UNIVERSIDADE DE SÃO PAULO

Predição de fenótipos bacterianos a partir de genomas por meio de aprendizado de máquina

BRUNO KOSHIN VÁZQUEZ IHA

São Paulo, Brasil

2025

Bruno Koshin Vázquez Iha

Predição de fenótipos bacterianos a partir de genomas por meio de aprendizado de máquina

Tese de Doutorado apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática da Universidade de São Paulo para obtenção do título de Doutor em Ciências.

Área de Concentração: Bioinformática

Orientador: Prof. Dr. João Carlos Setubal

São Paulo, Brasil

2025

Dedico esta obra à minha família. Mãe, Pai e Abuela.
E a Oba, Oji e Abuelo (*in memoriam*).

Agradecimentos

Agradeço ao meu orientador, **Prof. Dr. João Carlos Setubal**, por ter me aceito como seu aluno, por oferecer toda sua visão de pesquisa científica e paciência. Sem o senhor nada disto teria sido possível.

Ao **Herr Prof. Dr. Peter Florian Stadler** por ter me aceito em seu grupo em Leipzig, resultando em todo carinho, amizade além de aprendizado. Toda a experiência com ele e seu grupo de alunos e colegas ampliaram meu jeito de ver o mundo.

Aos **Profs. Drs. Regina Lúcia Baldini, Alexandre Dias Porto Chiavegatto Filho, André Carlos Ponce de Leon Ferreira de Carvalho** e **André Fujita** pelos conselhos valiosos.

À **Fundação Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES)** pela a concessão de bolsa de doutorado que financiou este trabalho.

Ao programa **Probral CAPES-DAAD** pela concessão de bolsa para estágio de 6 meses na Universidade de Leipzig, Alemanha.

À **FAPESP** por financiar a infraestrutura computacional do Laboratório de Bioinformática do IQ-USP, utilizada neste trabalho.

Agradecimento especial à minha parceira **Beatriz Toledo Akiti,** por ser quem é e estar ao meu lado.

Aos **meus amigos** por compreenderem os meus momentos de estresse e mesmo assim me darem sua amizade e apoio.

To **all my friends in Leipzig**, I miss you all.

"*Nosso modelo deve ser simples o suficiente para ser manejável, mas complexo o suficiente para capturar a ideia central do problema. Esta é a arte do cientista.*"

- Magnusson & Mourão (2005),

Estatística [sem] matemática: a ligação entre as questões e a análise.

Abstract

IHA, B. K. V. **PheMaLe-bacteria -Inferindo fenótipos de bactérias a partir de dados genômicos utilizando aprendizado de máquina**. 104 pages. Ph.D Thesis. Bioinformatics Graduate Program, Universidade de São Paulo, São Paulo. 2025.

The characterization of bacteria and the identification of genes related to the emergence of different bacterial characteristics are fundamental for understanding the complex microbial diversity and functionality. However, there is a significant gap in the knowledge of genetic patterns associated with genetically complex bacterial traits. The usual methodology for identifying and analyzing these patterns is time-consuming, labor-intensive, and requires highly specialized knowledge due to the large volume of data, broad genetic diversity, and lack of functional information for many genes. As an alternative to manual work in the screening stage for relevant genes, ad hoc models trained with machine learning techniques can be employed. To test the machine learning algorithms capability to learn genetic patterns correlated with bacterial traits, a tool called PheMaLe-bacteria was developed, derived from the terms "PHEnotype," "MAchine," "LEarning," and "bacteria." The tool employs a total of 21 machine learning algorithms for modeling and it was tested on three bacterial traits considered genetically complex: 1) the bacterium's ability to fix nitrogen, 2) the bacterium's survival pH range, 3) the bacterium's ability to form endospores. The results produced models with evaluation metrics superior to other studies in the same field, demonstrating the tool's ability to identify multiple genes correlated with these traits. With PheMaLe-bacteria, we aim to accelerate the characterization of new bacteria and assist in discovering genetic patterns underlying their traits. This tool, along with the data used and trained models, is available at https://github.com/LaboratorioBioinformatica/PHEMALE or upon request by contacting the author at <bkviha@gmail.com>.

Keywords: bacterial phenotypes, bacterial genomics, machine learning.

Resumo

IHA, B. K. V. **PheMaLe-bacteria -Inferindo fenótipos de bactérias a partir de dados genômicos utilizando aprendizado de máquina**. 104 páginas. Tese de Doutorado. Programa de Pós-Graduação Interunidades em Bioinformática, Universidade de São Paulo, São Paulo. 2025.

A caracterização de bactérias e a identificação dos genes relacionados à emergência das diferentes características bacterianas são fundamentais para a compreensão da complexa diversidade e funcionalidade microbiana. No entanto, há uma grande lacuna no conhecimento dos padrões genéticos associados a características bacterianas geneticamente complexas. A metodologia usual para identificar e analisar esses padrões demanda muito tempo, trabalho e conhecimento altamente especializado devido à grande quantidade de dados, ampla diversidade genética e ausência de informações funcionais de diversos genes. É possível, como alternativa ao trabalho manual da etapa de triagem de genes relevantes, utilizar modelos ad hoc treinados utilizando aprendizado de máquina. Para testar a capacidade de algoritmos de aprendizado de máquina em aprender padrões genéticos correlacionados a características bacterianas foi desenvolvida uma ferramenta denominada PheMaLe-bacteria, advinda dos termos "PHEnotype" "MAchine" "LEarning" "bacteria". A ferramenta emprega um total de 21 algoritmos de aprendizado de máquina para modelagem e foi testada em 3 características bacterianas consideradas complexas geneticamente: 1) capacidade da bactéria de fixarnitrogênio, 2) faixa de pH de sobrevivência da bactéria, 3) capacidade de endoesporulação. Os resultados foram modelos com métricas de avaliação superioras a outros trabalhos no mesmo tópico e com capacidade de identificação múltiplos genes correlacionados a essas características. Com a ferramenta PheMaLe-bacteria espera-se acelerar a caracterização de novas bactérias e auxiliar no descobrimento de padrões genéticos que geram suas características. Esta ferramenta, os dados utilizados e os modelos treinados se encontram disponíveis no website https://github.com/LaboratorioBioinformatica/PHEMALE ou mediante contato com o autor <bkviha@gmail.com>.

Palavras-chave: fenótipos bacterianos, genômica bacteriana, aprendizado de máquina

Lista de figuras

Figura 1 Fluxograma do processo de representação de relevância e quantificação de grupos ortólogos 14

Figura 2 Distribuição de ordens no conjunto de dados para fixação de nitrogênio 21

Figura 3 Quantidades de genomas completos no conjunto de dados com informação sobre capacidade de fixação de nitrogênio 22

Figura 4 EMD dos genomas com distinção em capacidade de fixação de nitrogênio 22

Figura 5 Métricas F1, AUC-ROC e AUC-PR dos modelos de predição de fixação de nitrogênio sob o conjunto de dados teste 24

Figura 6 Precisão e Recall dos dados de teste para capacidade de fixação de nitrogênio 25

Figura 7 Matrizes de confusão para predições de capacidade de fixação de nitrogênio 25

Figura 8 AUC-PR e AUC-ROC do modelo Regressão Logística com representação binária de COGs para classificação de capacidade de fixação de nitrogênio 29

Figura 9 Curva de precisão x recall e curva ROC do modelo Regressão Logística com representação quantitativa de COGs para classificação de capacidade de fixação de nitrogênio 29

Figura 10 Precisão e Recall e respectivas matrizes de confusão para o conjunto de teste corrigido 30

Figura 11 Precisão, Recall e F1 de Phenotrex 0.6.0 com o conjunto teste de fixação de nitrogênio e respectiva matriz de confusão 31

Figura 12 20 grupos ortólogos mais influentes para predição de fixação de nitrogênio no modelo treinado com Regressão Logística representação binária de COGs 32

Figura 13 20 grupos ortólogos mais influentes para predição de fixação de nitrogênio no modelo Regressão Logística com representação quantitativa de COGs 34

Figura 14 Distribuição de ordens no conjunto de dados para faixas de pH 41

Figura 15 Quantidade de espécies no conjunto de dados de acordo com seus limites inferior e superior de pH 42

Figura 16 EMD dos genomas com informação sobre faixas de pH associadas 43

Figura 17 Métrica 100×R^2 dos modelos de predição de limites inferior e superior de pH sob o conjunto de dados teste 45

Figura 18 Comparação dos valores reais de limites superior e inferior de pH com valores preditos pelo modelo XGBoost com representação quantitativa de COGs 46

Figura 19 Comparação dos valores reais de limites superior e inferior de pH com os valores preditos pelo modelo Random Forest utilizando representação binária de COGs e eNOGs 47

Figura 20 Quantidades de casos por distância do valor entre as predições e valores reais dos limites de pH 48

Figura 21 20 grupos ortólogos mais influentes para predição do limite superior de pH no modelo XGBoost com representação quantitativa de COGs 49

Figura 22 10 grupos ortólogos mais influentes para predição do limite inferior de pH no modelo Random Forest com representação binária de COGs e eNOGs 52

Figura 23 Distribuição de ordens no conjunto de dados para capacidade de endoesporulação 57

Figura 24 Quantidades de genomas completos no conjunto de dados com informação para capacidade de endoesporulação 58

Figura 25 EMD dos genomas com distinção sobre capacidade de endoesporulação 58

Figura 26 Métricas F1, AUC-ROC e AUC-PR dos modelos de predição de capacidade de endoesporulação sob o conjunto de dados teste 60

Figura 27 Precisão, Recall e F1 do modelo teinado com LightGBM e matriz de confusão com o conjunto teste de endoesporulação 61

Figura 28 Curva de precisão x recall e curva ROC do modelo LightGBM com representação binária de COGs e eNOGs para classificação de capacidade de fixação de nitrogênio 62

Figura 29 20 grupos ortólogos mais influentes para predição de capacidade de endoesporulação no modelo treinado com LightGBM utilizando representação binária de COGs e eNOGs para predição de capacidade de endoesporulação 63

Figura B.1 (Apêndice B) 20 grupos ortólogos mais influentes para predição de fixação de nitrogênio no modelo SVM treinado com representação binária de COGs e eNOGs 90

Figura B.2 (Apêndice B) 20 grupos ortólogos mais influentes para predição de fixação de nitrogênio no modelo SVM treinado com representação quantitativa de COGs e eNOGs 90

Lista de tabelas

Tabela 1 Espécies preditas incorretamente em relação a capacidade de fixação de nitrogênio 26

Tabela 2 Espécies preditas incorretamente em relação a capacidade de endoesporulação 61

Tabela A (Apêndice A) Ordens com no máximo 10 representantes no conjunto de dados utilizado para treino e avaliação da PheMaLe-bacteria para capacidade de fixação de nitrogênio 88

Tabela C (Apêndice C) Ordens com no máximo 10 representantes no conjunto de dados utilizado para treino e avaliação da PheMaLe-bacteria para faixas de pH toleradas 91

Tabela D (Apêndice D) Ordens com no máximo 10 representantes no conjunto de dados utilizado para treino e avaliação da PheMaLe-bacteria para capacidade de endoesporulação 94

Sumário

Capítulo 1 Introdução 1

1.1 Importância da identificação dos genes responsáveis por características bacterianas 2

1.2 Motivação 5

1.3 Revisão bibliográfica 6

Capítulo 2 Desenvolvimento da PheMaLe-bacteria 9

2.1 Informações dos fenótipos e genomas bacterianos utilizados 9

2.2 Representação simbólica dos genomas 10

2.3 Algoritmos de aprendizado de máquina selecionados 15

2.4 Treino e avaliação dos modelos treinados 16

2.5 Informações técnicas da PheMaLe-bacteria 17

Capítulo 3 Capacidade de fixação de nitrogênio 19

3.1 Informações de espécies bacterianas e sua capacidade de fixação de nitrogênio 21

3.2 Resultados da PheMaLe-bacteria 23

3.3 Análise de influência dos grupos ortólogos 31

3.4 Conclusão 37

Capítulo 4 Tolerância de bactérias a faixas de pH 39

4.1 Espécies bacterianas e as faixas de pH toleradas 41

4.2 Resultados da PheMaLe-bacteria 43

4.3 Análise de influência dos grupos ortólogos 48

4.4 Conclusão 53

Capítulo 5 **Endoesporulação** 55

5.1 Informações de espécies bacterianas e sua capacidade de endoesporulação 57

5.2 Resultados da PheMaLe-bacteria 59

5.3 Análise de influência dos grupos ortólogos 62

5.4 Conclusão 64

Capítulo 6 **Conclusão** 66

Referências 68

Glossário e abreviações 84

Apêndice A 87

Apêndice B 89

Apêndice C 90

Apêndice D 93

1. Introdução

As bactérias estão presentes em quase todos os aspectos da vida no planeta, podendo ser encontradas até mesmo nos ambientes mais extremos. Elas influenciam desde os ecossistemas naturais (desempenhando papéis essenciais para a ciclagem de nutrientes, como carbono e nitrogênio) até a saúde humana e de outros animais, apresentando também diversas aplicações biotecnológicas para a indústria (Nielsen *et al.*, 2022). Devido a essa grande relevância, entender as suas características funcionais é crucial para uma melhor compreensão tanto das bactérias quanto dos próprios ecossistemas em que estão inseridas, além de gerar oportunidades de avanços tecnológicos, principalmente nas áreas de saúde, indústrias e agricultura.

As características bacterianas são bem diversas, em função da influência de variados atributos, como a própria diversidade taxonômica e de ambientes ocupados pelo grupo. Alguns exemplos: morfologia (forma da célula: cocos, bacilos, espirilos), Gram-positiva ou negativa, presença de cápsula, mobilidade (presença de flagelos ou outro mecanismo de locomoção), metabolismo (aeróbia, anaeróbia, facultativa, quimioheterotrófico, fotossintético), capacidade de endoesporulação, produção de toxinas, capacidade de formação de biofilme, resistência a antibióticos, produção de metabólitos, capacidade de sobrevivência em ambientes extremos (acidófila, alcalófilas, termofílica, psicrofílica, mesofílica).

O modo mais confiável e mais utilizado atualmente para fenotipagem de bactérias é por meio de testes laboratoriais do cultivo isolado da cepa de interesse em diferentes condições físico-químicas. No entanto, esse processo constantemente se depara com obstáculos, incluindo a necessidade de múltiplos testes até se encontrar as condições de cultivo adequadas, equipamentos e reagentes financeiramente custosos, e constante monitoria humana. Ademais, nem sempre tal processo é factível, dado que diversos organismos exigem condições de cultivo complexas, custosas ou não completamente elucidadas. Por exemplo, muitas são dependentes de outras espécies de seres vivos (inseridas em relações de comensalismo, mutualismo ou parasitismo) ou exigem ambientes contrastantes ao ambiente laboratorial, como as extremófilas.

Outro fator relevante é que algumas bactérias podem apresentar plasticidade fenotípica, ou seja, certas características emergem apenas em determinadas condições ambientais (Tadrowski *et al.*, 2018; Smits *et al.*, 2006), portanto alguns fenótipos latentes podem passar despercebidos durante análises laboratoriais.

1.1. Importância da identificação dos genes responsáveis por características bacterianas

O mapeamento de funcionalidades genéticas e suas relações com fenótipos bacterianos abre diversas possibilidades práticas como a caracterização de novas bactérias baseada nos genes presentes, reconhecimento de resistência a antibióticos, terapias antimicrobianas mais específicas, identificação de bactérias que causam desregulações em ecossistemas e na saúde, seleção de bactérias com as qualidades desejadas no âmbito de indústria bioquímica ou conferir essas propriedades desejadas às bactérias por meio da manipulação genética.

Esse mapeamento dos genes relacionados a determinadas características em bactérias em geral é realizado por um processo composto por duas etapas: 1) a identificação dos genes candidatos 2) a validação experimental da relação entre a característica de interesse e os genes candidatos.

1) A identificação dos genes candidatos pode ser realizada de diversas formas:

    a) Genômica comparativa de bactérias com e sem a característica de interesse. Dessa forma são identificados os genes únicos ou conservados em bactérias que compartilham esse fenótipo. Isso pode ser realizado analisando os genes ortólogos presentes nos genomas, utilizando ferramentas como BLAST (Camacho *et al.*, 2009), HMMER (Eddy, 2020) e EggNOG (Huerta-Cepas *et al.*, 2018). Também há ferramentas, como o panX (Ding *et al.*, 2017), que identificam a presença e ausência de genes em múltiplos genomas. Outra técnica recentemente utilizada é o Genome-Wide Association Studies (GWAS) para analisar variações genéticas em populações microbianas com o intuito de identificar correlações com fenótipos (Read & Massey, 2014; Saber & Shapiro, 2020).

b) Análise transcriptômica, ou seja, a análise da expressão gênica de diferentes expressões fenotípicas para identificar os genes diferencialmente expressos. Isso pode ser realizado principalmente por via de RNA-Seq ou microarrays, os quais medem a quantidade de RNA mensageiro presentes (Mantione *et al.*, 2014).

c) Proteômica e Metabolômica, que identifica as proteínas e metabólitos presentes em diferentes condições fenotípicas e posteriormente possibilita inferir os genes relacionados a essas proteínas ou metabólitos. Isso pode ser realizado por via de espectrometria de massa, por microarrays (Sutandy *et al.*, 2013) ou cromatografia de afinidade (Chamrád *et al.*, 2024).

2) A validação experimental dos genes candidatos em relação a característica de interesse pode ser realizada de quatro formas:

a) Knockout, ou seja a deleção de gene candidato, para avaliação do impacto da sua ausência.

b) Mutagênese sítio-dirigida no gene candidato para verificar a ocorrência de alterações fenotípicas.

c) Mutagênese aleatória empregada por agentes químicos ou radiação para gerar mutações em genes aleatórios nas bactérias, na esperança de alguma das bactérias afetadas apresentar ou perder o fenótipo em questão.

d) Utilização de genes marcadores para monitorar a atividade de genes específicos. Pode ser realizado pela fusão de um marcador com os promotores do gene candidato, permitindo o monitoramento de quando o promotor é ativado.

e) Adição do gene candidato em um organismo modelo para verificar se o fenótipo passa a ser expresso.

Esse processo de mapeamento é eficaz para características geneticamente simples ou mendelianas, que são determinadas por um ou poucos genes, permitindo a identificação direta de sua relação com o fenótipo. No entanto, algumas

características dependem de múltiplos genes e envolvem interações gênicas complexas. Diferentemente das características simples, onde cada gene contribui de forma linear e independente para o fenótipo (ou seja, a soma dos efeitos individuais dos genes é suficiente para explicar a variação observada), as características geneticamente complexas frequentemente apresentam interações gênicas não aditivas (Lin *et al.*, 2022). Isso significa que a contribuição de um gene para o fenótipo pode depender da presença ou da atividade de outros genes, onde um gene pode potencializar, suprimir ou modificar a influência de outro. Essas interações tornam a relação entre genótipo e fenótipo mais difícil de prever, pois os efeitos combinados dos genes não podem ser simplesmente somados.

