultados, o que é crítico para o gerenciamento de memória RAM.

Na terceira função, ocorre uma tripla vinculação entre as matrizes, correlacionando a matriz 1 (Taxon x Amostra) com a matriz 2 (Taxon x KO) e mapeando essa junção com a PLaBAse (KO x PGPT) para unir aos PGPTs. Essa operação gera todas as combinações possíveis de Taxon × PGPT × Amostra onde existe contribuição funcional não-zero. O processamento ocorre em partes (*batches*) e faz as junções de tabelas massivas sem exceder memória disponível.

2.2.10 Estágio 2: Análises estatísticas multivariadas

O Estágio 2 implementa um conjunto abrangente de análises estatísticas aplicadas aos dados funcionais gerados na etapa anterior. Por utilizarem métodos especializados em dados metagenômicos, essas análises permitem a caracterização multidimensional dos perfis de PGPTs, bem como a identificação de padrões de variação associados aos metadados selecionados. Além disso, a arquitetura deste estágio foi projetada para maximizar eficiência computacional através de um sistema de pré-processamento centralizado que gera todas as transformações e formatos de dados necessários uma única vez.

Vale pontuar que todos os testes estatísticos realizados tiveram sua veracidade checada seguindo os métodos padrões de validação em programação, para isso foram criados diversos testes unitários e utilizado bibliotecas comuns para tal, como Pytest, Mock e MagicMock.

2.2.10.1 Aplicação de CLR no pré-processamento

A primeira etapa do Estágio 2 consiste na transformação CLR dos dados. O módulo clr_normalize.py atua como um orquestrador central, processando as tabelas geradas no estágio anterior. Sua função é gerar todas as saídas necessárias para as análises subsequentes, adequando os formatos de entrada às exigências específicas de cada teste estatístico. Esta abordagem modular evita transformações de matriz desnecessárias durante as análises, o que economiza poder de processamento.

O processo inicia-se pela aplicação do método multiplicativo de substituição, uma técnica essencial para dados composicionais que substitui valores zero por quantidades infinitesimais, preservando as proporções relativas dos componentes não-zero. Esta etapa é pré-requisito fundamental para a transformação CLR, visto que o cálculo envolve logaritmos naturais, matematicamente indefinidos para zero. Em seguida, aplica-se a transformação CLR propriamente dita. Ambas as operações foram executadas através de funções nativas da biblioteca Scikit-bio.

Visando atender à heterogeneidade das análises subsequentes, o módulo de pré-processamento gera oito arquivos de saída, organizados em quatro pares complementares. Inicialmente, são produzidas tabelas em formato amplo (*wide*) com orientação padrão NxD (amostras nas linhas e características nas colunas). Para esta configuração, o sistema exporta uma versão contendo as abundâncias brutas e outra com os dados transformados via CLR. Em paralelo, são criadas versões na orientação transposta DxN, essenciais para algoritmos específicos como análise de redes de co-ocorrência.

No processamento dos dados estratificados, que se apresentam em formato longo (*long*), realiza-se a pivotagem para o formato amplo, consolidando a combinação Táxon-PGPT nas colunas. Sobre esta nova estrutura, aplica-se a transformação CLR, resultando na exportação dos arquivos ‘raw_wide’ e ‘clr_wide’. Adicionalmente, é gerada uma variação da tabela ampla com colunas segregadas (*split*), preservando a estrutura tridimensional (Amostra x PGPT x Táxon) para análises tensoriais avançadas, que serão implementadas no PGPTracker no futuro.

Esta etapa é crítica pois diferentes análises requerem diferentes representações dos dados. Por exemplo, análises de diversidade beta baseadas em Aitchison necessitam dados CLR-transformados, enquanto distâncias de Bray-Curtis

operam sobre abundâncias relativas brutas; Da mesma forma, análises de ordenação como PCA requerem matrizes no formato amostras x características (NxD), enquanto certas análises de rede podem necessitar a orientação transposta características x amostras (NxD).

Esta estratégia de pré-processamento exaustivo apresenta vantagens computacionais significativas, já que a implementação integral em Polars mantém o consumo de memória RAM abaixo de 10 GB. Adicionalmente, ao executar todas as transformações uma única vez e armazená-las em disco, evita-se a replicação de DataFrames temporários em memória, que ocorreria caso cada análise realizasse suas próprias transformações independentemente. Além disso, dão a oportunidade ao usuário realizar outras análises em outros programas caso queira sem maiores dores de cabeça.

2.2.10.2 Caracterização da diversidade funcional

A diversidade funcional das comunidades microbianas foi quantificada através de métricas complementares de diversidade alfa e beta, aplicadas aos perfis de PGPTs. Assim, as métricas de diversidade alfa foram calculadas individualmente por amostra, incluindo riqueza funcional observada (número de PGPTs detectados), índice de Shannon (considerando simultaneamente riqueza e equitabilidade), índice de Simpson (enfatizando funções dominantes) e equitabilidade de Pielou (uniformidade da distribuição funcional).

A significância estatística de agrupamentos observados nas matrizes de dissimilaridade foi avaliada através de PERMANOVA (Análise Multivariada de Variância por Permutações), um teste não-paramétrico que determina se centróides de grupos definidos por metadados (por exemplo: pH do solo, tipo de bioma, temperatura etc.) são significativamente diferentes no espaço multivariado funcional. O teste opera através de permutações aleatórias das atribuições de grupo, comparando a variância entre grupos observada com a distribuição nula gerada por permutação, fornecendo p-valores robustos sem assumir normalidade multivariada.

2.2.10.3 Análises de ordenação

Para visualização e interpretação da estrutura multivariada dos perfis funcionais, foram implementadas análises de ordenação complementares. A Análise de Componentes Principais (PCA) foi aplicada aos dados CLR-transformados, identificando gradientes funcionais lineares dominantes que explicam a maior variância nos dados.

Complementarmente, o algoritmo t-SNE (t-Distributed Stochastic Neighbor Embedding) foi implementado para capturar estruturas de agrupamento não-lineares que podem não ser evidentes em projeções lineares como PCA. Além disso, o parâmetro de perplexidade, que controla o equilíbrio entre preservação de estrutura local versus global, pode ser ajustado pelo usuário, com valor padrão de 30 apropriado para a maioria dos conjuntos de dados de microbioma.

2.2.10.4 Testes estatísticos univariados

Com o objetivo de identificar características funcionais (PGPTs ou táxons) diferencialmente abundantes entre grupos ambientais, foram implementados testes estatísticos não-paramétricos adequados à natureza composicional dos dados. Estes testes constituem a base estatística necessária para a geração subsequente de visualizações, como os mapas de calor (*heatmaps*).

Especificamente, o teste de Kruskal-Wallis foi selecionado para comparações envolvendo três ou mais grupos, ao passo que o teste de Mann-Whitney U foi empregado nas comparações binárias. Ressalta-se que ambos operam individualmente sobre cada característica funcional, utilizando dados previamente transformados via CLR para mitigar vieses composicionais.

Dada a multiplicidade de testes realizados simultaneamente (potencialmente milhares de características testadas), todos os *p*-valores são sistematicamente ajustados para controle da taxa de falsas descobertas (FDR) através do método de Benjamini-Hochberg. Este procedimento calcula *q*-valores que representam a proporção esperada de falsos positivos ao rejeitar a hipótese nula para aquela característica, permitindo interpretação direta da robustez estatística das descobertas. No mais, características com $q<0,05$ são consideradas estatisticamente significativas após correção para testes múltiplos e mantidas.

2.2.10.5 Aprendizado de máquina e seleção de características

Algoritmos de aprendizado de máquina foram implementados para duas finalidades complementares: Predição de variáveis ambientais/fenotípicas a partir de perfis funcionais e a identificação de características preditivas que capturam relações não-lineares complexas. Dessa forma, o algoritmo Random Forest foi aplicado tanto para tarefas de classificação (variável-alvo categórica, como tipo de bioma) quanto regressão (variável-alvo contínua, como pH do solo), operando sobre dados CLR-transformados.

Para seleção robusta de características genuinamente preditivas versus ruído estocástico, foi implementado o algoritmo Boruta. Este método opera comparando a importância de características reais contra características shadow (versões permutadas aleatoriamente das características originais, que por construção não possuem relação com a variável-alvo). Apenas características cuja importância excede consistentemente a importância máxima das características shadow são confirmadas como genuinamente preditivas, controlando rigorosamente para falsas descobertas em contexto de alta dimensionalidade (Kursa et al., 2010).

Para tarefas de regressão, o algoritmo LASSO (Least Absolute Shrinkage and Selection Operator) com validação cruzada foi adicionalmente implementado. O LASSO realiza regularização L1, forçando os coeficientes de características não-informativas exatamente a zero, efetivamente realizando seleção automática de características enquanto ajusta um modelo linear. O parâmetro de regularização é otimizado via validação cruzada, balanceando capacidade preditiva versus parcimônia do modelo.

2.2.10.6 Visualizações integradas

Para cada análise, são geradas visualizações prontas para publicação através de bibliotecas Matplotlib e Seaborn, com opções de exportação em múltiplos formatos (PNG, PDF, SVG).

O PGPTracker também implementa duas abordagens estatisticamente distintas para plotagem de mapas de calor, cada uma respondendo às questões

científicas complementares, o que resulta na geração de dois *heatmaps* com lógicas de seleção diferentes.

Além disso, *Volcano plots* são gerados para visualização simultânea de significância estatística (q-valor) e magnitude de efeito (*fold-change* ou diferença de médias CLR) dos testes univariados, permitindo checagem visual rápida de como os testes se saíram. Já para análises de ordenação, são produzidos *biplots* que sobrepõem scores de amostras e vetores de *loadings* de características, facilitando interpretação biológica dos gradientes funcionais. Gráficos de caixa estratificados por grupo são gerados para métricas de diversidade alfa, com valores-p de testes estatísticos automaticamente anotados diretamente nos gráficos.

Adicionalmente, gráficos de importância de características do Random Forest são ordenados por ranking decrescente de importância, exibindo as principais características que diferenciam os grupos ou predizem a variável-alvo. Por fim, todas as visualizações implementam lógica condicional para coloração de pontos por variáveis de metadados (contínuas como pH ou categóricas como tipo de solo), permitindo exploração visual de associações ambiente-função.

2.2.10.7 Interface gráfica para exploração interativa

Complementarmente à interface de linha de comando, foi desenvolvida uma interface gráfica web interativa implementada em Streamlit, acessível via navegador executando em servidor local. Esta GUI (Interface Gráfica do Usuário) permite exploração visual dinâmica dos dados funcionais sem necessidade de conhecimento de programação, sendo especialmente útil para investigação exploratória e geração de hipóteses.

O módulo de upload implementa detecção automática inteligente da estrutura dos dados carregados, identificando se os dados estão em formato amplo ou longo, se foram CLR-transformados (detectado por presença de valores negativos característicos), e se contêm estratificação taxonômica. O sistema realiza junção (*merge*) automático entre dados funcionais e metadados através de identificação heurística de colunas de chave primária (padrões comuns como "Sample", "#SampleID", "sample_name") que o próprio usuário seleciona, visando a flexibilidade entre *datasets*.

O módulo de exploração gera visualizações dinâmicas através de Plotly, biblioteca que permite interatividade completa (*zoom, pan, tooltips* informativos, seleção de pontos etc.). Dessa forma, para dados não-estratificados, o usuário pode selecionar dois PGPTs diferentes para eixos X e Y através de menus *dropdown* interativos, gerando gráficos de dispersão onde cada ponto representa uma amostra individual. Os pontos podem ser coloridos dinamicamente por variáveis de metadados contínuas usando escalas de cor gradiente, ou por variáveis categóricas usando paletas de cores distintas. Para dados estratificados, a interface permite visualização de contribuições taxonômicas específicas para PGPTs selecionados, revelando quais táxons dominam funcionalmente em diferentes contextos ambientais.

Além disso, gráficos de caixa interativos permitem comparação visual de distribuições de PGPTs entre grupos definidos por metadados categóricos, com quartis e p-valores exibidos diretamente nos gráficos quando o usuário interage com eles. É importante ressaltar que a GUI não implementa filtragem por nível taxonômico ou PGPT dinamicamente durante a exploração; o usuário deve gerar tabelas pré-filtradas no nível desejado através da linha de comando dos estágios 1 e 2, carregando então essas tabelas específicas na GUI para exploração visual interativa.

3 RESULTADOS E DISCUSSÃO

3.1 CARACTERIZAÇÃO DOS DADOS PROCESSADOS

Para validar a funcionalidade da ferramenta, o PGPTracker processou com sucesso um subconjunto de validação do Earth Microbiome Project contendo 66 amostras de solos não salinos com metadados completos, abrangendo 12.257 ASVs após filtragem de rarefação. O processamento completo foi executado em 23 minutos em ambiente computacional com 64 GB de RAM e 8 vCPUs, demonstrando viabilidade de execução em hardware de especificações moderadas.

A classificação taxonômica identificou representação de 68 filos bacterianos distintos no *dataset* processado. A análise taxonômica revelou a presença de 697 famílias bacterianas. Esta diversidade taxonômica substancial oferece contexto robusto para subsequente decomposição funcional estratificada, demonstrando a capacidade da ferramenta de processar comunidades microbianas complexas características de ambientes edáficos.

As 10 famílias mais abundantes (*Baltobacteraceae, Beijerinckiaceae, Acetobacteraceae, Burkholderiales Burkholderiaceae_A_595421, Acidobacteriaceae, Paludibacteraceae, Xanthobacteraceae, Geobacterales Geobacteraceae_439684, Pseudopelobacteraceae, Limisphaerales UBA11358*) responderam coletivamente por 55,35% da abundância relativa total através das 66 amostras. As 100 famílias mais abundantes responderam por 95,89% da abundância total, enquanto as 597 famílias restantes contribuíram com apenas 4,11% da abundância total. Este padrão de distribuição é característico de estudos de microbioma do solo, onde organismos minoritários podem exercer impacto funcional desproporcional à sua abundância relativa (Zhang et al., 2022; Jousset et al., 2017).

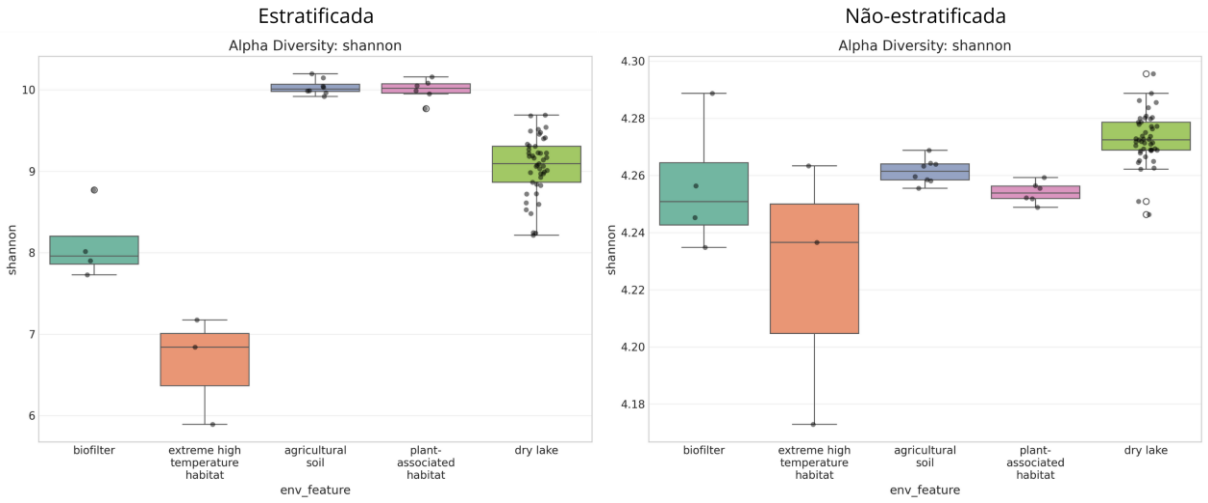
A métrica NSTI calculada para o *dataset* apresentou distribuição com média de 0,246 e desvio padrão de 0,194, indicando uma excelente proximidade filogenética aos genomas de referência. Aproximadamente 99,77% das ASVs apresentaram NSTI<1.0, correspondendo a confiabilidade moderada a alta das predições funcionais. As ASVs restantes (0,23%) apresentaram NSTI entre 1,0 e 1,7, o que ainda está dentro do limiar de aceitação estabelecido para predições, ou seja, apenas 42 ASVs das 12.257 originais foram cortadas das análises por não atenderem o NSTI. Esta distribuição favorável valida a aplicabilidade da abordagem

de inferência funcional baseada em filogenia para o *dataset* de solos não salinos, refletindo a cobertura genômica adequada deste ambiente na base de referência do PICRUSt2.

Entre os KOs preditos e os KOs específicos de metabólitos promotores de crescimento vegetal catalogados na PLaBAse, que contém mais de 7 milhões de KOs (KEGG Ortholog) pareados com 6.700 PGPTs diferentes, todos classificados em 5 níveis hierárquicos ontológicos, houve uma convergência de 6.214 KOs. Esta interseção resultou na detecção de 43 PGPTs diferentes no nível 3 selecionado para os testes, gerando uma tabela não estratificada de 66 colunas (amostras) e 43 linhas (PGPTs), presente nos apêndices. Com isso, a cobertura funcional adequada demonstra que o mapeamento entre inferência filogenética e base de dados especializada captura diversidade funcional relevante para processos de promoção de crescimento vegetal em solos.

A análise de diversidade alfa do potencial funcional, explorada no Nível 3 de PGPT, revelou uma alta riqueza de vias e elevada uniformidade na distribuição das funções. Os índices consistentemente altos, como o de Shannon (Figura 2), indicam que o PGPTracker mapeou uma vasta gama de PGPTs na comunidade, sugerindo uma redundância funcional. Esta característica é fundamental para a resiliência do sistema, pois garante que múltiplas vias PGP estejam disponíveis para manter processos críticos, como a ciclagem de nutrientes, mesmo diante de perturbações que possam afetar táxons específicos.