Diante dessa complexidade, abordagens integrativas são fundamentais para capturar as interações entre múltiplos genes e fatores regulatórios. Para organismos cujas funções gênicas e metabólicas ainda são pouco conhecidas, as redes de coexpressão gênica (Montenegro, 2022) são uma ferramenta valiosa, permitindo inferir interações funcionais com base em padrões de expressão semelhantes. Com organismos para os quais há mais conhecimento prévio sobre as funções dos genes presentes, podem-se utilizar redes metabólicas (Dräger & Planatscher, 2013), que permitem visualizar e identificar interações entre produtos de genes e metabólitos, fornecendo insights mais direcionados sobre a regulação e o metabolismo celular. Entretanto, a eficácia desses métodos depende fortemente da qualidade e da abrangência dos bancos de dados disponíveis para a anotação funcional de genes e metabólitos.

Quando estes métodos falham ou apresentam resultados incompletos, o aprendizado de máquina se mostra uma opção poderosa. Devido a sua flexibilidade em identificar múltiplos padrões de relações não aditivas sem a necessidade de informações prévias de como são essas relações, os métodos de aprendizado de máquina apresentam o potencial de revelar interações gênicas e metabólicas que passariam despercebidas com abordagens convencionais.

No entanto, a principal limitação da modelagem computacional de sistemas biológicos reside na incompletude e na qualidade dos dados disponíveis. Felizmente essa limitação tende a ser progressivamente superada com o crescimento contínuo dos bancos de dados biológicos e o aprimoramento das tecnologias de obtenção de dados, dessa forma aumentando a quantidade de informação e permitindo o avanço na modelagem e análise computacional. O avanço e a redução dos custos das

tecnologias de sequenciamento de DNA, aliado ao desenvolvimento de abordagens em metagenômica e reconstrução de genomas (Tully *et al.*, 2018) facilitou enormemente a obtenção de genomas com alta porcentagem de completude. Esses progressos impulsionaram a era do Big Data genômico (Stephens *et al.*, 2015; Land *et al.*, 2015; Zhao *et al.*, 2020). Além disso, agora é possível sequenciar genomas de microrganismos raros ou de baixa abundância, incluindo aqueles considerados incultiváveis em laboratório (Albertsen, 2013; Furtak *et al.*, 2020; Sogin *et al.*, 2006).

Na data de 17 de dezembro de 2024, o NCBI possuía em seu banco de dados público 39.997 sequências de genomas completos ou acima de 98% de completude de bactérias (*Genome List - Genome - NCBI*, s.d.), 4.877 grupos funcionais curados cujas funcionalidades são caracterizadas (Galperin *et al.*, 2021), e no banco de dados EggNOG 5.0 estão catalogados 206.782 grupos funcionais hipotéticos (Huerta-Cepas *et al.*, 2018) para os quais não há muitas informações definitivas sobre sua funcionalidade.

1.2. Motivação

O crescimento exponencial dos dados microbianos, tanto em volume (quantidade de dados) quanto em abrangência (número de genes identificados), tornou a análise manual inviável devido à sua complexidade e demanda intensiva de recursos. No entanto, o caráter de Big Data desses dados viabiliza a aplicação do aprendizado de máquina para sua modelagem e análise, metodologia essa especialmente útil no processamento de grandes volumes de informações com conhecimento prévio limitado ou inexistente. Dessa forma o aprendizado de máquina pode ser utilizado como metodologia para acelerar uma primeira etapa de triagem dos genes e padrões genéticos correlacionados com as características a serem estudadas.

O objetivo central deste trabalho é avaliar a capacidade de múltiplos algoritmos de aprendizado de máquina em identificar e aprender os padrões genéticos associados a fenótipos bacterianos, e se isso permitiria realizar predições com alta precisão e sensibilidade de características consideradas geneticamente complexas em bactérias somente levando em consideração informações genéticas presentes nas sequências genômicas. Para isso foi criada a ferramenta denominada

PheMaLe-bacteria (nome advindo dos termos Phenotype Machine Learning Bacteria) que realiza o treino padronizado de modelos de predição de fenótipos bacterianos a partir de seus padrões genéticos e engloba 21 algoritmos de aprendizado de máquina diferentes. Para a avaliação da ferramenta foram selecionados os fenótipos de capacidade para fixação de nitrogênio, faixa de pH tolerado e capacidade de endoesporulação, considerados fenótipos geneticamente complexos por emergirem devido a interações diretas e indiretas entre múltiplos genes.

Deve-se ressaltar no entanto que o uso de aprendizado de máquina não substitui a análise e conclusão final de um especialista na área, mas apenas propõe-se a servir de método que auxilie uma triagem dos genes a serem analisados mais profundamente. E deve se manter em mente que este trabalho se trata da modelagem baseado em uma abordagem de aprendizado de máquina, e como toda modelagem se trata de uma simplificação do mundo real, sendo possível predições erradas e falhas derivadas de uma incompletude nos parâmetros selecionados ou por limitações dos modelos escolhidos. Ainda assim estes resultados errados podem trazer informações que indiquem onde os modelos estão falhando e por quais motivos, sendo potencialmente útil para trabalhos futuros que abordem esta tema pela mesma abordagem.

1.3. Revisão bibliográfica

Os principais trabalhos científicos realizados nos últimos 10 anos relacionados ao tópico de aplicação de aprendizado de máquina a fenótipo-genótipos bacterianos generalizados são:

1) Phenotrex (Feldbauer *et al.*, 2015): trabalho que desenvolveu ferramenta de caráter generalista, treinada originalmente para predizer os fenótipos aeróbio, anaeróbio, facultativo, Gram-negativo, halofilia, termófilo, motilidade, fotossintético, psicrofílico e formação de endósporo. É uma reimplementação e extensão da ferramenta PICA (MacDonald & Beiko, 2010), que realizava classificação binária com o algoritmo Support Vector Machine (SVM), utilizando kernel linear. Os parâmetros analisados pela ferramenta para chegar a conclusão dos fenótipos são os grupos ortólogos presentes ou ausentes no genoma bacteriano

e os dados utilizados para seu treino foram obtidos no Joint Genome Institute (JGI) Integrated Microbial Genomes (IMG) (Markowitz *et al.*, 2011) e complementado com dados do NCBI. Os dados usados para essa segunda versão foram obtidos no JGI-IMG e complementados com literatura científica. A ferramenta continuou em desenvolvimento, e foi posteriormente renomeada Phenotrex. A limitação do PICA de devolver apenas valores binários foi contornada no Phenotrex realizando *Platt scaling* para devolver valores de confiança das predições. Em 10 de setembro de 2021 o Phenotrex teve a sua última atualização de acordo com seu repositórios online (*Phenotrex 0.6.0 Documentation*, s.d.). Nessa versão foi adicionado o modelo de predição baseado em XGBoost, alternativamente ao modelo prévio baseado em SVM, além da funcionalidade de informar os grupos ortólogos mais influentes nas predições utilizando o método SHapley Additive exPlanations (Lundberg & Lee, 2017).

2) Traitar (Weimann *et al.*, 2016): realiza predição binária de 67 características fenotípicas utilizando o algoritmo SVM com kernel linear. A ferramenta foi treinada com informações de fenótipos de 234 espécies microbianas obtidos no Global Infectious Disease and Epidemiology Online Network (Berger, 2005). Durante o treino foram simulados eventos probabilísticos de Transferência Horizontal de Genes para obter maior diversidade dos dados. Seus autores verificaram que ela realiza previsão com confiabilidade em 572 espécies de 8 filos. A ferramenta foi capaz de aprender relações entre algumas famílias de proteínas e fenótipos comprovados na literatura científica. Traitar teve sua última atualização em 25 de julho de 2016.

3) DiASPora (abreviação para Digital Approaches for the Synthesis of Poorly Accessible Biodiversity Information): um projeto que inclui uma ferramenta baseada em SVM para predição de 7 fenótipos (acidofilia, anaeróbia, motilidade, Gram-positivo, halofilia, capacidade de endoesporulação, termofilia) utilizando famílias de domínios de proteínas do banco de dados Pfam (Mistry *et al.*, 2020) como parâmetros analisados. Ela foi treinada utilizando dados fenotípicos obtidos no banco de dados BacDive (Reimer *et al.*, 2021) em fevereiro de 2023. A ferramenta não foi disponibilizada para o público e as únicas informações disponíveis sobre a ferramenta estão no próprio site (*DiASPora Project*, s.d.).

O método SVM, algoritmo de aprendizado de máquina utilizado em todas as 3 ferramentas, é uma boa escolha para conjuntos de dados pequenos e com grande quantidade de parâmetros devido a sua resiliência a sobreajuste. No entanto seu aprendizado é limitado a dados com separação de classes compatíveis com o kernel selecionado. Nos casos de Traitar e Phenotrex foi escolhido o kernel linear, provavelmente por ser o único kernel com o qual é possível analisar diretamente os parâmetros considerados importantes pelos modelos treinados. No entanto, essa escolha limita a flexibilidade de aprendizado dos modelos a apenas fenótipos separáveis linearmente no espaço dos parâmetros. Feldbauer *et al.* (2015) notaram limitações na escalabilidade do método SVM com os dados de treino a partir dos resultados de dois testes que realizaram. Um primeiro teste verificou que o tempo de treino da SVM com os 47.615 parâmetros originais da primeira versão utilizando o banco de dados eggNOG 2.0 (Muller *et al.*, 2009) demora cerca de 12 minutos. Aumentando para 192.421 parâmetros utilizando o banco de dados eggNOG 4.0 (Powell *et al.*, 2013) o tempo de treino aumenta para 43 minutos. A versão atual, o eggNOG 5.0, possui 206.782 de grupos ortólogos procariotos (Huerta-Cepas *et al.*, 2018). O tempo de treino em SVM cresce proporcional ao cubo do número de dados, dificultando a escalabilidade do método para conjunto de dados maiores ou exigindo uma seleção cuidadosa do conjunto de dados, a qual pode ser extensiva dependendo da complexidade e tamanho do conjunto de dados.

Nas ferramentas Phenotrex e Traitar é discutível a escolha de modelos unificados para realizar predição tanto de bactérias quanto de arqueas, dada a distância evolutiva entre ambos os domínios. Maior acurácia e menor sobreajuste provavelmente seriam alcançados em modelos separados para ambos.

A ferramenta do DiASPora é a mais recentemente atualizada e completa em relação à cobertura dos dados fenotípicos e genotípicos. Em face do aumento de dados fenotípicos e genômicos disponíveis atualmente, as últimas versões das ferramentas Phenotrex e Traitar utilizaram quantidades de dados relativamente baixas, e utilizando como referências as estimativas recentes estando na ordem de milhões (Louca *et al.*, 2019).

2. Desenvolvimento da PheMaLe-bacteria

Neste trabalho foi desenvolvida a ferramenta PheMaLe-bacteria, nome derivado dos termos Phenotype Machine Learning Bacteria, com o objetivo de averiguar a capacidade de algoritmos de aprendizado de máquina em identificarem nos genomas bacterianos os padrões genéticos correlacionados a determinados fenótipos bacterianos. A metodologia empregada na construção e treino dos modelos preditores da PheMaLe-bacteria para cada fenótipo é realizada em 5 etapas:

1. Coleta de informações de múltiplas bactérias e seus fenótipos.
2. Coleta de dados genômicos dessas bactérias.
3. Transformação dos dados genômicos para maior interpretabilidade para aprendizado de máquina.
4. Treino e avaliação de múltiplos algoritmos de aprendizado de máquina.
5. Análise dos parâmetros genéticos considerados mais relevantes nos modelos treinados.

Os detalhes de cada etapa serão detalhados nos subcapítulos a seguir.

2.1. Informações dos fenótipos e genomas bacterianos utilizados

As informações fenotípicas de diversas bactérias foram obtidas na data de 17 de dezembro de 2024 a partir de duas fontes principais: o banco de dados BacDive (Reimer *et al.*, 2021) e o catálogo compilado por Madin *et al.* (2020), que reúne metadados de 26 bancos de dados públicos.

Os genomas bacterianos foram coletados no formato FASTA a partir do banco de dados de genomas anotados do NCBI (Benson *et al.*, 2012) na data de 17 de dezembro de 2024. Para garantir a qualidade e consistência dos dados, foram selecionados apenas os registros que apresentavam o status de genoma completo no NCBI.

Devido à elevada quantidade de registros disponíveis e para evitar redundância genômica, a seleção foi restringida a um único registro representativo por espécie, escolhido aleatoriamente.

2.2. Representação simbólica dos genomas

Para permitir que os algoritmos de aprendizado de máquina identifiquem padrões relacionados a diferentes fenótipos nos genomas, é fundamental que estes,e suas informações relevantes, sejam representados de forma compreensível e adequada para esses algoritmos. Quanto mais claras e discerníveis forem as informações genéticas associadas aos fenótipos, maior será a capacidade dos algoritmos em aprender e identificar os padrões existentes. No caso específico de fenótipos bacterianos, algumas premissas podem ser consideradas para determinar a representação mais adequada dos genomas:

1. Bactérias apresentam a tendência evolutiva a terem genomas compostos principalmente por regiões codificadoras (Mira *et al.*, 2001), portanto a expressão de fenótipos em uma bactéria provavelmente se deve majoritariamente a expressões de genes codificadores.

2. As funções de um gene podem ser mapeadas pela identificação de similaridade do mesmo em relação a grupos de genes ortólogos (Tatusov *et al.*, 1997).

3. A repetição de genes no genoma bacteriano é relevante para a expressão fenotípica (Rocha *et al.*, 1999;).

Com base nessas premissas, é possível concluir que as informações codificadas nas milhares de sequências de nucleotídeos presentes nos genomas podem ser simplificadas para os grupos ortólogos nos quais os genes se enquadram, tornando assim as informações mais legíveis aos algoritmos de aprendizado de máquina. Esse tipo de representação simbólica da informação genética já foi utilizado com sucesso em estudos anteriores, como na predição de interações entre proteínas (Huerta-Cepas *et al.*, 2018) e na correlação entre os grupos ortólogos e fenótipos bacterianos (Jim *et al.*, 2004).

Para realizar o mapeamento dos grupos em cada genoma foi utilizada a ferramenta eggNOG-mapper 2.0.1 (Huerta-Cepas *et al.*, 2017) utilizando o banco de dados de *Cluster of Ortholog Group* (Galperin *et al.*, 2021) e o banco de dados de *Extended Clusters of Orthologous Groups 5.0* (Huerta-Cepas *et al.*, 2018). Este trabalho utiliza a sigla COG para referenciar um grupo ortólogo pertencente ao banco de dados de *Cluster of Ortholog Group* e a sigla eNOG para referenciar um grupo ortólogo hipotético pertencente ao banco de dados EggNOG (Huerta-Cepas *et al.*, 2018).

Após a identificação dos grupos ortólogos em um genoma, estes são estruturados em uma matriz onde as colunas correspondem aos grupos ortólogos e as linhas representam os genes identificados no respectivo genoma. Dessa forma, cada célula da matriz indica a relação entre um gene e um grupo. É importante ressaltar que, devido à multifuncionalidade gênica, um mesmo gene pode apresentar similaridade relevante com múltiplos grupos.

Neste trabalho, foram utilizados dois modos distintos de representação simbólica dos genomas. O primeiro modo consiste em uma representação binária de presença ou ausência de grupos ortólogos. O segundo modo incorpora uma abordagem mais detalhada, que leva em consideração a similaridade do genes com os grupos ortólogos e mantém informação da quantificação presente de cada grupo ortólogo no genoma. Esses dois modos permitem verificar se a informação da presença de genes repetidos altera os resultados ou se a simples indicação de ausência ou presença deles já é informação suficiente para o aprendizado de máquina.

Para obter essas representações, inicialmente é calculado o *e-value* (Pearson, 2013) resultante da comparação entre as sequências de aminoácidos de cada gene da matriz e os grupos ortólogos. O *e-value* representa o número esperado de alinhamentos que poderiam ocorrer ao acaso com uma similaridade igual ou superior àquela observada entre duas sequências de interesse. Dessa forma, quanto menor o *e-value*, menor a probabilidade do alinhamento ser aleatório, indicando uma maior relevância estatística e biológica. Em geral, considera-se *e-values* abaixo de 10^{-6} estatisticamente relevantes (Choudhuri, 2014), embora esse limiar possa variar dependendo do banco de dados utilizado. Foram testados diversos limiares nos treinos dos algoritmos de aprendizado de máquina e adotou-se 10^{-6} como valor máximo para relevância estatística por apresentar os resultados mais consistentes.

Associações entre genes e grupos ortólogos com valores superiores a esse limiar foram descartados.

Contudo, essa escala de relevância de *e-value* apresenta um conflito com a maioria dos algoritmos de aprendizado de máquina, que normalmente atribuem maior importância a valores numéricos mais altos. Para resolver esse problema, é necessária uma reescalonamento dos *e-values* para o padrão "maior é melhor", de forma que valores menores (mais significativos) sejam convertidos em valores mais altos, compatíveis com a interpretação dos algoritmos.

Com essa adaptação, é possível definir o modo de representação genômica binária baseado na ausência ou presença dos grupos ortólogos. Nesse modo, os grupos associados a um genoma são representados como um vetor binário, em que cada posição corresponde a um grupo ortólogo específico: o valor "1" indica a presença de pelo menos um gene associado a esse grupo ortólogo no genoma e o valor "0" indica a ausência. Esse formato simplificado facilita a interpretação e o processamento das informações pelos algoritmos de aprendizado de máquina.

Para o modo de representação de quantificação dos grupos ortólogos foi adotada uma escala contínua de “0” a “1” para representar a confiabilidade da associação entre um gene e um grupo ortólogo. Nesse modo, o valor “0” indica a ausência de relação entre o gene e o grupo, enquanto o valor “1” representa a máxima confiança nessa relação. Dessa forma, a relação entre genes e grupos assume um caráter quantitativo, permitindo medir a relevância dos grupos associados a cada gene. Essa representação numérica é especialmente útil para fenótipos que dependem de repetições de genes, conforme discutido anteriormente no ponto 3.

Para ajustar os *e-values* a essa nova escala, foi formulada uma função de transformação com comportamento sigmoidal (Equação 1). O objetivo dessa função é destacar os *e-values* que representam grande similaridade com grupos ortólogos, sem descartar informações potencialmente relevantes de valores intermediários. A escolha dos parâmetros constantes da Equação 1 foi realizada por meio de múltiplos testes empíricos, a fim de identificar os valores que apresentaram os melhores resultados nos modelos preditores treinados.

Essa abordagem garante que a informação capturada pela escala contínua preserve a relevância biológica e probabilística dos *e-values*, ao mesmo tempo em que se adapta às necessidades dos algoritmos de aprendizado de máquina.

$$f_{(e-value)} = \frac{1}{1 + e^{\frac{log_{10}(e-value)+15)}{5}}} \quad (1)$$

Os valores nessa escala tem o propósito de, quando realizada a soma das linhas matriciais, os *e-values* menos significativos possuírem pequena probabilidade de sobrepujar *e-values* mais significativos mesmo quando presentes em maior quantidade.