Figura 2 – Comparação entre os índices de shannon derivados das tabelas estratificadas e não estratificadas

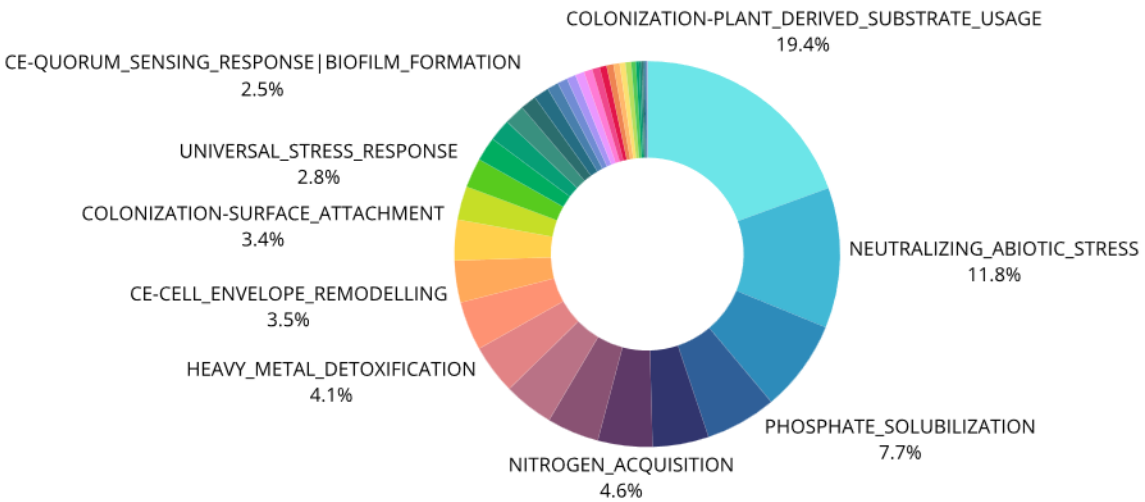


Fonte: A autora (2025)

Contudo, ao comparar as abordagens, observa-se uma diferença na magnitude dos valores do índice de Shannon nas análises não estratificada (média 4,2) e estratificada (média 9,0). Essa discrepância se deve à mudança na unidade de análise, enquanto a abordagem não estratificada limita-se a contar os tipos de funções presentes, a estratificada considera cada associação única entre uma família bacteriana e uma função como uma categoria distinta. Dessa forma, o aumento no índice reflete matematicamente a própria redundância funcional mencionada anteriormente, onde cada via metabólica é desmembrada em milhares de pares "táxon-função", evidenciando a diversidade de agentes biológicos que sustentam o repertório funcional da comunidade.

Adicionalmente, a quantificação agregada das vias (Figura 3) demonstrou que as funções mais representadas em termos de abundância total foram: Colonização e uso de substratos de origem vegetal (‘COLONIZATION PLANT_DERIVED_SUBSTRATE_USAGE’), neutralização de estresses abióticos (‘NEUTRALIZING_ABIOTIC_STRESS’) e solubilização de fosfato (‘PHOSPHATE_SOLUBILIZATION’). A identificação destas funções predominantes evidencia a capacidade do PGPTracker de mapear a estrutura da funcionalidade promotora de crescimento em comunidades complexas de solo. Métricas adicionais de diversidade alfa e uniformidade para esta análise encontram-se disponíveis no Apêndice A.

Figura 3 – Distribuição dos PGPTs em todas as amostras



Fonte: A autora (2025)

3.2 INVESTIGAÇÃO DA CAIXA PRETA MICROBIANA VIA ESTRATIFICAÇÃO

A análise funcional de microbiomas do solo enfrenta historicamente um desafio fundamental: conectar identidade taxonômica a capacidade funcional. Técnicas de sequenciamento de amplicons 16S rRNA tornaram acessível a caracterização taxonômica em larga escala, mas a ligação entre "quem está presente" e "o que fazem" permanece uma lacuna crítica na ecologia microbiana. Assim, a "caixa preta microbiana" limita a aplicabilidade de análises funcionais para seleção racional de inoculantes ou compreensão de resiliência funcional em sistemas agrícolas.

Estudos recentes demonstram as consequências práticas desta limitação. Niu et al. (2024), ao caracterizar comunidades bacterianas em solos de alta e baixa produtividade de trigo, recorreram a análises de rede de co-ocorrência para inferir táxons-chave baseados em conectividade topológica. Embora tenham identificado diferenças significativas na complexidade das redes entre os dois tipos de solo, a abordagem permanece indireta, ou seja, uma bactéria pode co-ocorrer com outras funcionalmente importantes sem ela mesma portar os genes relevantes. De forma similar, Martínez-Núñez e Orozco-Ramírez (2024) discutiram a presença de *Gemmatimonadetes* e sua importância para redução de N_2O, mas seus dados mostravam apenas a presença do filo e a presença da via metabólica separadamente, sem conexão causal direta na análise. A interpretação biológica dependeu de literatura externa, não de seus próprios dados integrados.

Tendo em vista essa problemática, ferramentas contemporâneas adotam estratégias distintas para solucioná-la. Por exemplo, Nagpal et al. (2019) desenvolveram o iVikodak, uma plataforma web que permite ao usuário investigar vias metabólicas específicas e visualizar contribuintes taxonômicos através de um módulo denominado "Local Mapper". Entretanto, esta abordagem opera sob demanda, o pesquisador deve selecionar manualmente cada via de interesse para inspeção, implicando que para análises exploratórias envolvendo dezenas ou centenas de funções, este processo torna-se laborioso e impraticável para análises multivariadas subsequentes.

Outra abordagem é a de Lu et al. (2023), que propuseram no MicrobiomeAnalyst 2.0 métodos como Análise de Procrustes, focado em integração visual. Estes métodos permitem identificar padrões de correlação entre taxonomia e

função através de redução de dimensionalidade, mas não fornecem a matriz bruta de contribuição taxonômica para cada função.

Assim, o PGPTracker aborda este problema através da geração nativa e automatizada de matrizes estratificadas completas (Táxon × PGPT × Amostra). Diferentemente de abordagens anteriores, o *pipeline* processa o fluxo completo de estratificação em lote (*batches*), viabilizado pela arquitetura de processamento em fluxo da biblioteca Polars.

Para explicar melhor como ocorre a formação da tabela estratificada, é preciso entender sua origem. Em resumo, o PICRUSt2 já produz nativamente uma tabela estratificada que conecta ASV com KOs, a ideia original era de unir essa saída nativa à PLaBAse, no entanto, o processo exigia mais de 512GB de RAM e horas de processamento para os dados de teste, logo, fez-se necessária uma otimização para que o PGPTracker pudesse cumprir seu objetivo final, dessa forma, a estratificação foi reimplementada utilizando Polars e três funções principais, operando através de três agregações principais, cada um otimizada com *lazy evaluation* e operações de *streaming* do Polars.

Logo, a primeira etapa de agregação resultou em 13.027 pares únicos de táxon (família) e amostra. Para a segunda etapa, referente à associação funcional, o produto entre todas as ASVs e os KOs únicos geraria uma matriz excessiva, estimada em 76 milhões de linhas. Para viabilizar o processamento, aplicou-se um filtro de esparsidade que removeu as entradas de abundância zero (casos em que a família não produz o KO), reduzindo o conjunto final para 2.267.517 pares únicos de táxon e KOs. Logo, a união dessas duas matrizes gerou a tabela estratificada com 516.059 linhas e quatro colunas, abaixo há um exemplo de como a Tabela 1 estratificada se parece:

Tabela 1 – Exemplo da tabela longa estratificada

Family	Lv3	Sample	Abundance
Abditibacteriaceae	IRON_ACQUISITION	1692.Biofilm.B.DC.1.W.2012	376.737630
Abditibacteriaceae	IRON_ACQUISITION	1692.Biofilm.B.E1.1.C.2012	83.799056
Abditibacteriaceae	IRON_ACQUISITION	1692.Biofilm.C.DC.1.W.2012	251.397169
Abditibacteriaceae	IRON_ACQUISITION	1692.Soil.BE.PolygonA.Medium.Center.7.15.2012	125.340460
Abditibacteriaceae	IRON_ACQUISITION	1692.Soil.BE.PolygonA.Shallow.Rim.7.15.2012	628.134809
Abditibacteriaceae	IRON_ACQUISITION	1692.Soil.BE.PolygonB.Medium.Rim.7.15.2012	4,808.060400
Abditibacteriaceae	IRON_ACQUISITION	1692.Soil.BE.PolygonB.Shallow.Rim.7.15.2012	1,237.647605
Abditibacteriaceae	IRON_ACQUISITION	1692.Soil.BE.PolygonC.Medium.Center.7.15.2012	83.799056

Fonte: A autora (2025)

Dessa forma, o tempo de execução dessa etapa foi de aproximadamente 3 minutos com consumo de memória RAM inferior a 10 GB, representando redução de mais de 98% no uso de memória comparado à implementação do PICRUSt2.

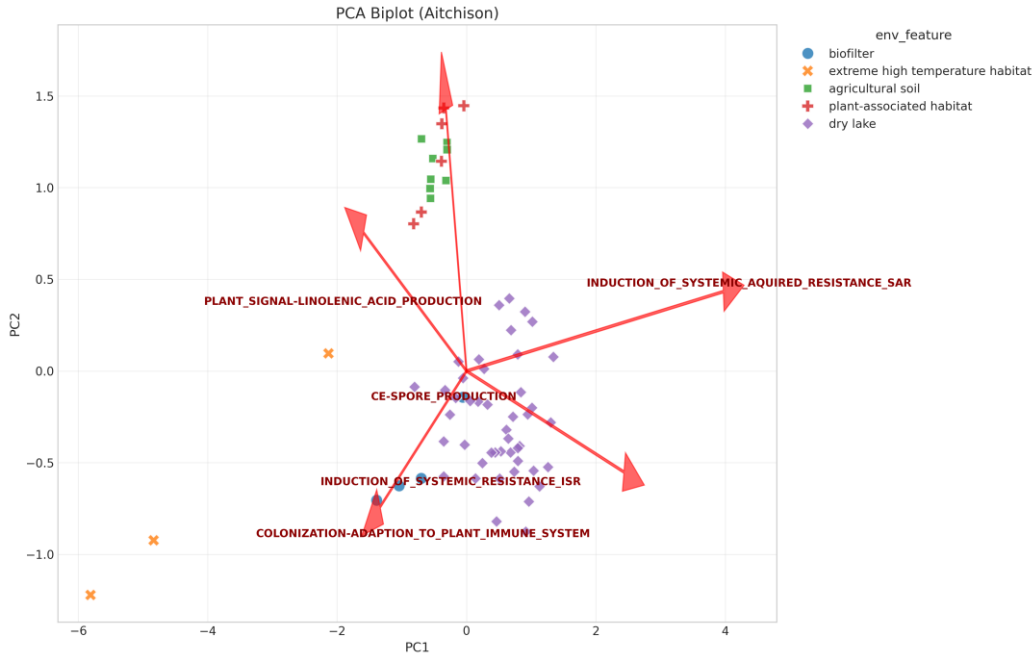
Esta otimização não apenas viabiliza a execução em hardware convencional, mas democratiza o acesso à análise funcional estratificada. A capacidade de gerar matrizes completas Táxon × PGPT × Amostra em poucos minutos permite que laboratórios com recursos limitados conduzam análises que anteriormente exigiriam infraestrutura computacional extensiva ou processos manuais laboriosos. O PGPTracker, dessa forma, transforma a estratificação funcional de uma etapa opcional ou visual pós-processamento em dado primário de saída, permitindo a descoberta não-supervisionada de biomarcadores funcionais e substituições de nicho que seriam invisíveis em análises de abundância agregada.

Além disso, esta abordagem oferece vantagens metodológicas fundamentais. Primeiro, a estratificação determinística supera métodos baseados em correlação estatística. Enquanto ferramentas como MicrobiomeAnalyst 2.0 utilizam correlação de Spearman para tentar vincular bactérias a metabólitos, o PGPTracker estabelece ligação direta através do banco de dados, onde se um ASV é mapeado filogeneticamente a um genoma de referência que contém um gene PGPT específico, a contribuição daquele táxon para aquela função é quantificada explicitamente na tabela estratificada.

Segundo, a geração matricial completa permite análises multivariadas avançadas que seriam inviáveis com extração manual. A aplicação de algoritmos de seleção de características, onde cada característica é uma combinação Táxon-PGPT específica, torna-se possível. Gonçalves et al. (2024) demonstraram manualmente que a capacidade de fixação de nitrogênio em *Acidobacteria* não é uniforme no filo, mas restrita a famílias específicas como *Acidobacteriaceae* e *Holophagaceae*. Esta análise, que exigiu montagem e anotação individual de 758 genomas metagenômicos, exemplifica o tipo de decomposição funcional que o PGPTracker automatiza para dados de amplicon 16S.

A importância dessa granularidade ficou evidente nas análises realizadas com o *dataset* de validação. Enquanto a análise global (não estratificada) apontou genericamente funções como produção de ácido linolênico como preditoras ambientais (Figura 4) a análise estratificada identificou especificamente quais famílias exerciam essas funções em cada ambiente (Figura 5).

Figura 4 – PCA biplot com vetores de loading da tabela não estratificada



Fonte: A autora (2025)

O primeiro *heatmap* estratificado (Figura 5), deriva de testes estatísticos univariados (Kruskal-Wallis com correção FDR) e opera através de análise independente de cada característica. Esta abordagem testa individualmente se a abundância de cada táxon ou PGPT difere significativamente entre grupos ambientais, sem considerar relações entre características. A pergunta científica subjacente é: "quais características variam em abundância entre os ambientes?" Cada teste estatístico avalia uma única característica isoladamente, determinando se suas abundâncias apresentam diferenças significativas entre grupos.

Características são retidas se atingirem significância estatística ($q<0,05$), independentemente de padrões de variação similares entre múltiplas características. Esta abordagem tende a gerar listas exaustivas incluindo características redundantes que apresentam padrões de resposta correlacionados, como múltiplas variantes de uma mesma família bacteriana respondendo similarmente a gradientes ambientais.

Para contornar esse problema, um mecanismo foi desenvolvido com a finalidade de que mapas de calor estatísticos fossem de fato úteis para inferir afirmações sobre os dados. Logo, implementou-se um sistema de filtragem inteligente que seleciona características para exibição através de um processo em

duas etapas: Primeiramente, retém-se apenas características estatisticamente significativas ou com alta importância preditiva; depois, aplica-se filtragem por diversidade taxonômica, selecionando as três características de maior variância dentro de cada grupo taxonômico até completar as 50 posições do gráfico.

Não aplicar essa estratégia faz com que a plotagem seja em ordem alfabética e sem filtragem prévia das características mais relevantes entre as relevantes. Por isso, esse processo assegura que todas as características exibidas são biologicamente muito relevantes, enquanto maximiza a representatividade taxonômica ao evitar dominância visual por grupos abundantes. Ademais, em *datasets* com menos de 51 características encontradas, todas são plotadas.

Figura 5 – Heatmap da tabela estratificada

Fonte: A autora (2025)

Dessa forma, observou-se um fenômeno de substituição taxonômica de nicho (*functional turnover*) na função de colonização e uso de substratos de origem vegetal (‘COLONIZATION-PLANT_DERIVED_SUBSTRATE_USAGE’). Embora esta função apresentasse alta abundância global, a decomposição estratificada revelou que no ambiente Dry Lake ela era desempenhada majoritariamente pela família *Baltobacteraceae*, enquanto em 'Agricultural Soil', a mesma função era mantida pela família *Rhizobiaceae.*

Ferramentas que operam com "sacolas de genes" (gene-bags), onde funções são agregadas sem atribuição taxonômica, seriam incapazes de detectar essa dinâmica ecológica, bem como a análise não estratificada também não seria.

Por fim, a exportação da matriz bruta para formatos tabulares padrão garante integração e capacidade de operar com ecossistemas analíticos externos. Pesquisadores podem aplicar métodos estatísticos customizados não implementados nativamente na ferramenta, contrastando com plataformas web que retêm dados internamente e oferecem apenas visualizações pré-configuradas.

3.2.1 Métodos estatísticos complementares para identificação de biomarcadores

O PGPTracker implementa múltiplas abordagens estatísticas que revelam aspectos complementares da estrutura funcional das comunidades microbianas. Diferentes métodos respondem a questões biológicas distintas, e a ferramenta permite que o pesquisador escolha a estratégia mais apropriada conforme seus objetivos analíticos.

A aplicação de análises de ordenação, como o PCA, a dados não estratificados permite identificar funções que explicam a máxima variância global na estrutura da comunidade. Frequentemente, essas funções são raras, mas apresentam alta variabilidade entre os grupos. É importante notar que raridade não implica irrelevância ecológica; uma função com baixa abundância, mas presente exclusivamente em um ambiente específico, pode atuar como um biomarcador altamente discriminatório (Jousset et al., 2017; Zhang et al., 2022) Essa abordagem é particularmente útil para gerar hipóteses sobre vias metabólicas pouco estudadas que merecem investigação mais aprofundada.

Por outro lado, o uso de algoritmos de aprendizado de máquina como Random Forest em dados estratificados permite identificar combinações específicas