Realizando esse processo para todos os genomas é construído então o conjunto de dados primário a ser utilizado. A Figura 1 demonstra um exemplo simplificado do processo de geração de uma matriz utilizando o modo de representação de relevância e quantificação de grupos ortólogos de um genoma hipotético.

Após construído o conjunto de dados para se testar a relevância de eNOGs o conjunto é replicado de duas formas: uma forma na qual o conjunto de dados contém exclusivamente COGs e outra forma na qual o conjunto de dados contém COGs e eNOGs inclusos. Isso permite verificar se a inclusão de grupos ortólogos hipotéticos auxilia no aprendizado de determinado fenótipo, o que potencialmente sugeriria uma relevância biológica desses grupos ortólogos hipotéticos ainda não caracterizados.

Portanto ao fim são geradas 4 representações de parâmetros diferentes:

1. Representação binária de presença e ausência de COGs.
2. Representação binária de presença e ausência de COGs/eNOGs.
3. Representação de relevância e quantificação de COGs.
4. Representação de relevância e quantificação de COGs/eNOGs.

A representação contendo apenas COGs permitirá uma validação em relação à literatura científica existente dos COGs identificados como mais relevantes nos modelos treinados, o que confirmará ou refutará a capacidade do modelo de aprender as intrincadas relações entre COGs e fenótipos. A representação que inclui eNOGs permitirá a identificação de grupos ortólogos hipotéticos que podem potencialmente estar relacionados aos fenótipos em questão.

A representação binária permite identificar grupos ortólogos relevantes independente de possíveis repetições nos genomas, enquanto a representação numérica dará maior relevância a grupos importantes e com menos repetições, além de discernir os fenótipos dependentes de repetições de grupos funcionais e qual seu limiar de repetições para se expressarem.

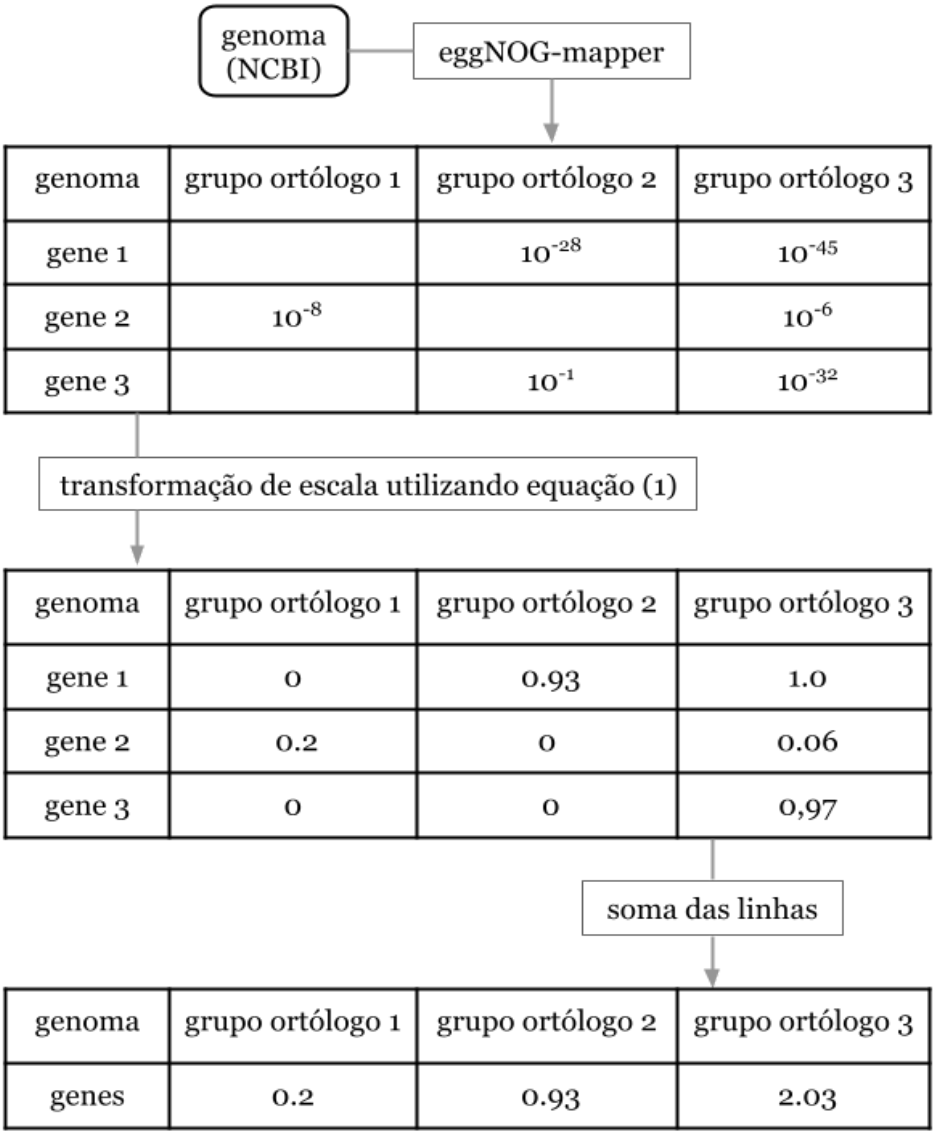

Figura 1: Fluxograma do processo de representação de relevância e quantificação de grupos ortólogos. Aplicado a um genoma hipotético de 3 genes e 3 grupos ortólogos hipotéticos para fins de exemplo.

2.3. Algoritmos de aprendizado de máquina selecionados

Os padrões genéticos relacionados a características geneticamente complexas não são conhecidos a priori, o que impossibilita a determinação antecipada de qual algoritmo de aprendizado de máquina seria o mais adequado para o aprendizado dos padrões genéticos associados a determinadas características bacterianas. Ademais há a possibilidade de mesmo apenas um fenótipo ser derivado de múltiplos padrões genéticos diferentes, além de ser possível que cada padrão siga um modelo de interações gênicas diferente dos outros. Nesse contexto, a estratégia mais eficaz e com cobertura maior é testar múltiplos algoritmos de aprendizado de máquina, cada qual com paradigmas de aprendizado próprios e modos de modelar relações entre os parâmetros diferentes. Esses múltiplos algoritmos podem variar de modelos simples como a Regressão Logística, que considera todos os parâmetros como independentes, até modelos mais complexos como o Random Forest, que leva em consideração interações entre parâmetros. O algoritmo de Redes Neurais Artificiais, apesar de demonstrar grande capacidade de aprendizado, não foi considerado adequado a este projeto pois a quantidade limitada e desbalanceada dos dados o torna suscetível a sobreajuste.

Para comparar os diferentes modelos treinados seus desempenhos são avaliados por meio de métricas em um conjunto de dados independente.

Para os fenótipos de carácter numérico, também conhecido como regressão, foram implementados 14 algoritmos de aprendizado de máquina:

- Regressão Linear
- K-Nearest Neighbours Regression
- Random Forest
- Support Vector Machine
- Ridge
- AdaBoost
- Theil-Sen
- Bernoulli Naive Bayes
- Tweedie
- Passive Aggressive Regression
- Quantile Regression
- Kernel Ridge
- XGBoost
- LightGBM

Para fenótipos de caráter classificatório foram implementados 7 algoritmos de aprendizado de máquina:

- Regressão Logística
- K-Nearest Neighbours
- Random Forest
- Support Vector Machine
- Bernoulli Naive Bayes
- XGBoost
- LightGBM

Os algoritmos foram implementados utilizando Python 3.10 e as bibliotecas de aprendizado de máquina utilizadas foram Scikit-learn 1.0.1, XGBoost 1.3.3 e LightGBM 3.3.0. Os treinamentos foram realizados em um servidor com processador NVIDIA GPU Quadros P6000.

2.4. Treino e avaliação dos modelos treinados

Um aspecto sensível para o aprendizado dos modelos é a configuração de treino, ou os chamados hiper-parâmetros de treino. Para encontrar o conjunto de hiper-parâmetros ideais foi implementada uma busca exaustiva de hiper-parâmetros que minimizem a métrica *log loss* (no caso de fenótipo de classificação) ou R^2 (no caso de fenótipo numérico) em um conjunto de validação independente do treino. Com esse propósito, foi utilizada a função GridSearchCV do Scikit-learn. Devido a grande quantidade de eNOGs catalogados, a busca exaustiva dos melhores hiper-parâmetros é realizada somente para os algoritmos treinados com os conjuntos de dados compostos exclusivamente por COGs, pois ao incluir eNOGs a busca se torna muito exigente computacionalmente. Para lidar com essa limitação, os treinos utilizando os conjuntos de dados que incluem eNOGs são realizados utilizando os mesmos hiper-parâmetros encontrados pela busca exaustiva com os treinos que utilizaram exclusivamente COGs.

Para estabelecer-se independência entre os fenótipos nos modelos treinados foi utilizada a estratégia *one-vs-rest* para todos os algoritmos. Dessa forma, é evitada a possibilidade de um resultado ser inferido pelo modelo a partir do resultado de outra predição realizada por ele, garantindo assim a independência

entre os resultados e apenas a dependência em relação aos grupos ortólogos apresentados.

Os modelos de multi-classificação serão avaliados de acordo com as métricas de Precisão, Recall, F1, AUC-ROC e AUC-PR. Os modelos de regressão serão avaliados de acordo com R^2. Vale ressaltar que o algoritmo Ridge não apresenta resultados baseados em probabilidades para suas predições, portanto não é possível calcular AUC-ROC e AUC-PR para modelos baseados nele.

Posteriormente aos treinos, foi catalogado o fator de influência de cada parâmetro para os modelos, ou seja, a importância de cada COG (e eNOG caso se aplique ao modelo) para a tomada de decisão. Isso foi realizado para cada modelo, salvo os algoritmos SVM com kernel polinomial com grau acima de 1, k-Nearest Neighbors, Ridge Classifier e Kernel Ridge para os quais não é possível uma estimação direta dos fatores de influência dos parâmetros.

2.5. Informações técnicas da PheMaLe-bacteria

A versão atual da PheMaLe-bacteria é capaz de realizar predições para os seguintes fenótipos:

1. Capacidade de fixação de nitrogênio;
2. Faixa de pH que a bactéria consegue sobreviver;
3. Capacidade de endoesporulação;

Para realizar a predição de uma bactéria é necessário o arquivo FASTA da sequência nucleotídica do genoma bacteriano. O resultado da predição é salvo em uma pasta gerada automaticamente juntamente a uma lista de grupos ortólogos encontrados no genoma e que foram relevantes para o resultado da predição.

Para cada fenótipo há 4 opções de preditores, cada um utilizando uma das representações de grupos ortólogos citadas anteriormente. Também é disponibilizado um pipeline para o treino de novos preditores e outros fenótipos.

Para o uso da ferramenta é necessário uma máquina Linux, Python 3.10 e espaço em memória digital mínimo de 22 MB para o uso dos modelos já treinados. Caso o usuário decida gerar novos modelos preditores faz se necessária mais

memória, a depender dos dados genômicos e catálogos fenotípicos utilizados. As predições e treinos podem ser realizados tanto usando CPU quanto GPU. A única dependência externa necessária é a ferramenta eggNOG-mapper 2.0.1 (Huerta-Cepas *et al.*, 2017), a qual atualmente requer 581 MB.

A PheMaLe-bacteria e suas instruções para uso podem ser encontradas no repositório online https://github.com/LaboratorioBioinformatica/PHEMALE.

3. Capacidade de fixação de nitrogênio

O nitrogênio é um elemento crucial a todos os seres vivos, componente necessário aos principais blocos essenciais à vida, como por exemplo aminoácidos, ácidos nucleicos, adenosina trifosfato (ATP) e adenina nicotinamida (NAD) (Nelson & Cox, 2021). No entanto o N_2, a forma molecular mais abundante de nitrogênio, é composta por dois átomos de nitrogênio unidos por uma ligação covalente tripla, uma configuração que torna a molécula altamente inerte, não reativa e, logo, não assimilável por uma grande parcela dos organismos multicelulares (Pi *et al.*, 2022). Portanto, organismos que conseguem transformar o nitrogênio atmosférico em moléculas utilizáveis, como alguns tipos de bactérias e arqueas, são essenciais ao ciclo do nitrogênio na biosfera (Pi *et al.*, 2022). A fixação de nitrogênio é realizada principalmente pelas bactérias diazotróficas, que possuem a enzima nitrogenase, capaz de catalisar a conversão do nitrogênio atmosférico N2 em amônia (NH3) por meio da quebra da ligação covalente e a adição de três átomos de hidrogênio a cada átomo de nitrogênio (Pi *et al.*,2022).

O principal conjunto de genes que faz emergir a propriedade fixadora de nitrogênio em bactérias é conhecido como genes *Nif* (Nitrogen Fixation Genes), que codificam componentes ou componentes precursores da nitrogenase (Lie *et al.*, 2022). Diversos estudos genômicos e transcriptômicos, no entanto, têm revelado que a fixação de nitrogênio não é determinada apenas pelos genes *Nif*, mas também é modulada por fatores metabólicos adicionais, como genes relacionados ao transporte de metais cofatores (molibdênio, ferro, vanádio), mecanismos de proteção contra o estresse oxidativo e genes auxiliares envolvidos na maturação e proteção da nitrogenase (Dos Santos *et al.*, 2012). Além disso, os mecanismos de fixação de nitrogênio são sensíveis a múltiplos fatores ambientais (Dixon & Kahn, 2004), como a concentração oxigênio, os níveis de carbono e a concentração de amônia presentes no ambiente, o que impõe restrições fisiológicas aos diazotróficos. Essas complexas interações entre diferentes grupos de genes e fatores ambientais dificultam a identificação precisa de todos os determinantes genéticos e regulatórios que controlam a capacidade de fixação de nitrogênio, tornando-o um fenótipo geneticamente complexo.

Uma melhor compreensão das vias e mecanismos envolvidos na fixação de nitrogênio por bactérias abriria a possibilidade de aprimoramentos significativos em aplicações biotecnológicas e industriais (Bennett *et al.*, 2023). Atualmente, a agricultura convencional depende majoritariamente do uso de fertilizantes industriais, os quais envolvem a redução de nitrogênio pelo processo Haber-Bosch. No entanto, o uso excessivo desses fertilizantes nitrogenados artificialmente gera uma perturbação no ciclo do nitrogênio que altera quimicamente o solo e a água, resultando na sua degradação (Penuelas *et al.*, 2023). Esse processo altera a composição microbiológica do solo, tornando-o mais dependente de produtos químicos para manter uma composição química saudável para a agricultura. Por outro lado, na água o principal processo de degradação é a eutrofização, ou seja, a proliferação de microrganismos que causam a diminuição dos níveis de oxigênio dissolvido, resultando em uma alta mortalidade de organismos aquáticos e criando as chamadas zonas mortas (Galloway *et al.*, 2003).

Com o desenvolvimento de inoculantes microbianos mais eficientes na fixação de nitrogênio para a agricultura, os impactos ambientais associados à agricultura intensiva seriam minimizados e se tornaria possível a biorremediação de solos e corpos de água já degradados. Isso permitiria também a redução da dependência de fertilizantes artificiais e contribuiria para a redução da poluição gerada no processo e dos custos de produção (Ambrosio *et al.*, 2017; Imran *et al.*, 2021; Kaschuk & Hungria, 2017; Zahran, 1999; Curatti *et al.*, 2005).

Até a presente data, muitos trabalhos investigaram a temática de genes correlacionados à capacidade fixadora de nitrogênio utilizando métodos estatísticos e experimentais, mas apenas a ferramenta Phenotrex (*Phenotrex 0.6.0 Documentation*, s.d.) explorou a aplicação de aprendizado de máquina no estudo desse tópico. Esse trabalho gerou modelos preditivos treinados com os algoritmos SVM utilizando kernel linear e XGBoost utilizando dados de 313 espécies de bactérias e arqueas.

3.1. Informações de espécies bacterianas e sua capacidade de fixação de nitrogênio

Os dados para capacidade de fixação de nitrogênio em espécies de bactérias foram coletados do catálogo compilado por Madin *et al.* (2020), totalizando registros referentes a 3783 bactérias. Desses registros, o NCBI possui genomas completos para 937 espécies. O conjunto de dados obtido apresenta uma diversidade bacteriana de 77 ordens, 160 famílias e 530 gêneros. Na Figura 2 está representada a distribuição percentual das ordens presentes e a quantidade de representantes em cada ordem. As ordens com no máximo 10 representantes são apresentadas em detalhes na tabela presente no Apêndice A.

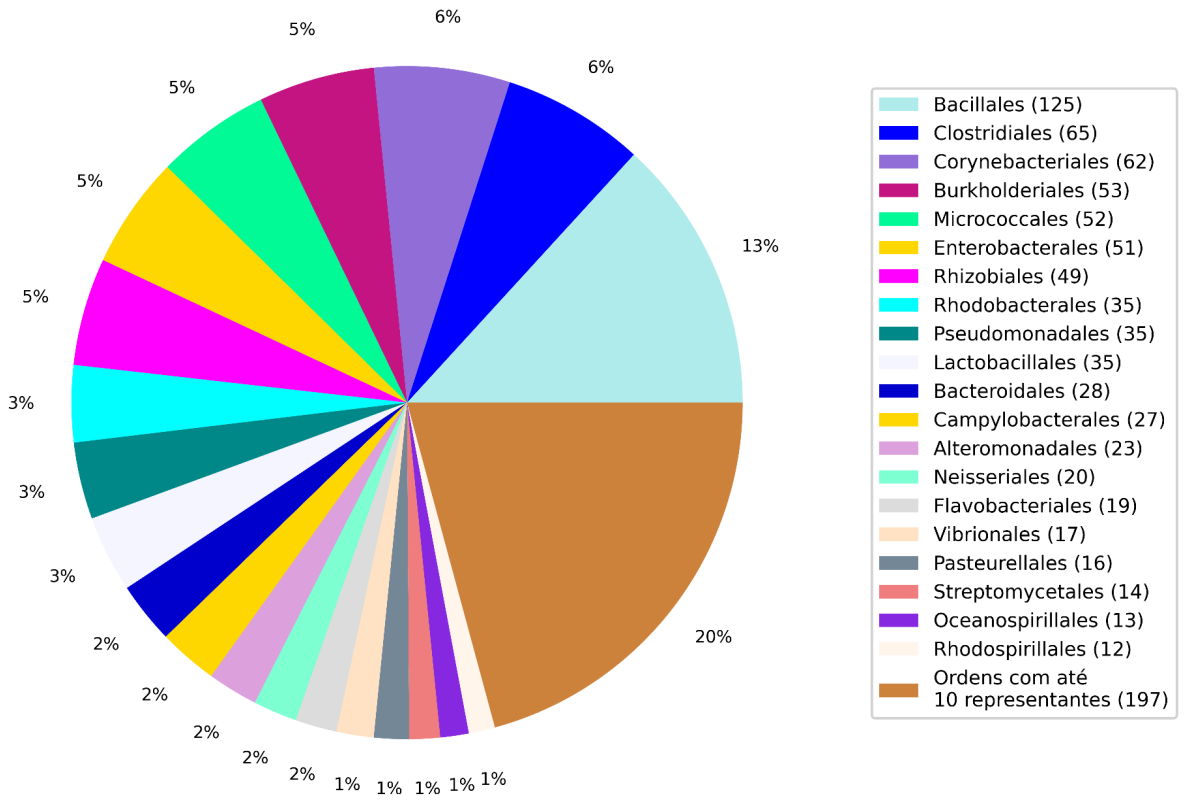

Figura 2: Distribuição de ordens no conjunto de dados para fixação de nitrogênio.