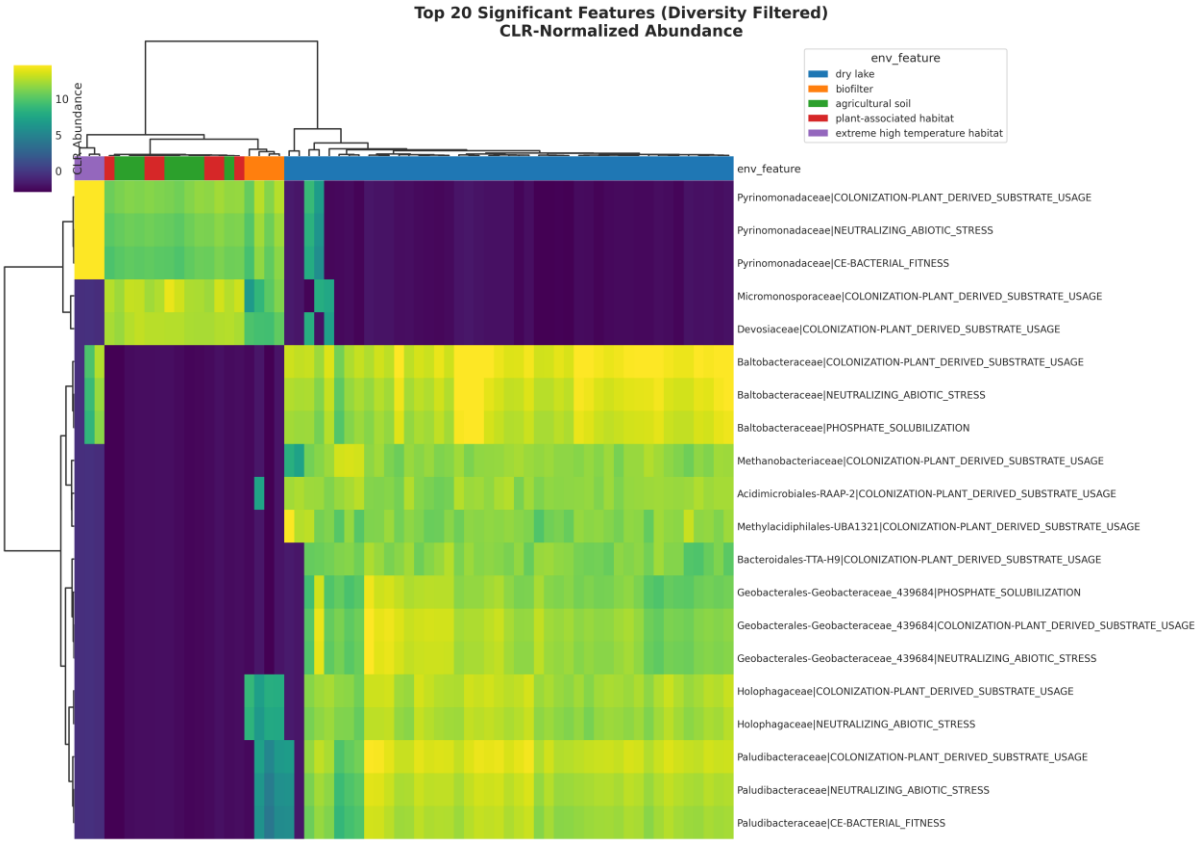
de Táxon-PGPT com alto poder discriminatório, levando em conta as interdependências entre as características, isso é possível porque ele opera através de análise conjunta de todas as características simultaneamente.

Esta abordagem identifica o subconjunto mínimo de características necessário para predição acurada da identidade ambiental de amostras, considerando explicitamente interdependências entre características. A pergunta científica subjacente é: "qual conjunto mínimo de características é suficiente para discriminar os ambientes?" O algoritmo avalia características no contexto de todas as demais, quantificando quanto cada característica contribui para redução de incerteza preditiva dado o conhecimento das características já selecionadas.

Assim, quando múltiplas características apresentam padrões correlacionados, o modelo retém aquela com maior poder discriminativo e descarta características redundantes que não adicionam informação preditiva incremental. Esta abordagem identifica sinalizadores eficientes, revelando características que capturam padrões complexos de co-variação, como combinações específicas de abundâncias que caracterizam unicamente determinados ambientes.

Desse modo, essa abordagem é capaz de detectar funções que, embora abundantes globalmente, apresentam substituição taxonômica entre ambientes, como visto na Figura 6 abaixo. A distinção conceitual aqui é fundamental, pois abundância não é sinônimo de utilidade preditiva. Uma função pode ser rara, mas discriminar bem entre grupos (detectada pelo PCA), ou pode ser abundante globalmente, mas executada por diferentes táxons em cada local (detectada pelo Random Forest estratificado).

Figura 6 – Mapa de calor estratificado originado de aprendizado de máquina



Fonte: A autora (2025)

Esse princípio foi ilustrado pela função de colonização (‘COLONIZATION-PLANT_DERIVED_SUBSTRATE_USAGE’), onde a análise global indicou que essa função seria onipresente com baixo poder discriminatório. No entanto, a análise estratificada revelou uma substituição de nicho, onde a família *Pyrinomonadaceae* desempenhava a função em Dry Lake, enquanto a *Baltobacteraceae* assumia esse papel no Biofilter.

Esse padrão taxonômico seria perdido em análises de dados agregados. A estratificação permite, portanto, estabelecer um vínculo determinístico entre o agente taxonômico e o processo funcional através de diferentes condições ambientais. Logo, evita-se a interpretação simplista de que funções com abundância constante são ecologicamente inafetadas por diferentes ambientes ou tratamentos. Isso demonstra que a aparente estabilidade de uma variável funcional entre tratamentos pode mascarar reestruturações taxonômicas, uma compreensão que permanece inacessível em análises globais sem o recurso da curadoria manual extensiva.

Dessa forma, o PGPTracker foi desenhado para integrar essas diferentes perspectivas analíticas, oferecendo flexibilidade metodológica. O pesquisador pode optar por explorar biomarcadores de abundância via Kruskal-Wallis para uma caracterização descritiva, ou empregar Random Forest para construir modelos preditivos multivariados, dependendo da pergunta de pesquisa.

3.3 COMPARAÇÃO COM FERRAMENTAS EXISTENTES

O PGPTracker situa-se em um cenário complexo de ferramentas para análise funcional de microbiomas, marcado por importantes decisões metodológicas. As ferramentas atuais oscilam entre abordagens como sequenciamento amplicon ou *shotgun*, plataformas web ou execução local, e bancos de dados amplos ou focados definem as capacidades de cada software. Entender esse panorama de escolhas é fundamental para posicionar a contribuição da ferramenta desenvolvida.

3.3.1 Automação da triagem e anotação funcional

A inferência funcional a partir de dados de amplicon 16S gera, tipicamente, milhares de KOs preditos, dos quais apenas uma fração possui relevância para processos específicos de interesse. Nityagovsky et al. (2025) ilustram esta limitação em seu estudo de comunidades endofíticas de abeto. Os autores utilizaram PICRUSt2 para prever KOs e depois aplicaram manualmente a ontologia PGPT da PLaBAse, encontrando que 59,5% dos 8.653 KOs previstos estavam associados a traços de promoção de crescimento. Este processo funciona biologicamente, mas não escala metodologicamente, ou seja, cada novo *dataset* requer curadoria manual para filtrar funções agronomicamente relevantes dos milhares de KOs generalizados.

O PGPTracker supera este desafio ao integrar nativamente a base de dados PLaBAse durante a etapa de mapeamento funcional. Aplicada ao *dataset* de validação, essa abordagem permitiu refinar o vasto universo de KOs preditos pelo PICRUSt2, filtrando automaticamente 6.214 KOs especificamente associados à promoção de crescimento vegetal. Esse processo resultou na identificação direta de 43 PGPTs interpretáveis, eliminando a necessidade de curadoria manual

subsequente, além de concluir a tarefa em 20 minutos, tempo que o estágio 1 levou para ser executado com esse conjunto de dados.

Essa capacidade de filtrar o "ruído" transforma a inferência funcional de uma análise genérica em uma ferramenta útil para a agronomia. O foco em características que já possuem validação experimental garante que os resultados tenham sentido biológico, facilitando o uso dessas informações no desenvolvimento de produtos biotecnológicos, como novos inoculantes ou práticas de manejo do microbioma.

3.3.2 Amplicon 16S versus Shotgun Metagenomics

O PGPg_finder, desenvolvido por Pellegrinetti et al. (2024), opera na mesma lógica conceitual do PGPTracker ao utilizar a PLaBAse para inferência de PGPTs, mas baseia-se em dados de sequenciamento *shotgun* metagenômico. Essa diferença vai de encontro ao trade-off tradicional da metagenômica, já que dados *shotgun* oferecem resolução funcional superior por detectarem a presença real dos genes, o que permite análises de expressão independente da inferência filogenética. Contudo, apesar da precisão inerente ao sequenciamento *shotgun,* esse método enfrenta duas barreiras: A primeira é monetária, sequenciar genomas inteiros é caro e pouco acessível quando comparado ao sequenciamento Amplicon 16S. A segunda é o tempo, como relatado pelos próprios autores do PGPg_finder, foram necessários 267 minutos de processamento para apenas 8 amostras.