A Figura 3 apresenta a distribuição fenotípica de capacidade de fixação de nitrogênio para os dados coletados. Nestes, apenas 8% do total é composto por bactérias fixadores de nitrogênio. Esse desbalanço é representativo do desbalanço natural existente, onde apenas uma minoria das bactérias apresenta a capacidade de fixação de nitrogênio (Boyd & Peters, 2013).

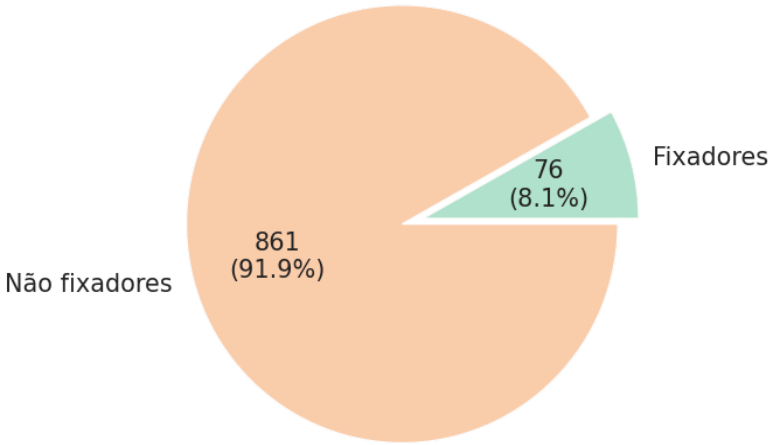

Figura 3: Quantidades de genomas completos no conjunto de dados com informação sobre capacidade de fixação de nitrogênio.

Para avaliar as distâncias euclidianas entre os genomas do conjunto de dados coletados e verificar se há ocorrência de agrupamentos de bactérias, foram gerados gráficos de escalonamento multidimensional (EMD) para cada uma das 4 representações utilizadas para os grupos ortólogos, apresentados na Figura 4.

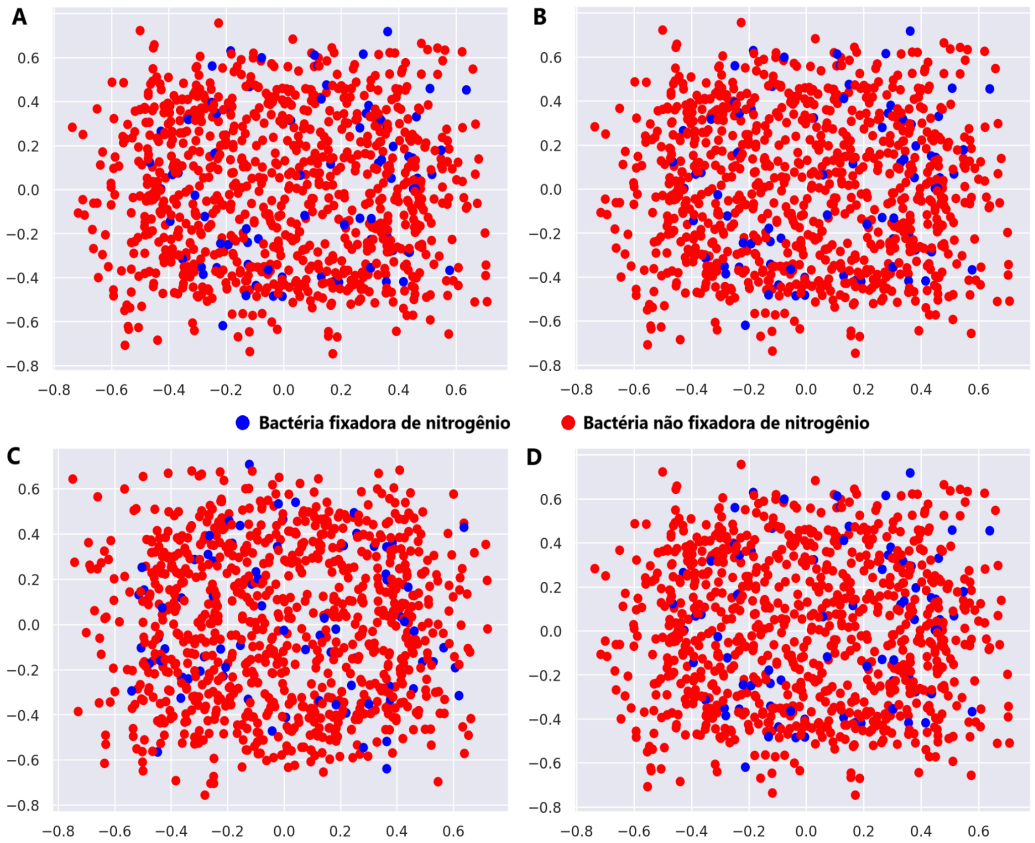

Figura 4: EMD dos genomas com distinção em capacidade de fixação de nitrogênio. Genomas utilizando representações A) binária de COGs; B) binária de COGs e eNOGs; C) quantificação de COGs; D) quantificação de COGs e eNOGs;

Percebe-se uma distribuição homogênea dos genomas pelo espaço de parâmetros dos gráficos e não é possível distinguir uma separação entre os genomas das bactérias fixadoras de nitrogênio e das não fixadoras.

3.2. Resultados da PheMaLe-bacteria

O conjunto de dados, por meio de seleção randômica, foi separado em 3 grupos: 64% dos exemplares do conjunto de dados, ou seja 600 espécies, foram selecionadas para formar o grupo de dados de treino; 16% (149) selecionados como exemplos para validação de treino e 20% do conjunto de dados (188 espécies) foi selecionados para formar o conjunto de dados de teste. Os COGs identificados totalizam 4641, enquanto os eNOGs identificados totalizam 102464.

Na Figura 5 são apresentadas as métricas de avaliação F1, AUC-ROC e AUC-PR dos modelos preditivos da PheMaLe-bacteria no conjunto de dados com as 4 diferentes representações: A) representação binária de presença e ausência de COGs, B) representação binária de presença e ausência de COGs e eNOGs, C) representação de relevância e quantificação de COGs, e D) representação de relevância e quantificação de COGs e eNOGs.

Os 4 grupos apresentaram resultados semelhantes, onde os algoritmos de Regressão Logística e o SVM com kernel linear performaram melhor que os outros algoritmos com a métrica F1 variando na faixa de 78% a 83%, indicando pouca variedade entre os diferentes modos de representação dos dados e a inclusão ou exclusão de eNOGs. O desempenho superior dos modelos baseados em métodos de decisão linear pode indicar que as representações paramétricas dos grupos ortólogos proporcionaram uma separação clara e bem definida, no espaço de parâmetros, entre as bactérias fixadoras e não fixadoras de nitrogênio. Além disso, esse resultado sugere que os genes envolvidos na fixação de nitrogênio possuem um comportamento predominantemente aditivo, facilitando a distinção pelos algoritmos.

Em termos de predição de capacidade fixadora de nitrogênio, o modelo com melhor Precisão foi o treinado com Regressão Logística utilizando a representação binária de COGs, que obteve um valor de 79%. O modelo com melhor Recall foi o treinado com Regressão Logística utilizando a representação quantitativa de COGs,

com um valor de 95%. Na Figura 6 são apresentadas as métricas de Precisão e Recall de ambos os modelos.

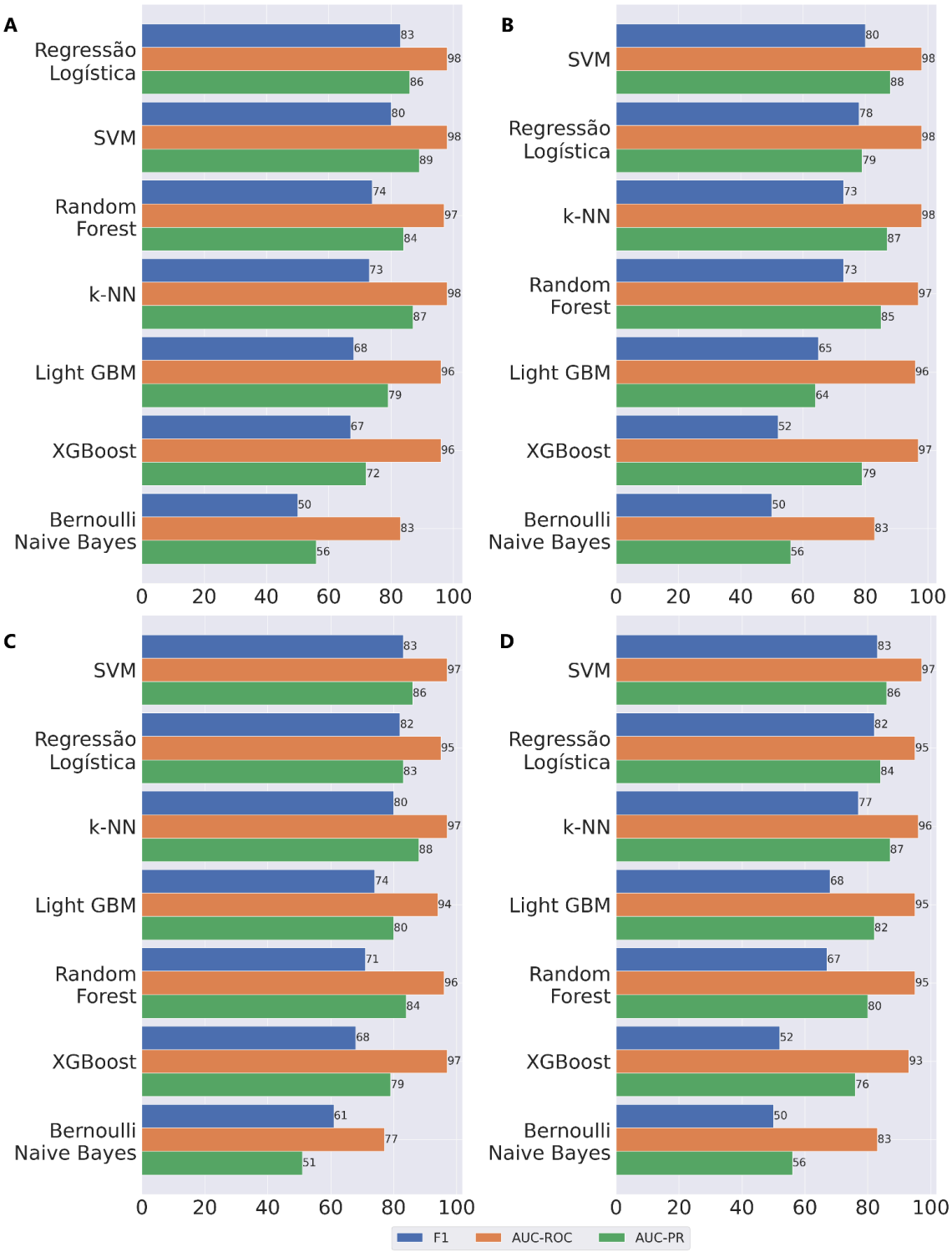

Figura 5: Métricas F1, AUC-ROC e AUC-PR dos modelos de predição de fixação de nitrogênio sob o conjunto de dados teste. Modelos utilizando representação: A) binária de COGs; B) binária de COGs e eNOGs; C) quantitativa de COGs; D) quantitativa de COGs e eNOGs;

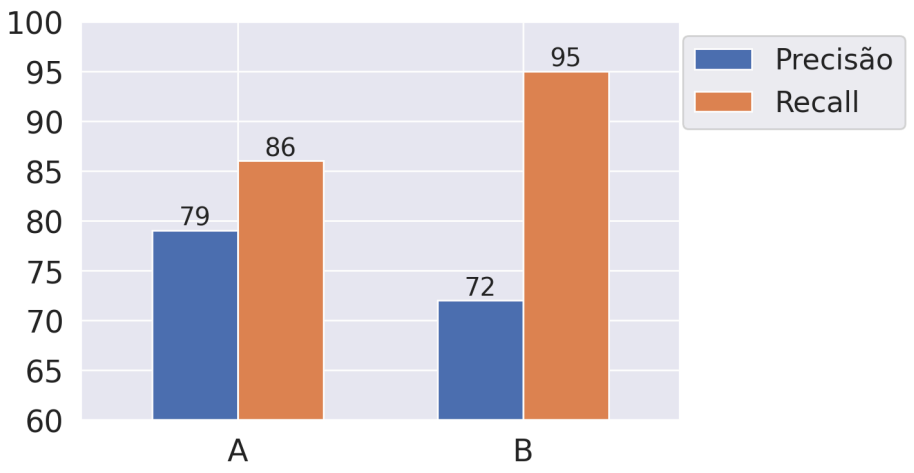

Figura 6: Precisão e Recall dos dados de teste para capacidade de fixação de nitrogênio. Modelos treinados com: A) Regressão Logística com representação binária de COGs; B) Regressão Logística com representação quantitativa de COGs.

Dos 188 exemplos do conjunto de dados de teste, pode-se verificar na Figura 7 as quantidades de predições corretas e erradas realizadas pelo modelo treinado com Regressão Logística e representação binária de COGs e pelo modelo treinado com Regressão Logística representação quantitativa de COGs. As predições corretas se enquadram entre os negativos verdadeiros, ou seja, bactérias não fixadoras que foram preditas corretamente como não fixadoras, e os positivos verdadeiros, bactérias fixadoras que foram preditas corretamente como fixadoras. Nas predições erradas se enquadram os falsos positivos, ou seja, bactérias não fixadoras que foram preditas erroneamente como fixadoras, e os falsos negativos, composto por bactérias fixadoras preditas erroneamente como não fixadoras. Na Tabela 1 estão listadas as espécies do conjunto de testes que foram preditas erroneamente pelos modelos.

Figura 7: Matrizes de confusão para predições de capacidade de fixação de nitrogênio. Modelos treinados com: A) Regressão Logística com representação binária de COGs; B) Regressão Logística com representação quantitativa de COGs.

Tabela 1: Espécies preditas pelos modelos incorretamente em relação a capacidade de fixação de nitrogênio.

Espécie	Modelo(s)	Resultado	Observações
Paraburkholderia tropica	• Reg. Log. com COGs binários • Reg. Log. com COGs quantitativos	Falso positivo	Erroneamente catalogado, verificou-se ser uma fixadora de nitrogênio (Silva *et al.*, 2018).
Methylorubrum zatmanii	• Reg. Log. com COGs binários	Falso positivo	Não há estudos sobre capacidade de fixação de nitrogênio nesta espécie, no entanto apresenta diversos genes relacionados a fixação de nitrogênio como: *NifH* (Gaby & Buckley, 2014), *NifD* (Igarashi & Seefeldt, 2003), *NifA* (Berrabah *et al.*, 2015; Demtröder *et al.*, 2019), *FixX* (Gubler *et al.*, 1989; Sperotto *et al.*, 2004), *NifV* (Nouwen *et al.*, 2017), *NifB* (Li *et al.*, 2021), *IscN* (Sullivan *et al.*, 2013), *ArgH* (Fu *et al.*, 2023; Chen *et al.*, 2024).
Bradyrhizobium daqingense	• Reg. Log. com COGs binários • Reg. Log. com COGs quantitativos	Falso positivo	Erroneamente classificada no catálogo do Madin *et al.* (2020) pois se trata de uma fixadora de nitrogênio (Wang *et al.*, 2013).
Pseudolabrys taiwanensis	• Reg. Log. com COGs binários	Falso positivo	Não há estudos sobre capacidade de fixação de nitrogênio nesta espécie, no entanto apresenta diversos genes relacionados a fixação de nitrogênio como: *FixX* (Gubler *et al.*, 1989; Sperotto *et al.*, 2004), *NifV* (Nouwen *et al.*, 2017), *NifB* (Li *et al.*, 2021), *IscN* (Sullivan *et al.*, 2013), *ArgH* (Fu *et al.*, 2023; Chen *et al.*, 2024).

Clostridium saccharoperbutylacetonicum	• Reg. Log. com COGs binários • Reg. Log. com COGs quantitativos	Falso positivo	Não foram encontradas informações sobre a capacidade desta espécie para fixar nitrogênio, no entanto diversos membros do gênero Clostridium são reconhecidamente capazes de fixar nitrogênio (Chen, 2006).
Achromobacter denitrificans	• Reg. Log. com COGs quantitativos	Falso positivo	Erroneamente classificada no catálogo do *Madin et al.* (2020) pois se trata de uma fixadora de nitrogênio (Mu *et al.*, 2022).
Ca. Methylacidiphilum infernorum	• Reg. Log. com COGs quantitativos	Falso positivo	Não foi confirmado experimentalmente, mas se supõe ser capaz de fixar nitrogênio por possuir genes de fixação de nitrogênio suficientes (Hou *et al.*, 2008).
Cupriavidus oxalaticus	• Reg. Log. com COGs quantitativos	Falso positivo	Não há estudos sobre capacidade de fixação de nitrogênio nesta espécie, no entanto apresenta diversos genes relacionados a fixação de nitrogênio como: *NifA* (Berrabah *et al.*, 2015; Demtröder *et al.*, 2019), *FixX* (Gubler *et al.*, 1989; Sperotto *et al.*, 2004), *NifV* (Nouwen *et al.*, 2017), *NifB* (Li *et al.*, 2021), *IscN* (Sullivan *et al.*, 2013), *ArgH* (Fu *et al.*, 2023; Chen *et al.*, 2024).
Cupriavidus necator	• Reg. Log. com COGs quantitativos	Falso positivo	Erroneamente classificada no catálogo do *Madin et al.* (2020) pois se trata de uma fixadora de nitrogênio (da Silva *et al.*, 2012).

Microcystis aeruginosa	• Reg. Log. com COGs binários • Reg. Log. com COGs quantitativos	Falso Negativo	Erroneamente classificada no catálogo do Madin *et al.* (2020) pois se trata de uma não fixadora de nitrogênio (Yue *et al.*, 2014; Yang *et al.*, 2023).
Herbaspirillum rubrisubalbicans	• Reg. Log. com COGs binários	Falso Negativo	Trata-se de um fixador de nitrogênio real (Baldani *et al.*, 1996). A probabilidade calculada pelo preditor do genoma dessa bactéria ser um fixador de nitrogênio foi de 48%, indicando a necessidade do limiar de probabilidade ser reajustado.
Frankia casuarinae	• Reg. Log. com COGs binários	Falso Negativo	Trata-se de um fixador de nitrogênio real (Kucho *et al.*, 2017). A probabilidade calculada pelo preditor do genoma dessa bactéria ser um fixador de nitrogênio foi de 49%, indicando a necessidade do limiar de probabilidade ser reajustado.