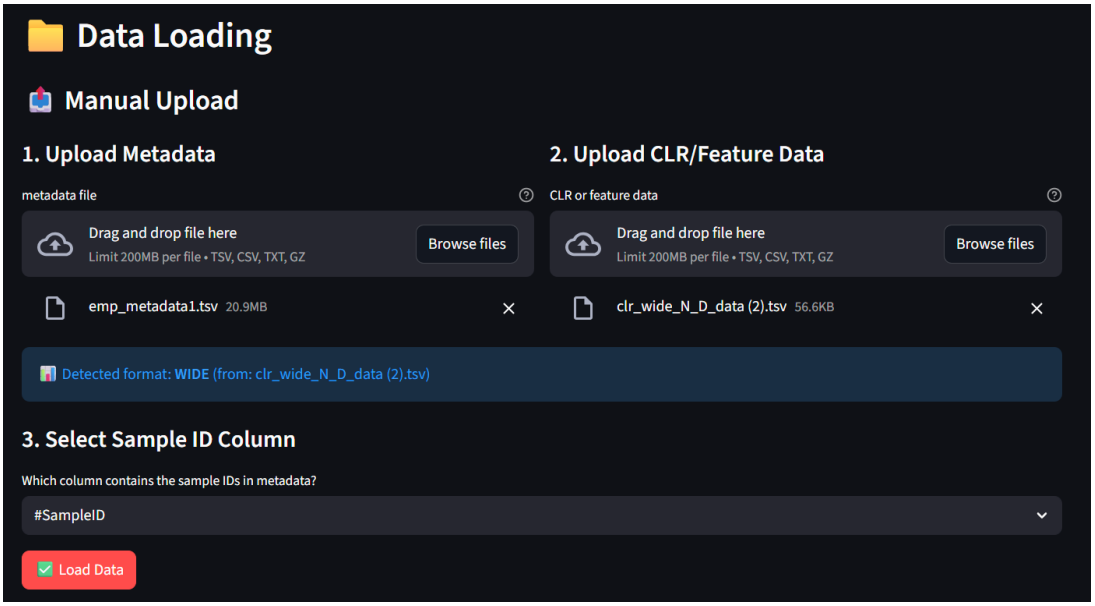
Diante disso, o PGPTracker serve como uma ferramenta complementar de triagem. A inferência por 16S, embora sujeita a limitações de bancos de referência, oferece custo-benefício favorável para identificação de padrões funcionais em larga escala. Dessa forma, propõe-se que o PGPTracker seja usado para direcionar os esforços, identificando as amostras que realmente justificam o investimento em uma análise mais profunda via *shotgun*, maximizando assim a eficiência dos recursos de pesquisa.

3.3.3 Arquitetura local versus plataformas web

A arquitetura do PGPTracker, desenhada como uma ferramenta de linha de comando (CLI) com uma interface gráfica (GUI) opcional, reflete uma escolha

deliberada para equilibrar acessibilidade e escalabilidade. É verdade que plataformas baseadas na web, como o iVikodak (Nagpal et al., 2019) e o MicrobiomeAnalyst 2.0 (Lu et al., 2023), facilitam o acesso inicial ao dispensar instalações locais e apresentar uma curva de aprendizado mais suave. No entanto, a dependência de infraestrutura compartilhada impõe limitações técnicas significativas. Os próprios autores do MicrobiomeAnalyst 2.0 reconhecem essa questão, explicitando que o tráfego de usuários e a largura de banda restringem o tamanho dos arquivos de entrada a 50 MB no servidor público.

Figura 7 – Interface de usuário inicial para análise exploratória



Fonte: A autora (2025)

Em contrapartida, a abordagem CLI do PGPTracker oferece autonomia computacional, eliminando tanto as latências de transferência de dados quanto as restrições de tamanho de arquivo inerentes ao processamento remoto. Para não comprometer a acessibilidade, a GUI opcional desenvolvida em Streamlit permite que usuários menos familiarizados com a linha de comando explorem os resultados de forma interativa.

Figura 8 – Exemplo de análise exploratória realizada pela GUI

Fonte: A autora (2025)

Além disso, embora ferramentas como o iVikodak ofereçam módulos para investigar contribuintes taxonômicos de vias específicas, essa análise opera sob demanda, exigindo a seleção manual de cada função. Esse processo torna-se impraticável para análises exploratórias em larga escala. O PGPTracker supera essa barreira ao gerar a matriz estratificada completa em lote, viabilizando a aplicação direta de análises multivariadas complexas, como Random Forest estratificado e seleção de características por Boruta, que seriam inviáveis com a extração manual via por via. Abaixo na Tabela 2, uma comparação entre as ferramentas citadas:

Tabela 2 – Comparação entre as ferramentas apresentadas

Característica	PGPTracker (Este trabalho)	PICRUSt2 (Douglas et al., 2020)	iVikodak (Nagpal et al., 2019)	MicrobiomeAnalyst 2.0 (Lu et al., 2023)	PGPg finder (Pellegrinetti et al., 2024)
Dado de Entrada	Amplicon 16S (ASV/OTU)	Amplicon 16S (ASV/OTU)	Amplicon 16S (ASV/OTU)	16S, Shotgun, Metabolômica	Shotgun / Genomas Montados
Infraestrutura	Local (CLI + GUI)	Local (CLI)	Web Server	Web Server	Local (CLI)
Foco Funcional	PGPTs (PLaBAse)	Generalista (KEGG/EC)	Generalista (KEGG/COG)	Saúde humana (KEGG/Sazonal)	PGPTs (PLaBAse)
Estratificação	Matriz completa otimizada	Matriz completa	Visualização pontual por taxon/PGPT	Visualização em Redes/Correlação	N/A (Focado em contagem de genes)
Performance	Alta e ilimitada	Média e ilimitada	Média e limitada	Alta e limitada (Upload máx. 50MB)	Baixa e ilimitada

Análise Estatística	Integrada (ML/Random Forest)	Externa (Requer R/Python à parte)	Integrada (Visual)	Integrada (Estatística Clássica)	Focado em Anotação

Fonte: O autor (2025), baseado em: Douglas et al. (2020), Nagpal et al. (2019), Lu et al. (2023), Pellegrinetti et al. (2024).

3.4 LIMITAÇÕES E PERSPECTIVAS FUTURAS

3.4.1 Limitações intrínsecas da inferência por Amplicon 16S

A abordagem de inferência funcional implementada no PGPTracker herda limitações fundamentais do método PICRUSt2, as quais devem ser explicitadas para uso apropriado da ferramenta. Douglas et al. (2020) destacam que qualquer análise baseada em Amplicon só consegue diferenciar táxons na medida em que diferem no gene marcador amplificado, resultando em perfis funcionais idênticos para linhagens cripticamente diferentes, mas com sequências de 16S rRNA altamente similares.

Logo, a inferência funcional baseada em marcadores filogenéticos atribui potencial metabólico através de reconstrução ancestral ou média de genomas relacionados. Este processo assume conservação filogenética de características funcionais, o que é uma premissa válida para genes *housekeeping*, mas menos robusta para características de promoção de crescimento vegetal.

Além disso, Traços PGP podem ser cepa-específicos, frequentemente adquiridos via transferência horizontal de genes através de plasmídeos ou elementos móveis genéticos, e portanto não necessariamente compartilhados por todo o gênero ou família. Consequentemente, resultados do PGPTracker devem ser interpretados como potencial funcional herdado, não como garantia de atividade fenotípica ou presença confirmada de genes específicos.

Ademais, a inferência detecta potencial funcional genômico, não atividade metabólica real. Gonçalves et al. (2024) utilizaram dados metatranscriptômicos para validar expressão gênica de PGPTs em *Acidobacteria*, demonstrando que nem todos os genes detectados genomicamente são expressos em condições ambientais específicas.

Afinal, são fatores como regulação gênica responsiva a sinais químicos do hospedeiro, disponibilidade de substratos e condições abióticas quem determinam

se um gene presente será efetivamente transcrito e traduzido em atividade funcional. Portanto, o PGPTracker identifica capacidade funcional potencial, cabendo ao pesquisador determinar quais funções preditas merecem validação experimental prioritária.

3.4.2 Posicionamento como ferramenta de triagem

Ferramentas baseadas em sequenciamento shotgun, como o PGPg_finder desenvolvido por Pellegrinetti et al. (2024), detectam a presença real de genes com maior precisão através de montagem e anotação de sequências metagenômicas. Esta abordagem oferece resolução funcional superior mas permanece financeiramente inacessível para triagens de larga escala. Como já abordado anteriormente, o PGPTracker ao operar com Amplicons 16S, é significativamente mais acessível mas menos preciso funcionalmente.

Dessa forma, o *pipeline* serve como triagem inicial que identifica amostras, ambientes ou grupos taxonômicos de interesse elevado, direcionando recursos de sequenciamento aprofundado para validação de alvos selecionados.

Esta estratégia híbrida maximiza a eficiência de recursos em projetos com restrições orçamentárias. Por exemplo, em estudos de monitoramento de microbioma do solo em múltiplas propriedades agrícolas, o PGPTracker pode processar centenas de amostras via 16S para identificar os 10 a 20% com maior potencial funcional para PGPTs de interesse, justificando investimento subsequente em sequenciamento *shotgun* apenas destas amostras prioritárias.

3.4.3 Validação experimental e aplicabilidade biotecnológica

A tradução de perfis funcionais preditos para aplicações biotecnológicas requer validação experimental em sistemas planta-solo. Glick et al. (2012) ressalta que nem todas as bactérias que contêm genes PGPT são eficazes promotoras de crescimento in vivo, devido a fatores como eficiência de colonização radicular, competência rizosférica e regulação gênica responsiva a exsudatos vegetais específicos do hospedeiro. Por exemplo, a presença de genes *nif* (nitrogenase) não garante fixação biológica de nitrogênio efetiva se as condições ambientais não forem

adequadas ou se a bactéria não conseguir estabelecer associação estável com raízes.

O PGPTracker serve como ferramenta de priorização para pesquisa experimental. Identificando táxons com maior densidade de PGPTs ou combinações funcionais sinérgicas (exemplo: fixação de nitrogênio + solubilização de fosfato + produção de sideróforos), pesquisadores podem direcionar esforços de cultivo microbiano e caracterização fenotípica para isolados com maior probabilidade de eficácia como inoculantes. Esta abordagem racional superaria triagens empíricas extensivas, reduzindo custos e acelerando desenvolvimento de bioinsumos.