A taxa de Falsos Negativos reportados pode ser reduzida ajustando o limiar de decisão dos modelos, o qual se encontra no padrão de 0.5, para um limiar mais baixo, o que aumentaria a detecção de Verdadeiros Positivos e diminuiria a taxa de Falsos Negativos. Esse ajuste permitiria encontrar um equilíbrio entre precisão e recall e pode ser determinado pelas curvas ROC e a curva de precisão x recall apresentados nas Figuras 8 e 9 para os dois modelos selecionados. No entanto, a determinação exata deste limiar foge ao escopo deste trabalho, e fica a cargo de um trabalho futuro.

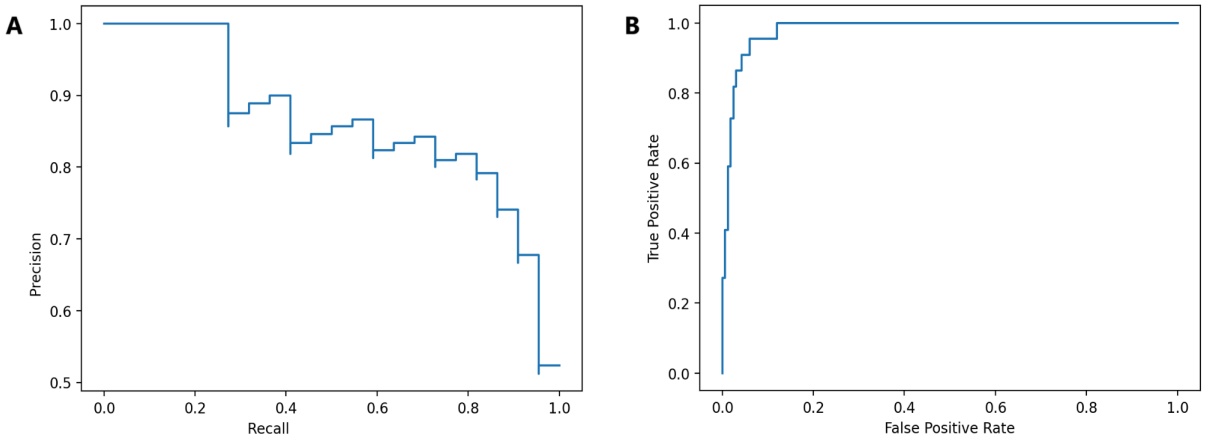

Figura 8: Métricas do treino do modelo treinado com Regressão Logística utilizando representação binária de COGs para classificação de capacidade de fixação de nitrogênio. A) curva de precisão x recall; B) curva ROC.

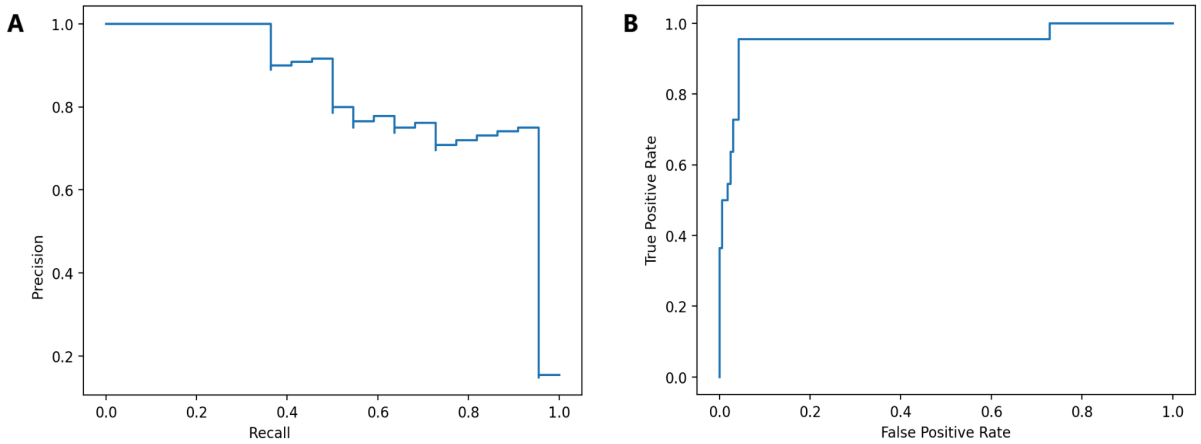

Figura 9: Métricas do treino do modelo treinado com Regressão Logística utilizando representação quantitativa de COGs para classificação de capacidade de fixação de nitrogênio. A) curva de precisão x recall; B) curva ROC.

Levando em conta os erros experimentalmente confirmados presentes no catálogo compilado por Madin *et al.* (2020), as métricas de avaliação e a matriz de confusão para o conjunto teste se alteram de acordo com a Figura 10.

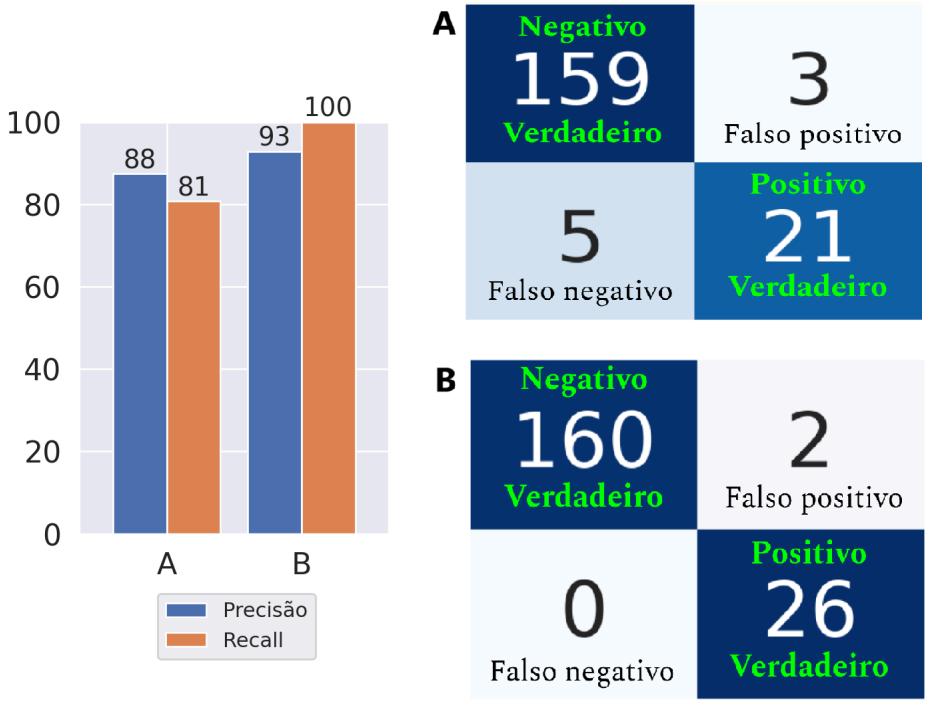

Figura 10: Precisão e Recall e respectivas matrizes de confusão para o conjunto de teste corrigido. Modelos treinados com A) Regressão Logística com representação binária de COGs; B) Regressão Logística com representação quantitativa de COGs de fixação de nitrogênio.

Para comparar os resultados obtidos com os modelos treinados verificou-se as predições da ferramenta Phenotrex 0.6.0 para o mesmo conjunto de dados de teste, após realizadas as correções identificadas. As métricas de performance do Phenotrex 0.6.0 são apresentadas na Figura 11. Pode-se notar pela matriz de confusão e pelas métricas de Precisão e Recall que a ferramenta Phenotrex 0.6.0 apresenta uma tendência a classificar as bactérias erroneamente como fixadoras de nitrogênio, indicando um provável sobreajuste do modelo utilizado pela ferramenta.

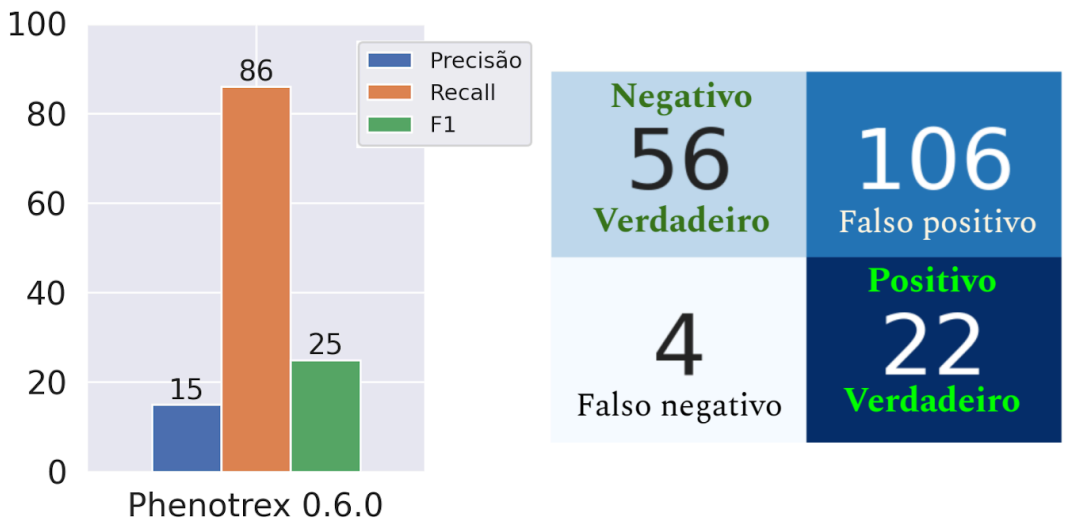

Figura 11: Precisão, Recall e F1 de Phenotrex 0.6.0 com o conjunto teste de fixação de nitrogênio e respectiva matriz de confusão.

3.3. Análise de influência dos grupos ortólogos

A partir dos modelos já treinados, foram extraídos os fatores de influência que cada grupo ortólogo tem sobre a decisão final dos modelos. Um fator de influência positivo indica que o grupo em questão contribui favoravelmente para a previsão de que uma bactéria possui capacidade de fixação de nitrogênio. Por outro lado, um fator de influência negativo sugere que o grupo exerce a influência contrária, indicando que a bactéria provavelmente não possui essa capacidade. A escala dos fatores de influência varia de acordo com a estrutura interna do modelo, portanto o fator de influência de um grupo ortólogo em um modelo não é comparável ao de um grupo ortólogo em outro modelo.

Na Figura 12 são apresentados os 20 grupos ortólogos mais influentes e os seus fatores de influência no modelo treinado com Regressão Logística e representação binária exclusivamente de COGs. Desses COGs dá-se especial atenção à presença do COG1348, COG2710 e COG3604, todos COGs relacionados a genes *Nif*, um grupo de genes diretamente responsáveis por atividades de fixação de nitrogênio (Dos Santos *et al.*, 2012).

Com a ausência do gene *NifB* na lista dos 10 grupos ortólogos mais relevantes, o modelo conseguiu identificar 5 dos 6 genes considerados essenciais para a atividade de fixação de nitrogênio (Dos Santos *et al.*, 2012). O gene *NifB* pode ser encontrado tanto em bactérias fixadoras de nitrogênio quanto em bactérias não

fixadoras sendo, portanto, essa a provável causa para o modelo não considerar esse gene como relevante para o discernimento entre fixadoras e não fixadoras.

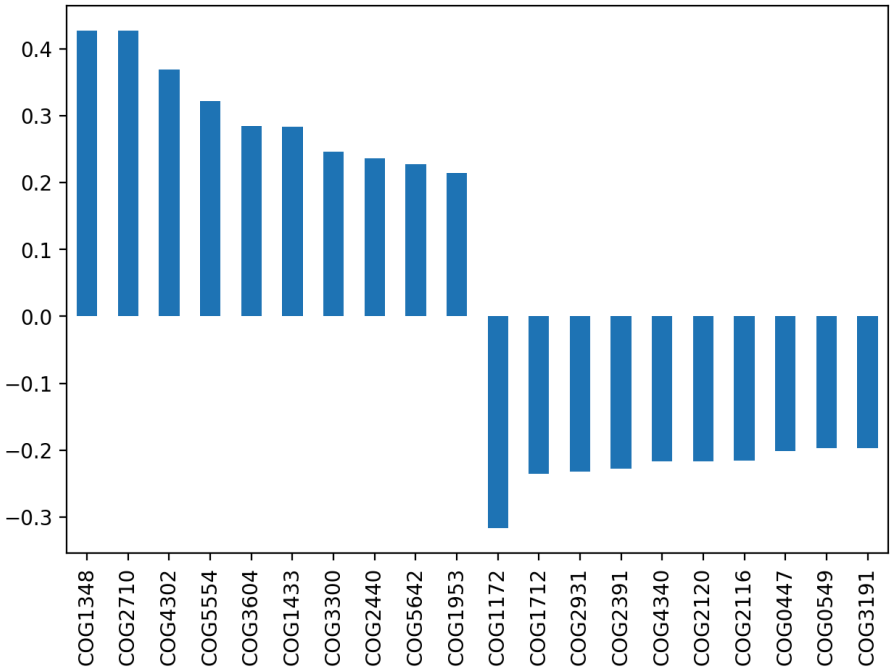

Figura 12: 20 grupos ortólogos mais influentes para predição de fixação de nitrogênio no modelo treinado com Regressão Logística representação binária de COGs.

Desses COGs foram encontradas as seguintes informações na literatura científica detalhando sua função e suas relações em relação à capacidade de fixação de nitrogênio:

- COG1348 (COG1348 - NCBI, s.d.; KEGG ORTHOLOGY: K02588, s.d.) é responsável pelo gene *NifH*, que codifica a componente redutase da dinitrogenase. Ela é responsável por transferir elétrons para a componente catalítica da nitrogenase, desempenhando um papel crucial na redução do nitrogênio atmosférico em amônia (Gaby & Buckley, 2014).

- COG2710 (KEGG ORTHOLOGY: K02586, s.d.; KEGG ORTHOLOGY: K02587, s.d.; KEGG ORTHOLOGY: K02591, s.d.; KEGG ORTHOLOGY: K02592, s.d.) é responsável pela proteína *NifD*, que codifica a subunidade α da dinitrogenase, a componente catalítica da nitrogenase, essencial para a atividade enzimática, facilitando a ligação e a redução do nitrogênio atmosférico em amônia (Igarashi & Seefeldt, 2003).

● COG4302 (KEGG ORTHOLOGY: K03736, s.d.) está associado à uma subunidade menor da etanolamina amônia-liase, *EutC*, envolvida na degradação da etanolamina em etanol e amônia, o que permite que seja utilizada como fonte de carbono e nitrogênio (Tsoy *et al.*, 2009).

● COG5554 (KEGG ORTHOLOGY: K02593, s.d.) está associado à família *NifT* (Wen *et al.*, 2020). Sua função ainda é desconhecida, mas verificou-se experimentalmente que a desativação do *NifT* diminui a atividade de fixação de nitrogênio na *Klebsiella pneumoniae,* apesar de não a anular totalmente (Simon *et al.*, 1996). Uma hipótese em voga é que a proteína possa estar envolvida na biossíntese do cofator ferro-molibdênio (FeMo) da nitrogenase (Stricker *et al.*, 1997).

● COG3604 (COG3604 - NCBI, s.d.) é responsável pela proteína *Nif*A, que codifica um ativador transcricional essencial para a expressão dos genes *Nif* (Berrabah *et al.*, 2015; Demtröder *et al.*, 2019). O *NifA* ativa a transcrição dos genes de fixação de nitrogênio em resposta a estímulos ambientais, como baixa disponibilidade de nitrogênio.

● COG1433 (COG1433 - NCBI, s.d.) está associado à proteína *nifX*, que desempenha um papel na biossíntese e estabilização do cofator FeMo da nitrogenase, essencial para a atividade do complexo enzimático nitrogenase responsável pela fixação de nitrogênio em bactérias (Cort *et al.*, 2000).

● COG3300 (COG3300 - NCBI, s.d.) está associado ao domínio MHYT, que é um domínio sensor de membrana, possivelmente envolvido na transdução de sinais relacionados ao óxido nítrico (Galperin et al., 2001).

● COG2440 (COG2440 - NCBI, s.d.) é responsável pela proteína FixX, tipo ferredoxina, essencial para a atividade da nitrogenase devido à sua provável atuação em transferência de elétrons (Gubler *et al.*, 1989; Sperotto *et al.*, 2004).

● COG5642 (COG5642 - NCBI, s.d.) é responsável por um componente antitoxina previsto de um sistema antitoxina-toxina, contendo domínio HTH tipo Xre. Não foram identificadas na literatura relações deste COG com a capacidade de fixação de nitrogênio.

- COG1953 (COG1953 - NCBI, s.d.) é responsável pela permease citosina/uracil/tiamina/alantoína associada ao catabolismo de purinas, as quais são fontes de nitrogênio (Huynh & Stewart, 2023).

Na Figura 13 são apresentados os 20 grupos ortólogos mais influentes e seus fatores de influência no modelo treinado com Regressão Logística e representação quantitativa de COGs. Nota-se que há uma concordância entre este modelo e o modelo treinado com Regressão Logística e representação binária dos COGs: COG5554, COG3604, COG4302 e COG1348.

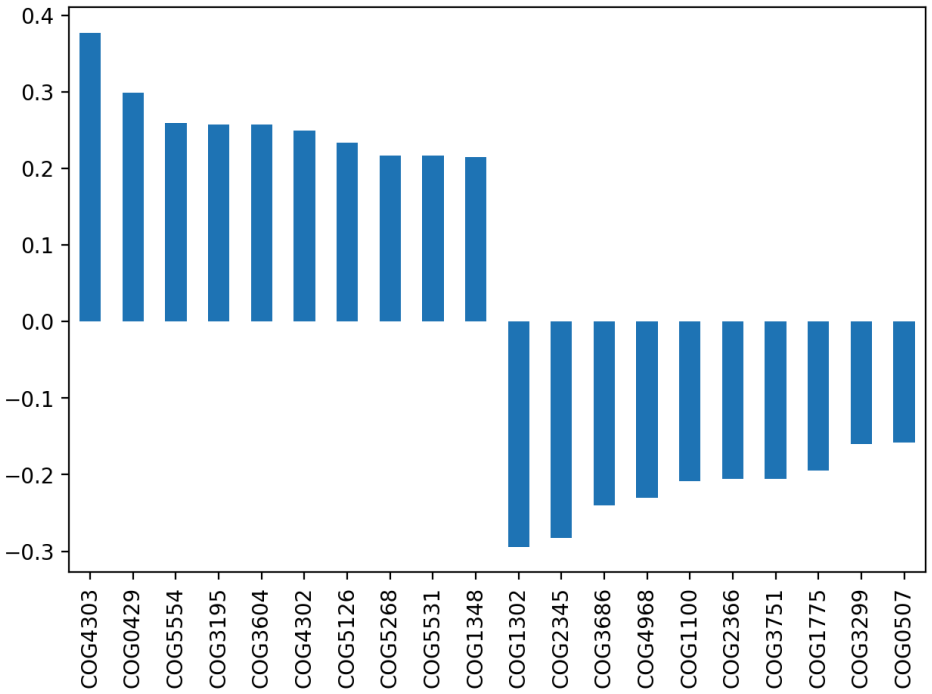

Figura 13: 20 grupos ortólogos mais influentes e seus fatores de influência para predição de fixação de nitrogênio do modelo treinado com Regressão Logística e representação quantitativa de COGs.

Desses COGs selecionados pelo modelo como mais influentes para predição positiva para capacidade de fixação de nitrogênio foram encontradas as seguintes informações na literatura científica detalhando sua função e suas relações em relação à capacidade de fixação de nitrogênio:

- COG4303 (COG4303 - NCBI, s.d.) é responsável pela enzima *EutB*, semelhante ao COG4302 apresentado anteriormente se trata de uma etanolamina amônia-liase, responsável pela metabolização de etanolamina em etanol e amônia (Tsoy *et al.*, 2009).

● COG0429 (COG0429 - NCBI, s.d.) se trata de uma família de hidrolases previstas caracterizadas por uma dobra α/*β*-hidrolase. Os papéis das proteínas associadas ao COG0429 permanecem não caracterizados. Sua presença em diversas espécies bacterianas sugere que desempenham papéis fundamentais no metabolismo celular. Sua estrutura é semelhante a enzimas que catalisam a hidrólise de várias ligações químicas, incluindo ligações éster, tioéster e amida (Holmquist, 2000; Shoda *et al.*, 2016).

● COG3195 (COG3195 - NCBI, s.d.) é responsável pelo 2-oxo-4-hidroxi-4-carboxi-5-ureidoimidazolina descarboxilase, importante para a decomposição do ácido úrico, umas das principais fontes de *N* excretadas por animais (Wright, 1995). A decomposição conduz a formação de ureia, que por hidrólise catalisada pela urease produz NH_4+ e CO_2 (LeVeen *et al.*, 1994; Wright, 1995).

● COG5126 (COG5126 - NCBI, s.d.) também chamado *FRQ1* verificou-se atividade elevada em ambiente rico em nitrogênio (Rosman *et al.*, 2024).

● COG5268 (COG5268 - NCBI, s.d.) está relacionado ao componente *TrbD* da via secretora tipo IV, que está envolvida no transporte de macromoléculas para a membrana celular. Não há conhecimento de relação direta entre o COG5268 e fixação de nitrogênio.

● COG5531 (COG5531 - NCBI, s.d.) se trata do domínio SWIB/MDM2 de ligação ao DNA. Não foi encontrada literatura relacionando este COG com atividade fixadora de nitrogênio.

No modelo treinado com Regressão Logística e representação binária de COGs o COG1348 ocupa a posição de maior influência, enquanto no modelo treinado com Regressão Logística e representação quantitativa de COGs é rebaixado para a 10ª posição dos COGs mais influentes, com seu fator de influência diminuindo para aproximadamente 0.2, aproximadamente metade do fator de influência apresentado no modelo utilizando representação binária. Isto indica possivelmente a necessidade de repetições deste COG nos genomas para apresentar influência relevante a fixação de nitrogênio.

Há também uma presença maior de COGs não diretamente associados à fixação de nitrogênio. É teorizado que estes COGs possam estar relacionados a processos metabólicos ou vias regulatórias que apoiem indiretamente o metabolismo de fixação de nitrogênio.

Há alguns COGs que não se configuram entre os 10 mais influentes nos modelos analisados mas que se destacaram entre os 20 mais influentes que são interessantes de destacar:

- COG3633 (COG3633 - NCBI, s.d.) que se trata de *SstT*, simportador Na+/serina e Na+/treonina, e está relacionado a sistemas de transporte para fontes de nitrogênio (Doranga & Conway, 2024).

- COG0119 (R9VGY9 - UniProt., s.d.) se trata da *NifV*, que codifica uma homocitrato sintase que catalisa a condensação de acetil coenzima A e 2-oxoglutarato. O homocitrato é um componente do cofator FeMo presente no centro catalítico da dinitrogenase (Nouwen *et al.*, 2017). Em estudos com bactérias fixadoras de nitrogênio, como *Klebsiella pneumoniae* e *Azotobacter vinelandii*, mutações no gene *NifV* apresentaram redução ou ausência de atividade nitrogenase (Kennedy & Dean, 1992).

- COG0535 (KEGG ORTHOLOGY: K02585, s.d.) se trata da *NifB*, uma enzima radical S-adenosil metionina que catalisa a formação de NifB-co, precursor comum da síntese dos cofatores FeMo para nitrogenase FeMo, o cofator FeV para nitrogenase V e o cofator FeFe da nitrogenase Fe (Li *et al.*, 2021).

- COG0715 (P73452 - UniProt, s.d.) é responsável pela proteína de ligação de nitrato/nitrito NrtA. Faz parte do complexo transportador ABC NrtABCD envolvido na captação de nitrato e provavelmente também está envolvido no transporte de nitritos (Koropatkin *et al.*, 2006).

- COG5420 (COG5420 - NCBI, s.d.) é responsável por uma proteína não caracterizada associada a fixação de nitrogênio (Wen *et al.*, 2020; Yang *et al.*, 2021).

- COG0316 (COG0316 - NCBI, s.d.) é responsável pela proteína IscN, verificado experimentalmente por optimizar a fixação de nitrogênio em até 90% em algumas espécies de bactérias (Sullivan *et al.*, 2013).

● COG0165 (COG0165 - NCBI, s.d.) responsável pela proteína ArgH, argininosuccinato liase. Foi verificado experimentalmente que o aumento de sua expressão está correlacionado à presença de nitrogênio (Fu *et al.*, 2023; Chen *et al.*, 2024).

● COG0830 (COG0830 - NCBI, s.d.) responsável pela proteína urease UreF, que atua na incorporação de íons níquel na enzima urease, que metaboliza ureia, uma fonte de nitrogênio para bactérias (Liu *et al.*, 2022).

● COG0462 (EggNOG Database | COG0462, s.d.) Domínio *N*-terminal da ribose fosfato pirofosfoquinase. Não foi encontrada literatura relacionando este COG com atividade fixadora de nitrogênio.

Os modelos que foram treinados utilizando eNOGs juntamente de COGs também foram analisados na questão de seus eNOGs mais influentes, os quais podem ser encontrados no Apêndice C. Em ambos os modelos, o grupo ortólogo hipotético eNOG28T9W se posicionou entre os eNOGs mais influentes, apresentando um fator de influência aproximadamente 5 vezes maior que os outros grupos, COG ou eNOG, e para o qual há confirmação experimental de atividade fixadora de nitrogênio em bactérias. Trata-se de uma proteína regulatória HutP relacionada à degradação de L-histidina, permitindo seu uso como fonte de carbono ou nitrogênio (Kumarevel *et al.*, 2004). Não foram encontradas informações sobre os outros eNOGs analisados.

3.4. Conclusão

Pode-se averiguar pelos resultados das métricas de avaliação dos modelos treinados que os algoritmos de aprendizado de máquina foram capazes de discernir e aprender os padrões genéticos relacionados ao fenótipo de capacidade de fixação de nitrogênio, mesmo em face a um conjunto de dados altamente desbalanceado. Além disso, os modelos foram capazes de perceber erros contidos no conjunto de dados de onde foram obtidos os dados de treino.

Com uma precisão de 93% e recall de 100%, o modelo treinado com algoritmo de Regressão Logística utilizando representação quantitativa de COGs apresentou as melhores métricas de avaliação. O modelo treinado com o mesmo algoritmo,

Regressão Logística, mas utilizando dados com representação binária de COGs apresentou métricas de avaliação piores, indicando que possivelmente ocorreu alguma perda de informação relevante ao utilizar dados binários.

Ainda no tópico de representação de dados, o fato de algoritmos de aprendizado simples como a Regressão Logística e o SVM com kernel linear figurarem com as melhores métricas de avaliação entre os modelos indica que a representação do genoma como grupos ortólogos, seja binário ou quantizado, representa as bactérias fixadoras e não fixadoras de modo discernível no espaço de parâmetros.

Os modelos analisados conseguiram discernir os grupos ortólogos relacionados ao fenótipo de fixação de nitrogênio como diversos pertencentes ao grupo de genes *nif* (Dos Santos *et al.*, 2012). Os modelos treinados utilizando grupos ortólogos hipotéticos conjuntamente com os grupos ortólogos curados identificaram corretamente o grupo ortólogo hipotético 28T9W como relacionado a fixação de nitrogênio, tornando os outros grupos ortólogos hipotéticos identificados potenciais focos de pesquisa no tópico de fixação de nitrogênio.

Por fim, os modelos com a nova metodologia de representação obtiveram métricas superiores em relação a ferramenta Phenotrex (Feldbauer *et al.*, 2015) na predição de fixação de nitrogênio, indicando um avanço nessa área.

4. Tolerância de bactérias a faixas de pH

Compreender a faixa de pH que as bactérias podem tolerar é fundamental para diversas áreas da microbiologia, pois o pH do ambiente influencia diretamente diversos aspectos do crescimento, da sobrevivência e das atividades metabólicas.

A maioria das bactérias conhecidas prospera em pH neutro, aproximadamente entre 6,5 e 7,5, sendo classificadas como neutrófilas. No entanto, algumas espécies adaptaram-se a condições extremas, como as acidófilas, que crescem em pH ácido, e as alcalófilas, que preferem ambientes alcalinos. Essa capacidade de tolerar ou crescer em diferentes valores de pH depende de mecanismos de homeostase celular que mantêm o equilíbrio ácido-base intracelular mesmo em condições externas adversas (Krulwich *et al.*, 2011). Essa habilidade é fundamental para a adaptação das bactérias em ambientes extremos, como solos ácidos, reservatórios alcalinos, o trato gastrointestinal de organismos hospedeiros e ambientes industriais.

As bactérias se utilizam de alguns mecanismos de proteção para lidar com flutuações de pH ou valores de pH extremos. Um desses mecanismos é a capacidade tampão do citoplasma para manter a homeostase interna da célula. Alguns prótons entrando ou saindo da célula já seriam suficientes para causar um grande impacto no pH interno se não houver capacidade tampão suficiente (por exemplo, fosfatos inorgânicos e orgânicos) para absorvê-los (Poolman, 2023). Outro mecanismo utilizado pelas bactérias é a manutenção da homeostase pH por meio da remodelação dos padrões metabólicos, como por exemplo o aumento da expressão de enzimas cujas reações consomem prótons citoplasmáticos, incluindo hidrogenases específicas e descarboxilases de aminoácidos (Krulwich *et al.*, 2011). No caso de bactérias acidófilas, a respiração é um dos principais modos de consumo dos prótons, pelo consumo durante a redução de NAD(P) + e O_2 para formar NAD(P)H e H_2O (Quatrini *et al.*, 2009). Outra opção são os múltiplos sistemas de transporte para diferentes sensibilidades de pH (Masrati *et al.*, 2018; Winkelmann *et al.*, 2022). Tais transportadores incluem bombas de prótons (prótons-ATPases) em casos de acidificação externa, e as bombas de K^+ ou Na^+ (sódio-potássio-ATPases) em casos de em ambientes alcalinos, além de transportadores ativos secundários, como antiportadores de prótons e cátions. Esses mecanismo descritos são mecanismos ativos, mas há também mecanismos

passivos de homeostase envolvidos, como as mudanças na permeabilidade dos prótons da membrana e na carga elétrica da superfície celular para retardar a entrada ou perda de prótons no citoplasma (Krulwich *et al.*, 2011).

Também há a necessidade de mecanismos de proteção contra outros tipos de estresse correlacionados a processos de acidificação ou alcalinização, como os estresses osmótico e oxidativo. Por exemplo, o cloreto causa tanto a acidificação quanto um desbalanço osmótico e estresse oxidativo (Rivera-Araya *et al.*, 2019).

Com toda essa miríade de mecanismos necessários, pode-se notar que, do ponto de vista genético a capacidade de bactérias de lidarem com diferentes valores de pH é derivada de uma ampla gama de mecanismos complexos, originada de múltiplos genes, sinalizadores e redes regulatórias sofisticadas. A identificação dos determinantes genéticos responsáveis pela tolerância ao pH é ainda mais desafiadora devido à interação de múltiplos fatores ambientais e metabólicos que influenciam a expressão gênica. Estudos baseados em abordagens genômicas e transcriptômicas têm revelado que bactérias utilizam vias redundantes e complementares para se adaptar a diferentes valores de pH, o que reforça a natureza multifatorial dessa característica (Slonczewski *et al.*, 2009). O melhor entendimento desses mecanismos pode ser útil para diversas áreas da vida humana e ambiental. Pode, por exemplo, informar como ajustar o pH do ambiente para que bactérias desempenhem suas funções metabólicas de modo otimizado. Ademais, pode auxiliar na seleção de bactérias que possuam maior capacidade de sobrevivência em ambientes extremos, como usualmente é o caso na área industrial (Qi *et al.*, 2019). Por fim, é um conhecimento imprescindível também no combate às bactérias patogênicas ou invasoras.

Até a data desta publicação, o trabalho mais recente relacionado à aplicação de aprendizado de máquina à predição de pH tolerado por bactérias foi publicado por Ramoneda *et al.* (2023). Neste trabalho, os autores reuniram dados de bactérias coletadas dos solos do Panamá, da Austrália e da Carolina do Sul nos EUA, e em água doce do Canadá. Estes dados foram utilizados para treinar um modelo utilizando o algoritmo XGBoost para a predição do pH teórico preferido pelas bactérias dentro da faixa de pH de 4 a 9. O pH teórico preferido foi definido como a média da faixa de pH registrada dos ambientes de coleta. Esse modelo obteve uma métrica de performance de $R^2 = 0.21$ em um conjunto de dados independente coletado no Reino Unido.

Não foram encontrados trabalhos que se propusessem a predizer as faixas de tolerância de pH.

4.1. Espécies bacterianas e as faixas de pH toleradas

Os dados para faixas de pH toleradas por espécies de bactérias foram coletados do banco de dados BacDive, totalizando registros referentes a 5659 espécies bacterianas. Desses registros, o NCBI possui genomas completos para 1021 espécies. O conjunto de dados obtido apresenta uma diversidade bacteriana de 93 ordens, 218 famílias e 559 gêneros. Na Figura 14 está apresentada a distribuição percentual das ordens e a quantidade de representantes em cada ordem. As ordens com no máximo 10 representantes são apresentadas em detalhes no Apêndice D.

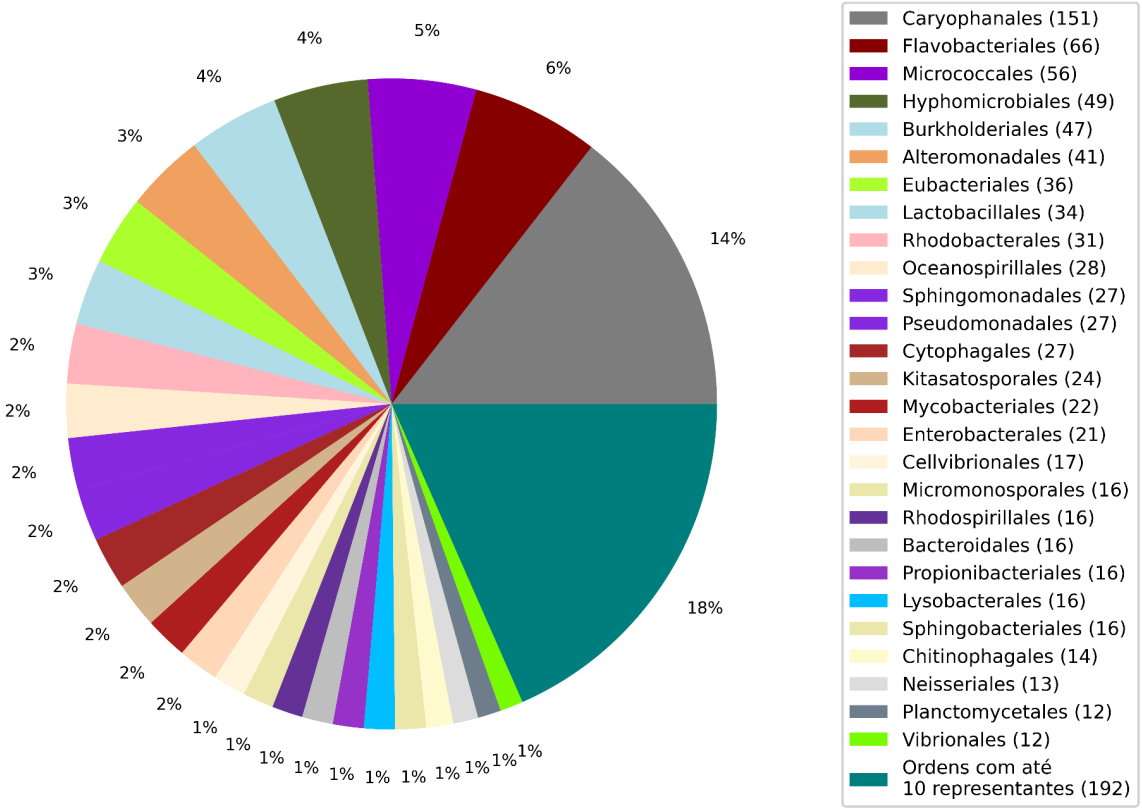

Figura 14: Distribuição de ordens no conjunto de dados para faixas de pH.

Na Figura 15 são apresentadas as distribuições de quantidade de espécies em relação aos valores de seus limites inferior e superior de pH tolerados. Pode-se notar que a maior parte dos dados apresenta limite inferior em 6 e limite superior

em 9 e menos nos extremos, uma distribuição esperada dado que as bactérias mais estudadas e de fácil obtenção se tratam das encontradas em pH próximo ao neutro.