Vale ressaltar que o PGPTracker complementa, mas não substitui, esta validação experimental, fornecendo fundamento quantitativo para seleção inicial de candidatos.

3.4.4 Expansões futuras

Niu et al. (2024) sugeriram que tecnologias multi-ômicas modernas como metagenômica, metatranscriptômica e metabolômica podem fornecer melhor compreensão das dinâmicas funcionais microbianas. A arquitetura modular do PGPTracker permite expansões futuras para incorporar múltiplas camadas de validação funcional.

Primeiro, uma evolução natural seria a integração de suporte para dados de metagenômica *shotgun*. Conforme demonstrado por Pellegrinetti et al. (2024) com o PGPg_finder, a detecção direta de genes funcionais oferece maior resolução do que a inferência. Uma futura versão do PGPTracker poderia incluir um módulo para processar tabelas de contagem de genes *shotgun*, aplicando a mesma lógica de filtragem pela PLaBAse e estratificação taxonômica já implementada, permitindo aos usuários validar seus resultados de 16S com dados de maior precisão quando disponíveis.

Segundo, a base de dados PLaBAse, embora robusta, é estática. Uma melhoria significativa seria implementar um mecanismo de atualização automática ou permitir que o usuário forneça dicionários personalizados de "Gene-para-Função". Isso garantiria que a ferramenta acompanhasse a rápida descoberta de novos mecanismos de promoção de crescimento vegetal, além de permitir a

adaptação da análise para nichos específicos não cobertos pela PLaBAse original, como funções de biocontrole contra patógenos regionais específicos.

Terceiro, a interface gráfica (GUI) atual opera de forma independente do *pipeline* de processamento principal, servindo apenas para visualização de resultados pré-calculados, logo poderia ser implementado botões e formulários para executar a *pipeline* completa a partir da GUI. Outra expansão viria a ser a integração completa, onde a GUI comandasse a execução do *pipeline* em segundo plano, o que aumentaria significativamente a acessibilidade da ferramenta para pesquisadores sem familiaridade com a linha de comando. Isso exigiria, no entanto, um refatoramento da arquitetura para gerenciar filas de processos e feedback de execução em tempo real.

Além disso, desenvolvimentos futuros podem explorar métodos alternativos de inferência funcional. Mongad et al. (2021) propõem abordagem que realiza classificação taxonômica refinada antes da inferência funcional, potencialmente aumentando precisão. Outra alternativa seria expandir a *pipeline* para prever enzimas (E.C) e inferir função a partir de suas vias metabólicas.

4 CONCLUSÃO

Este trabalho apresentou o desenvolvimento e validação do PGPTracker, uma ferramenta bioinformática composta da tradução de dados taxonômicos de Amplicon 16S rRNA em caracterização funcional relacionada à promoção de crescimento vegetal. A lacuna entre identificação taxonômica e capacidade funcional tem limitado historicamente a aplicabilidade prática de estudos de microbioma em contextos agrícolas. Por isso, com o objetivo de conectar estes dois domínios críticos de informação, o PGPTracker ofereceu uma solução metodologicamente robusta e computacionalmente acessível.

A validação técnica demonstrou viabilidade de processamento em hardware convencional. O *dataset* de 66 amostras do Earth Microbiome Project, contendo 12.257 ASVs, foi processado em aproximadamente 23 minutos, com a etapa crítica de estratificação funcional concluída em 3 minutos e consumo de RAM inferior a 10 GB. Esta eficiência, alcançada através da reimplementação do módulo de estratificação com a biblioteca Polars, representa redução superior a 98% no requisito de memória comparado à implementação original do PICRUSt2, que exigiria mais de 512 GB de RAM para o mesmo conjunto de dados.

A otimização computacional viabilizou a execução em computadores pessoais de especificações modestas, eliminando a dependência de infraestrutura de *cluster* de alto desempenho e democratizando o acesso à análise funcional estratificada, a principal contribuição científica desse trabalho. Além disso, a integração com a PLaBAse conferiu ao PGPTracker especificidade agronômica superior a ferramentas generalistas que utilizam bancos de dados amplos como KEGG ou COG.

Assim, a estratificação revelou dinâmicas ecológicas que seriam invisíveis em análises de abundância agregada, como o fenômeno de substituição taxonômica de nicho funcional observado na função de colonização e uso de substratos vegetais.

A comparação com ferramentas existentes posicionou o PGPTracker como solução complementar no ecossistema de análise funcional de microbiomas. Ferramentas baseadas em sequenciamento shotgun, como o PGPg finder, oferecem resolução funcional superior mas a custos computacionais e financeiros significativamente maiores. Já plataformas web como MicrobiomeAnalyst 2.0 e

iVikodak oferecem acessibilidade mas impõem restrições de infraestrutura compartilhada, incluindo limites de tamanho de arquivo que inviabilizam análises de larga escala. Logo, o PGPTracker, ao operar como ferramenta de linha de comando com interface gráfica opcional, combinou autonomia computacional com acessibilidade, permitindo processamento local de *datasets* massivos sem latências de transferência ou limitações de servidor.

As limitações identificadas são transparentes e comuns a abordagens de inferência funcional baseadas em Amplicons 16S rRNA. Isso significa que a ferramenta detecta potencial funcional herdado baseado em proximidade filogenética a genomas de referência, não atividade fenotípica real. Além disso, características específicas de cepa, frequentemente adquiridas via transferência horizontal de genes em plasmídeos, podem não ser capturadas se não correlacionadas com a filogenia do gene marcador.

O PGPTracker representa contribuição metodológica significativa para pesquisas de agricultura, biotecnologia e biologia. A ferramenta viabiliza triagem funcional em larga escala de comunidades microbianas do solo, fornecendo fundamentação quantitativa para seleção de inoculantes microbianos e direcionamento de recursos de sequenciamento aprofundado.

Ressaltando a capacidade de processar dezenas de amostras em minutos com hardware acessível transforma a análise funcional estratificada de um procedimento especializado em ferramenta rotineira de caracterização de microbiomas. Assim, a identificação automatizada de táxons portadores de funções promotoras de crescimento vegetal acelera a descoberta de candidatos biotecnológicos, reduzindo a dependência de triagens empíricas extensivas e possibilitando desenvolvimento mais eficiente de bioinsumos para sistemas agrícolas sustentáveis.

Assim, o PGPTracker alcançou seu objetivo de conectar taxonomia a função através da estratificação funcional automatizada e otimizada computacionalmente. Transformando dados de sequenciamento 16S rRNA em conhecimento prático sobre o potencial funcional dos microbiomas do solo. Essa ferramenta democratiza o acesso a análises funcionais estratificadas, impulsionando a compreensão e a manipulação racional das comunidades microbianas na agricultura.

Por fim, o PGPTracker não deve ser visto apenas como uma ferramenta de análise de dados, mas como um instrumento de apoio à decisão em projetos de

microbiologia agrícola. Em estudos de inoculantes, a *pipeline* permite priorizar táxons com combinações sinérgicas de PGPTs, direcionando esforços de isolamento e testes de estufa para candidatos com maior probabilidade de sucesso. Em avaliações de impacto de manejo, o PGPTracker pode quantificar ganhos ou perdas de funções promotoras de crescimento vegetal associados a diferentes práticas de uso do solo. Finalmente, na triagem de solos para cultivos específicos, a ferramenta possibilita identificar perfis funcionais mais adequados às exigências fisiológicas de cada cultura, auxiliando na seleção de áreas e na formulação de estratégias de manejo integradas.

REFERÊNCIAS

ADOMAKO, Michael Opoku; ROILOA, Sergio; YU, Fei-Hai. Potential roles of soil microorganisms in regulating the effect of soil nutrient heterogeneity on plant performance. Microorganisms, v. 10, n. 12, p. 2399, 2022.

AITCHISON, John. The statistical analysis of compositional data. Journal of the Royal Statistical Society: Series B (Methodological), v. 44, n. 2, p. 139-160, 1982.

AMIR, Amnon et al. Deblur rapidly resolves single-nucleotide community sequence patterns. MSystems, v. 2, n. 2, p. 10.1128/msystems. 00191-16, 2017.

BOKULICH, Nicholas A. et al. Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2's q2-feature-classifier plugin. Microbiome, v. 6, n. 1, p. 90, 2018.

BOLYEN, Evan et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature biotechnology, v. 37, n. 8, p. 852-857, 2019.

BREITKREUZ, Claudia et al. Can we estimate functionality of soil microbial communities from structure-derived predictions? A reality test in agricultural soils. Microbiology Spectrum, v. 9, n. 1, p. 10.1128/spectrum. 00278-21, 2021.

CALLAHAN, Benjamin J.; MCMURDIE, Paul J.; HOLMES, Susan P. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME journal, v. 11, n. 12, p. 2639-2643, 2017.

CHEN, Huaihai et al. Functional redundancy in soil microbial community based on metagenomics across the globe. Frontiers in microbiology, v. 13, p. 878978, 2022.

DOUGLAS, Gavin M. et al. PICRUSt2 for prediction of metagenome functions. Nature biotechnology, v. 38, n. 6, p. 685-688, 2020.