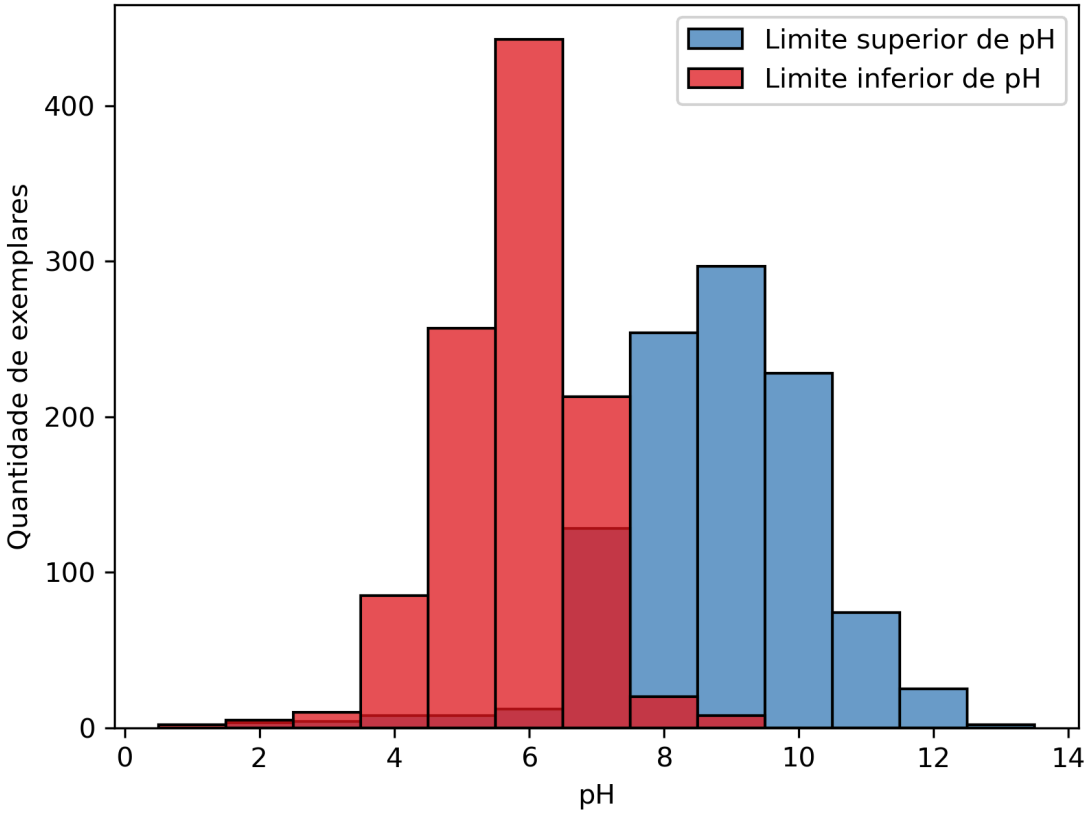

Figura 15: Quantidade de espécies no conjunto de dados de acordo com seus limites inferior e superior de pH.

Para analisar as distâncias e por agrupamentos entre os genomas do conjunto de dados, foram gerados gráficos EMD utilizando distância euclidiana para cada uma das 4 representações utilizadas para os grupos ortólogos. Os gráficos são apresentados na Figura 16 a seguir. Percebe-se uma distribuição homogênea dos genomas pelo espaço de parâmetros dos gráficos e não é possível distinguir uma separação entre bactérias que toleram pH alcalino, neutro e ácido.

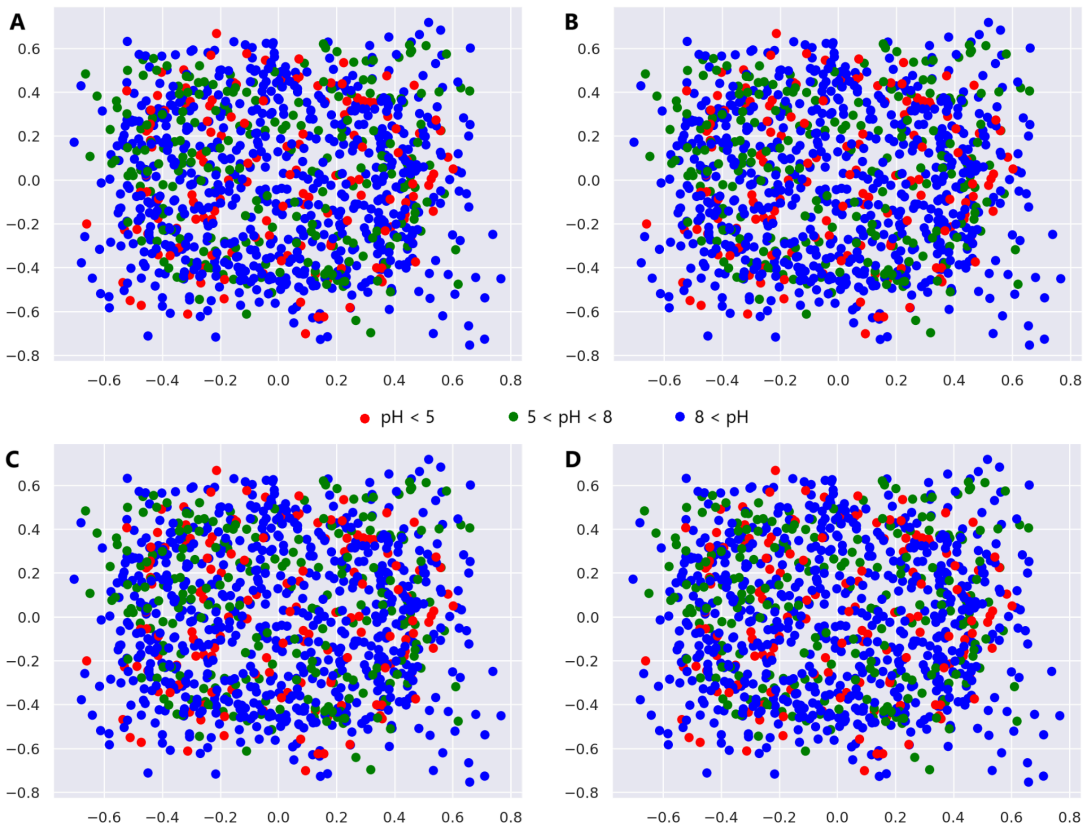

Figura 16: EMD dos genomas com informação sobre faixas de pH associadas. Genomas utilizando representações: A) binária de COGs; B) binária de COGs e eNOGs; C) quantificação de COGs; D) quantificação de COGs e eNOGs;

4.2. Resultados da PheMaLe-bacteria

Do conjunto de dados, por via de seleção randômica, 20% das espécies (204 espécies) foram selecionados para formar o conjunto de dados de teste, 64% (653) foram selecionadas para fazer parte do conjunto de dados de treino e 16% (164) selecionados como exemplos para validação de treino. Os COGs identificados totalizam 4638, enquanto os eNOGs identificados totalizam 96239.

Também definiu-se o pH neutro, de valor 7, como o valor nulo para a lógica interna dos algoritmos, ou seja zero. Para isso, configurou-se os algoritmos para internamente subtraírem o valor 7 dos valores de pH dos dados de treino. O objetivo dessa definição é facilitar o aprendizado dos algoritmos, ressaltando a diferença entre dados associados a pH alcalinos e ácidos por meio da sua separação, grosseiramente, em duas faixas/grupos: valores positivos associados a pH alcalino e valores negativos associados a pH ácido. Teoricamente, desse modo os algoritmos

aprendem que os grupos ortólogos presentes nos genomas relacionados a pH próximos a 7 são associados como padrão nulo, desse modo associando a esse grupos um valor de influência nulo para as predições. Com isso, ressaltam-se os grupos ortólogos presentes em genomas associados a valores distantes do pH neutro e confere-se a eles maior influência, seja positiva em caso de pH alcalino, ou negativo em caso de pH ácido. Isso também permite uma escalabilidade simples caso repetições de determinados genes nos genomas indique maior capacidade para lidar com valores de pH mais extremos, útil para o caso de modelos treinados com dados utilizando representação quantitativa de grupos ortólogos. Os modelos treinados foram configurados para após realizarem as predições somarem a elas o valor 7, retornando assim a predição à escala padrão de pH.

Na Figura 17 são apresentadas as métricas de avaliação $100 \times R^2$ dos modelos preditivos da PheMaLe-bacteria para as predições dos limites inferiores e superiores de tolerância de bactérias para pH, utilizando o conjunto de dados com as 4 diferentes representações: A) representação binária de presença e ausência de COGs, B) representação binária de presença e ausência de COGs e eNOGs, C) representação de relevância e quantificação de COGs, e D) representação de relevância e quantificação de COGs e eNOGs.

Podemos verificar que o R^2 máximo obtido não ultrapassa o valor de 0,43, indicando que os algoritmos não conseguiram uma boa métrica de predição para pH. Os valores mais altos de R^2 para o limite superior de pH tolerado estavam próximos de 0,40, enquanto os valores mais altos de R^2 para o limite inferior de pH tolerado estavam próximos de 0,29. Estes resultados indicam uma baixa capacidade de predição dos modelos treinados. O algoritmo com melhor métrica para R^2 para predição dos valores superiores da faixa de pH tolerada foi o XGBoost com representação quantitativa de COGs, enquanto o algoritmo com melhor métrica para R^2 para predição dos valores inferiores da faixa de pH tolerada foi o Random Forest com representação binária de COGs e eNOGs.

Muitos algoritmos lineares, como Regressão Linear, Theil-Sen, Elastic Net, Passive Aggressive, apresentaram métricas baixas ou falharam em convergir para uma solução, enquanto algoritmos não lineares como Random Forest, XGBoost e Kernel Ridge com kernel cossenoide apresentaram os melhores resultados. Isso indica que as relações entre os genes nesse fenótipo possivelmente não são aditivas, com seus genes agindo independentemente entre sí.

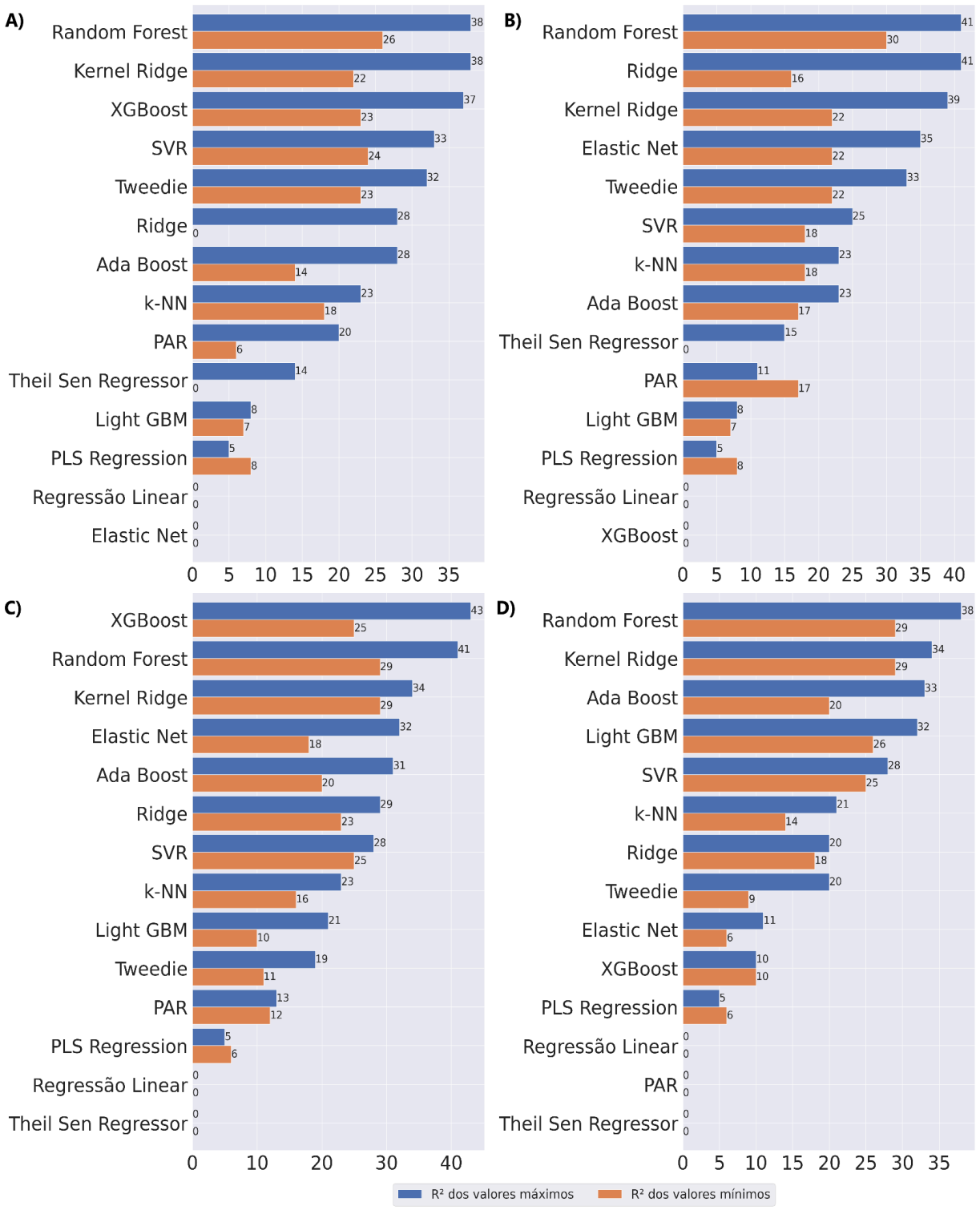

Figura 17: Métrica 100×R² dos modelos de predição de limites inferior e superior de pH sob o conjunto de dados teste. Modelos treinados utilizando A) representação binária de COGs; B) representação binária de COGs e eNOGs; C) representação numérica de COGs de acordo com sua relevância; D) representação numérica de COGs e eNOGs de acordo com sua relevância;

Na Figura 18 apresenta-se um gráfico de comparação entre os valores preditos pelo modelo treinado com XGBoost utilizando representação quantitativa de COGs e os valores reais para o conjunto teste dos valores de limites inferior e superior para pH tolerado por bactérias. A linha cinza corresponde à linha de identidade, ou seja, a linha que mostra onde os pontos preditos estariam localizados se houvesse 100% de acertos em relação aos valores reais registrados no banco de dados. Para comparação com a linha de identidade foi traçada a linha azul, que representa a regressão linear calculada a partir dos valores preditos pelo modelo XGBoost para o limite superior de pH tolerado pelas bactérias. Ela apresenta um coeficiente angular de 0.51 e coeficiente linear de 3.98, indicando que o modelo conseguiu identificar apenas uma parcela dos padrões genéticos relacionados ao limite inferior de pH.

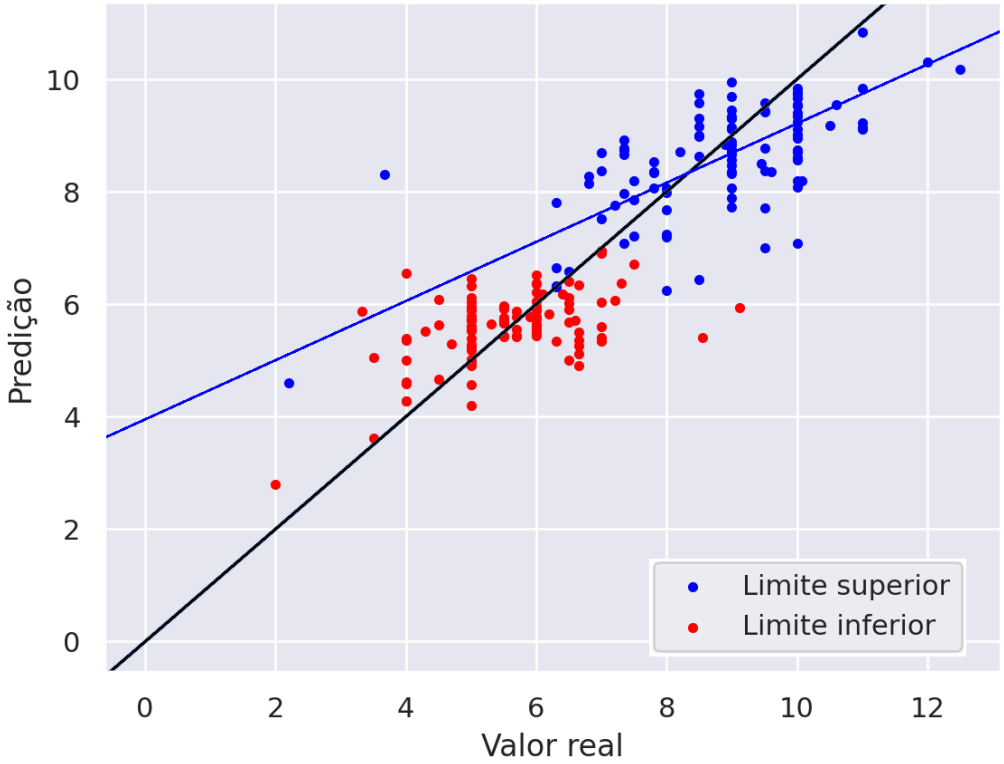

Figura 18: Comparação dos valores reais de limites superior e inferior de pH com valores preditos pelo modelo treinado com XGBoost com representação quantitativa de COGs. A linha negra representa a linha identidade entre valores preditos e valores reais. A linha azul representa a regressão linear calculada a partir dos valores preditos pelo modelo treinado com XGBoost para o limite superior de pH.

A Figura 19 apresenta um gráfico de comparação entre os valores preditos pelo modelo treinado com Random Forest utilizando representação binária de COGs e

eNOGs e os valores reais para o conjunto teste dos valores de limites inferior e superior para pH tolerado por bactérias. Também está presente a linha de identidade, representada pela linha cinza. Da mesma forma, para comparação com a linha de identidade foi traçada a linha vermelha, que representa a regressão linear calculada a partir dos valores preditos pelo modelo treinado com Random Forest para o limite inferior de pH tolerado pelas bactérias. Ela apresenta um coeficiente angular de 0.33 e coeficiente linear de 3.92, indicando que o modelo conseguiu identificar apenas uma parcela pequena dos padrões genéticos relacionados ao limite inferior de pH.

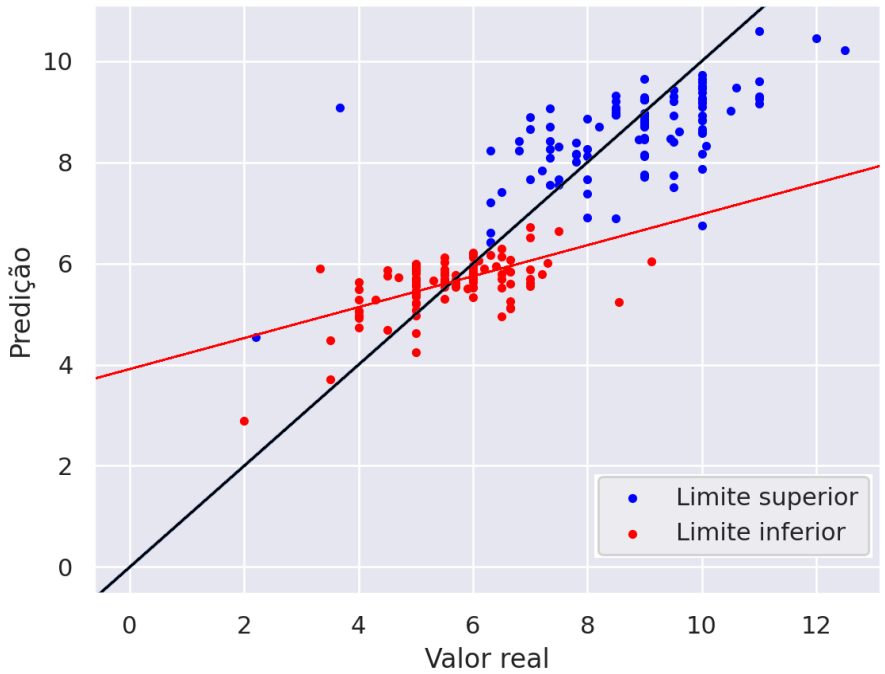

Figura 19: Comparação dos valores reais de limites superior e inferior de pH com valores preditos pelo modelo treinado com Random Forest utilizando representação binária de COGs e eNOGs. A linha negra representa a linha identidade entre valores preditos e valores reais. A linha vermelha representa a regressão linear calculada a partir dos valores preditos do limite inferior de pH.