ESCOBAR-ZEPEDA, Alejandra; VERA-PONCE DE LEÓN, Arturo; SANCHEZ-FLORES, Alejandro. The road to metagenomics: from microbiology to DNA sequencing technologies and bioinformatics. Frontiers in genetics, v. 6, p. 348, 2015.

GLICK, Bernard R. Plant growth-promoting bacteria: mechanisms and applications. Scientifica, v. 2012, n. 1, p. 963401, 2012.

GONÇALVES, Osiel S. et al. Insights into plant interactions and the biogeochemical role of the globally widespread Acidobacteriota phylum. Soil Biology and Biochemistry, v. 192, p. 109369, 2024.

HARTMANN, Martin et al. Distinct soil microbial diversity under long-term organic and conventional farming. The ISME journal, v. 9, n. 5, p. 1177-1194, 2015.

JOOS, Lisa et al. Daring to be differential: metabarcoding analysis of soil and plant-related microbial communities using amplicon sequence variants and operational taxonomical units. BMC genomics, v. 21, n. 1, p. 733, 2020.

JOUSSET, Alexandre et al. Where less may be more: how the rare biosphere pulls ecosystems strings. The ISME journal, v. 11, n. 4, p. 853-862, 2017.

KUO, Jimmy; LIU, Daniel; LIN, Chorng-Horng. Functional prediction of microbial communities in sediment microbial fuel cells. Bioengineering, v. 10, n. 2, p. 199, 2023.

KURSA, Miron B.; JANKOWSKI, Aleksander; RUDNICKI, Witold R. Boruta–a system for feature selection. Fundamenta informaticae, v. 101, n. 4, p. 271-285, 2010.

LEINONEN, Rasko et al. The European nucleotide archive. Nucleic acids research, v. 39, n. suppl_1, p. D28-D31, 2010.

LU, Yao et al. MicrobiomeAnalyst 2.0: comprehensive statistical, functional and integrative analysis of microbiome data. Nucleic Acids Research, v. 51, n. W1, p. W310-W318, 2023.

MACRAE, Andrew. The use of 16S rDNA methods in soil microbial ecology. Brazilian Journal of microbiology, v. 31, p. 77-82, 2000.

MARTÍNEZ-NÚÑEZ, Mario Alberto; OROZCO-RAMÍREZ, Quetzalcoátl. Characterizing bacterial communities in agroecosystems of the UNESCO global geopark Mixteca Alta, Oaxaca. Agriculture, v. 14, n. 12, p. 2180, 2024.

MATCHADO, Monica Steffi et al. On the limits of 16S rRNA gene-based metagenome prediction and functional profiling. Microbial Genomics, v. 10, n. 2, p. 001203, 2024.

MCDONALD, Daniel et al. Greengenes2 unifies microbial data in a single reference tree. Nature biotechnology, v. 42, n. 5, p. 715-718, 2024.

MONGAD, Dattatray S. et al. MicFunPred: A conserved approach to predict functional profiles from 16S rRNA gene sequence data. Genomics, v. 113, n. 6, p. 3635-3643, 2021.

NAGPAL, Sunil et al. iVikodak—A platform and standard workflow for inferring, analyzing, comparing, and visualizing the functional potential of microbial communities. Frontiers in Microbiology, v. 9, p. 3336, 2019.

NITYAGOVSKY, Nikolay N. et al. Endophytic Bacterial and Fungal Communities of Spruce Picea jezoensis in the Russian Far East. Plants, v. 14, n. 16, p. 2534, 2025.

NIU, Hongjin et al. Deciphering the differences of bacterial communities between high-and low-productive wheat fields using high-throughput sequencing. Frontiers in Microbiology, v. 15, p. 1391428, 2024.

NOBU, Masaru K. et al. Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. The ISME journal, v. 9, n. 8, p. 1710-1722, 2015.

OSTOS, Iván; FLÓREZ-PARDO, Luz Marina; CAMARGO, Carolina. A metagenomic approach to. 2024.

PARKS, Donovan H. et al. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature biotechnology, v. 36, n. 10, p. 996-1004, 2018.

PATZ, Sascha et al. mgPGPT: Metagenomic analysis of plant growth-promoting traits. BioRxiv, p. 2024.02. 17.580828, 2024.

PATZ, Sascha et al. PLaBAse: A comprehensive web resource for analyzing the plant growth-promoting potential of plant-associated bacteria. BioRxiv, p. 2021.12. 13.472471, 2021.

PELLEGRINETTI, Thierry Alexandre et al. PGPg_finder: a comprehensive and user-friendly pipeline for identifying plant growth-promoting genes in genomic and metagenomic data. Rhizosphere, v. 30, p. 100905, 2024.

SHAYANTHAN, Ambihai; ORDOÑEZ, Patricia Ann C.; ORESNIK, Ivan John. The role of synthetic microbial communities (SynCom) in sustainable agriculture. Frontiers in Agronomy, v. 4, p. 896307, 2022.

THOMPSON, L.R., Sanders, J.G., McDonald, D. et al. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. Nature, 551, 457-463.

VINK, Ritchie et al. Polars: Blazingly fast DataFrames in Rust and Python. Versão 1.35.2 [S.l.]: Zenodo, 2025. DOI: https://doi.org/10.5281/zenodo.7697217. Disponível em: https://github.com/pola-rs/polars. Acesso em: 26 nov. 2025.

ZHANG, Yiqian et al. Review and revamp of compositional data transformation: A new framework combining proportion conversion and contrast transformation. Computational and Structural Biotechnology Journal, v. 23, p. 4088-4107, 2024.
ZHANG, Zhengqing et al. Rare species-driven diversity–ecosystem multifunctionality relationships are promoted by stochastic community assembly. MBio, v. 13, n. 3, p. e00449-22, 2022.

APÊNDICE A – REPOSITÓRIOS DIGITAIS E ACESSO AO CÓDIGO FONTE

O código-fonte completo, a documentação de instalação, os scripts de automação e o histórico de versões do PGPTracker está hospedado publicamente na plataforma GitHub para garantir a transparência e a reprodutibilidade do estudo.

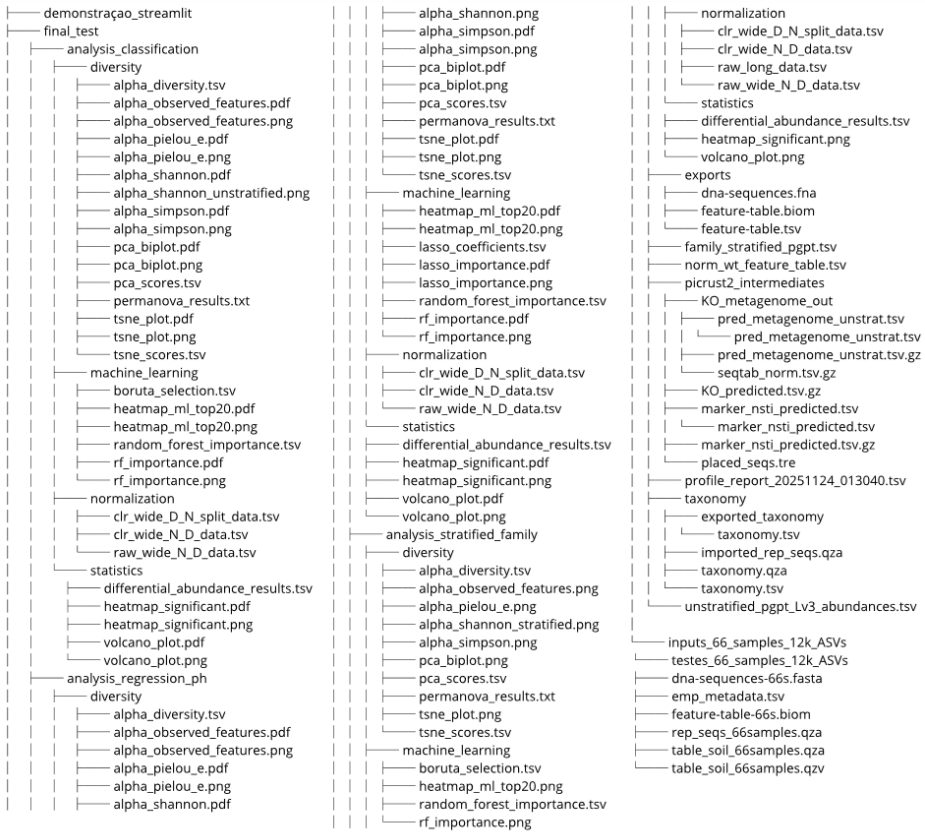
- Repositório oficial (GitHub): https://github.com/kiuone/PGPTracker.

Adicionalmente, o conjunto de dados completo utilizado para a validação da ferramenta contendo os arquivos de entrada e todos os resultados processados apresentados neste trabalho, encontra-se disponível em nuvem no *link* abaixo:

- Repositório de dados e resultados: https://drive.google.com/drive/folders/1WjeMhvXvHmz8Vc3jQtuyfnHEMVQSjCVz?usp=sharing.

Abaixo apresenta-se a árvore de diretórios completa gerada durante a execução do PGPTracker para o conjunto de dados de validação. Esta estrutura ilustra a organização dos dados brutos, resultados intermediários e saídas finais estatísticas e gráficas disponíveis no repositório de dados.

Figura 9 – Estrutura de dados do repositório com os exemplos



Fonte: A autora (2025)