Na Figura 20 são apresentados os histogramas de valores de distância entre os valores preditos pelos modelos e os valores reais de limites de pH inferior e superior. O desvio padrão do erro do modelo treinado com Random Forest está em 1.02, enquanto o desvio padrão do erro do modelo treinado com XGBoost está em 0.96.

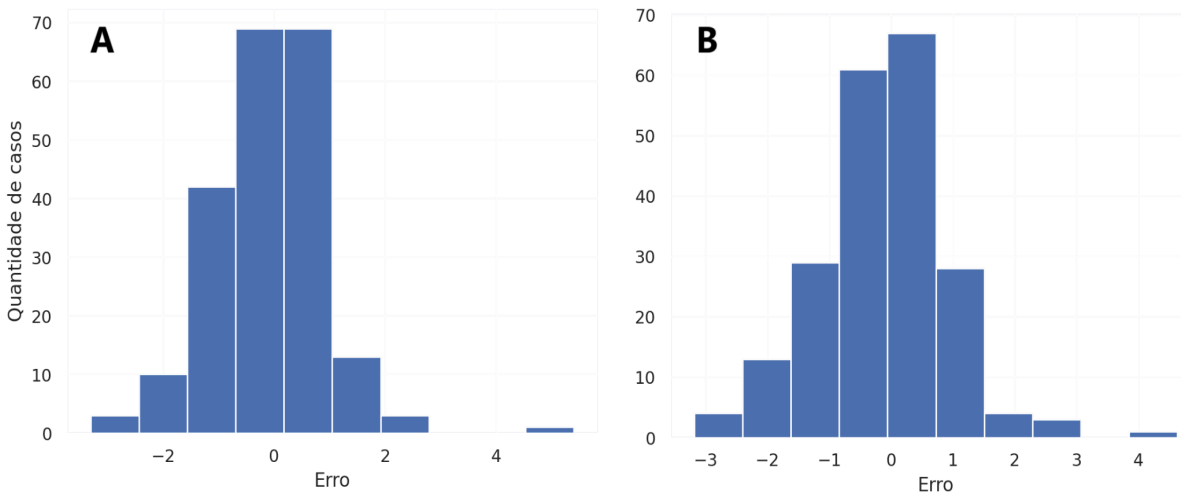

Figura 20: Quantidades de casos por distância do valor entre as predições e valores reais dos limites de pH. Modelos treinados com: A) Random Forest com representação binária de COGs e eNOGs; B) XGBoost com representação quantitativa de COGs.

4.3. Análise de influência dos grupos ortólogos

Apesar da baixa capacidade preditiva dos modelos treinados, os grupos ortólogos identificados como relevantes podem fornecer insights valiosos sobre mecanismos potencialmente associados ou, no mínimo, correlacionados à proteção contra extremos de pH.

A Figura 21 apresenta os 20 grupos ortólogos considerados mais relevantes e seus fatores de influência no modelo treinado com XGBoost utilizando a representação quantitativa de COGs para a predição do limite superior de pH.

É importante esclarecer a interpretação desses fatores de influência no contexto do XGBoost. Diferentemente de modelos treinados com Regressão Linear, onde os valores dos coeficientes impactam diretamente no cálculo da predição final, no XGBoost o fator de influência indica a relevância do parâmetro na tomada de decisão, sem uma relação direta com o cálculo do valor numérico da predição. O XGBoost, assim como o LightGBM e o Random Forest, é um algoritmo baseado em conjuntos de Árvores de Decisão, e neste tipo de abordagem, o fator de influência de um parâmetro reflete a quantidade de vezes que esse parâmetro surge como relevante no conjunto de Árvores, dessa forma ele só pode assumir valores ou nulo ou positivo.

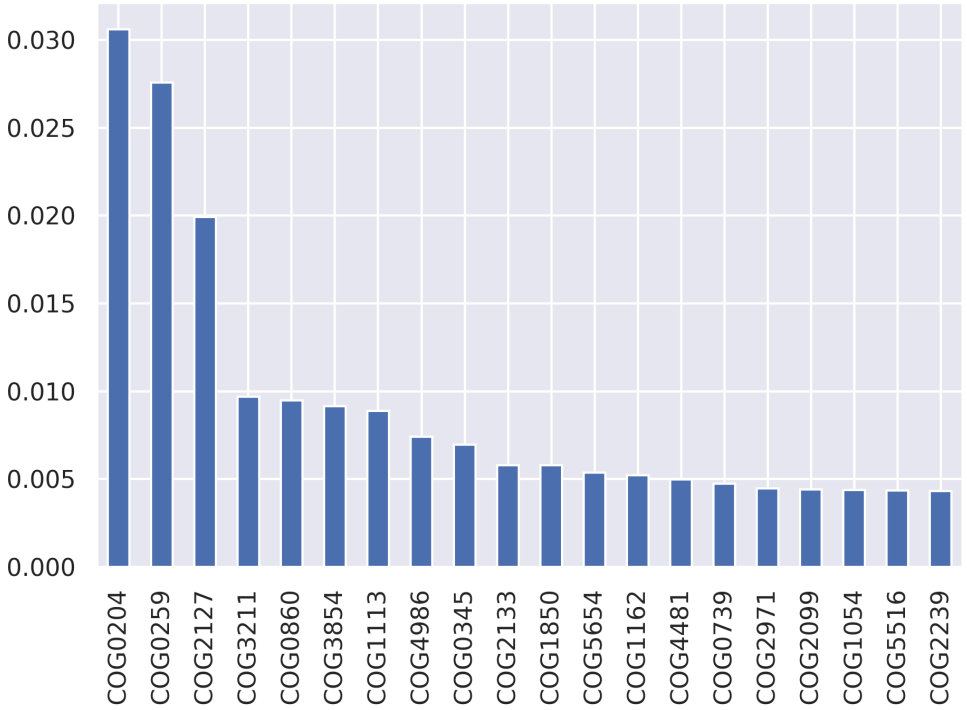

Figura 21: 20 grupos ortólogos mais influentes para predição do limite superior de pH no modelo treinado com XGBoost com representação quantitativa de COGs.

Desses COGs mais influentes uma parcela considerável é relacionada à membrana celular, onde muitos dos mecanismos de proteção contra acidificação ou alcalinização ocorrem, seja por transportadores ativos, pela composição da membrana em si ou por sinalizadores. Outra parcela considerável também está relacionada à proteção contra estresse osmótico e estresse oxidativo. Pode-se ver mais detalhes dos 10 COGs mais influentes no modelo a seguir.

- COG0204 (COG0204 - NCBI, s.d.) é responsável pela 1-acil-sn-glicerol-3-fosfato aciltransferase, uma enzima envolvida na biossíntese de fosfolipídios, componentes essenciais da membrana celular. Ela catalisa a acilação do ácido lisofosfatídico e é responsável pela produção de ácido fosfatídico, um precursor para a síntese de vários glicerofosfolipídios de membrana (Ogawa *et al.,* 2018).

- COG0259 (COG0259 - NCBI, s.d.) responsável pela piridoxina/piridoxamina 5'-fosfato oxidase, ou PdxH. Está envolvido na via de biossíntese do fosfato de piridoxal (forma ativa da vitamina B6), que atua como cofator de várias reações bioquímicas relacionadas ao metabolismo de aminoácidos

(Ankisettypalli *et al.*, 2016), incluindo as que envolvem a modulação de proteínas e enzimas que ajudam na adaptação bacteriana a estresses ambientais, como variações no pH.

- COG2127 (COG2127 - NCBI, s.d.) é responsável pela proteína adaptadora de protease Clp, dependente de ATP, ou ClpS, envolvida na via de degradação de proteínas, por meio da desestabilização de resíduos do *N*-terminal (Schmidt *et al.*, 2009; Queraltó *et al.*, 2023). O sistema Clp, do qual ClpS faz parte, desempenha um papel fundamental na resposta das bactérias a estresses térmico, osmótico e oxidativo (Capestany *et al.*, 2008; Frees *et al.*, 2004). Verificou-se experimentalmente nas bactérias *Salmonella enterica* e *Escherichia coli* que a ClpS é reprimida quando há baixo nível de Mg^{2+} citoplasmático (Yeom *et al.*, 2018).

- COG3211 (COG3211 - NCBI, s.d.; D8TXF6 - UniProt, s.d.) se trata de uma fosfatase alcalina secretada, da família Phox, envolvida na obtenção de fósforo pela bactéria, sendo geralmente sintetizadas em maior quantidade quando há falta do elemento no ambiente (Ragot *et al.*, 2015; Zaheer *et al.*, 2009). As fosfatases alcalinas possuem melhor atividade em pH alcalino (Zaheer *et al.*, 2009), possivelmente também justificando sua presença na faixa superior da predição.

- COG0860 (COG0860 - NCBI, s.d.) responsável pela proteína N-acetilmuramil-L-alanina amidase, associada às enzimas encontradas no periplasma AmiA, AmiB, AmiC e AmiD e também associada a enzima citosólica AmpD. As enzimas AmiA, AmiB, AmiC e AmiD são responsáveis pela degradação da parede celular bacteriana, especialmente envolvidas na na hidrólise de peptidoglicanos (Heidrich *et al.*, 2001). Os peptidoglicanos definem o formato da célula e possuem papel fundamental na proteção contra estresse osmótico (Zhang *et al.*, 2023). A AmpD está envolvida na reciclagem dos fragmentos de peptidoglicano gerados durante o crescimento e divisão bacteriana (Kerff *et al.*, 2009).

- COG3854 (COG3854 - NCBI, s.d.) responsável pela proteína SpoIIIAA, envolvida na transição do estágio III de endoesporulação e relacionada ao controle de proteínas de comunicação intracelular e à regulação de respostas

de estresse. Em alguns estudos, é possível observar que a endoesporulação pode ocorrer como resposta a estresses ambientais, incluindo variações de pH (Lowe *et al.*, 1989; Wetzel & McBride, 2020). Para bactérias do gênero *Bacillus*, por exemplo, a endoesporulação é um mecanismo crucial de sobrevivência para condições ambientais adversas como modificações extremas de pH (Zeytuni *et al.*, 2017).

- COG1113 (COG1113 - NCBI, s.d.) responsável pelo transportador de L-asparagina ou permease relacionada, AnsP. Não foram encontrados estudos relevantes associando este transportador à adaptabilidade de uma bactéria a variações de pH. No entanto, os transportadores podem ter um papel na adaptação a condições de estresse ambiental por manterem a concentração citoplasmática estável (Elston *et al.*, 2023).

- COG4986 (COG4986 - NCBI, s.d.; B8DL56 - UniProt, s.d.) se trata de uma subunidade transmembranar encontrada em transportadores ABC dependentes de proteína de ligação periplasmática. Há registros de um transportador ABC, relacionado a regulação de transporte de K^+, essencial em ambientes alcalinos (Lin *et al.*, 2009).

- COG0345 (COG0345 - NCBI, s.d.) Pirrolina-5-carboxilato redutase que catalisa a síntese de prolina. Este aminoácido é capaz de regular o transporte equivalente redox, resultando na proteção das células contra o estresse oxidativo (Xiao *et al.*, 2020). Foi encontrada também relação com a proteção a estresse osmótico na bactéria *Bacillus subtilis* (Morawska *et al.*, 2022).

- COG2133 (COG2133 - NCBI, s.d.) responsável pela glicose/arabinose desidrogenase, com dobra da hélice beta. Apesar de não ter sido encontrado um estudo conclusivo sobre o COG em questão e a faixa de pH tolerado, desidrogenases no geral, são enzimas que atuam em processos metabólicos variados da célula. A glicose desidrogenase oxida a glicose em ácido glucônico, assumindo um papel regulatório e bioenergético para a célula, uma vez que os prótons gerados pela oxidação apoiam a força próton motriz da membrana necessária para obtenção de compostos exógenos. Isso a torna importante para sobrevivência em situações diversas, incluindo em ambientes com pouco fósforo (Tripura *et al.*, 2007).

Na Figura 22 são apresentados os 10 COGs considerados mais relevantes pelo modelo treinado com Random Forest para predição do limite inferior de pH utilizando representação binária de COGs e eNOGs e os seus fatores de influência. Percebe-se que uma parcela considerável dos COGs mais influentes estão relacionados a formação de biofilmes e a resistência antimicrobiana.

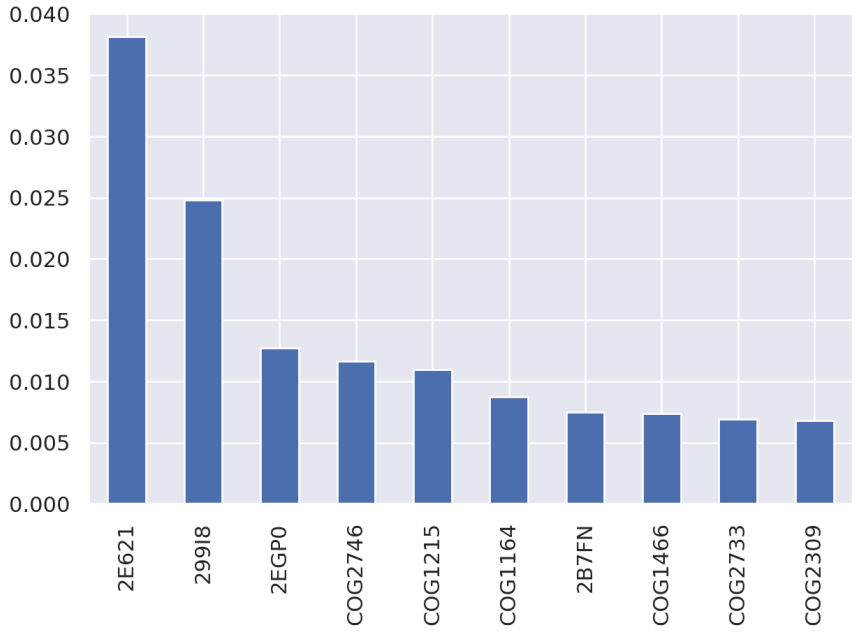

Figura 22: 10 grupos ortólogos mais influentes para predição do limite inferior de pH no modelo treinado com Random Forest com representação binária de COGs e eNOGs.

Mais detalhes dos COGs são apresentados a seguir.

- COG2746 (COG2746 - NCBI, s.d.) responsável pelo aminoglicosídeo N3'-acetiltransferase, para resistência antimicrobiana (Vliegenthart *et al.*, 1991).

- COG1215 (COG1215 - NCBI, s.d.) responsável pela glicosiltransferase, componente chave do peptidoglicano da parede celular bacteriana. Age como sinalizador que promove a formação de biofilme e regula o processo de endoesporulação em ambientes com baixa quantidade de nutrientes (Konopka, 2012).

- COG1164 (COG1164 - NCBI, s.d.; P54124 - UniProt, s.d.) responsável pela Oligoendopeptidase F, um inibidor de endoesporulação que hidrolisa

peptídeos que contém entre 7 e 17 aminoácidos, possuindo especificidade bastante ampla (Kanamaru *et al.*, 2002).

- COG1466 (COG1466 - NCBI, s.d.; P28630 - UniProt, s.d.) associado à proteína DNA polimerase III subunidade delta, uma enzima complexa, formada por múltiplas cadeias, que é responsável pela maior parte da síntese replicativa em bactérias (Jeruzalmi *et al.*, 2001).
- COG2733 (COG2733 - NCBI, s.d.) responsável pela proteína não caracterizada YjiN, cuja função prevista indica semelhança com parte de um sistema transportador localizado na membrana bacteriana interna (Krizsan *et al.*, 2015).
- COG2309 (COG2309 - NCBI, s.d.) responsável pela leucil aminopeptidase, associada à formação de biofilmes e à resistência a antibióticos (Singh *et al.*, 2012).

4.4. Conclusão

Os modelos treinados apresentaram baixa capacidade de predição das faixas de pH toleradas. As métricas de avaliação apresentaram 0,30 para o R^2 nas predições para o limite inferior de pH e 0,43 para o R^2 nas predições para o limite superior de pH. Os coeficientes angulares calculados para a relação entre valores preditos e valores reais indicam que os modelos conseguiram detectar alguns padrões relacionados às faixas de pH de forma correta, mas ainda não o suficiente para predições com métricas satisfatórias.

Os motivos para esse problema podem ser diversos, desde as escolhas de representação simbólicas não expressarem claramente a diferença entre as bactérias com diferentes faixas de pH toleradas, tornando necessário algum outro tipo de representação simbólica dos genomas, ou uma mera incapacidade dos algoritmos de aprendizado de máquina em modelar os padrões genômicos para esse fenótipo. Adicionalmente, uma baixa diversidade de padrões genéticos no conjunto de dados, ou até mesmo a incompletude dos registros no banco de dados BacDive podem também ter afetado a competência dos modelos na predição relacionada a esse fenótipo. Possivelmente uma quantidade considerável dos dados catalogados para

os limites superior e inferior de pH se trata apenas de valores referentes aos ambiente de coleta, e não foi realizada uma experimentação extensiva para determinação dos limites verdadeiros de pH tolerados,

No entanto, estes resultados ainda representam um avanço no estado da arte dentro desse escopo dado ser o primeiro trabalho com o intuito de predizer as faixas de tolerância de pH. Ademais, os modelos do PheMaLe-bacteria ainda identificaram múltiplos COGs que, mesmo de forma indireta, possivelmente influenciam a faixa de pH tolerado e a capacidade de sobrevivência de bactérias em um ambiente de pH variável ou extremo. Isso é evidenciado pelas categorias funcionais aos quais muitos desses genes pertencem identificadas pela literatura como relacionadas a capacidade de tolerância a extremos de pH (Cortez *et al.*, 2022), destacando-se: composição da membrana celular bacteriana, sinalizadores celulares, replicação de DNA, reparo, transporte de aminoácidos e tráfego intracelular, categorias relacionadas à proteção da célula contra estresses ambientais associados a mudanças de pH.

5. Endoesporulação

A endoesporulação, também conhecida como esporulação bacteriana, é um dos mais notáveis exemplos de adaptação de bactérias, conferindo a elas uma incrível capacidade de sobrevivência não apenas à passagem do tempo, podendo resistir por milhares de anos, mas também a condições ambientais letais, como escassez de nutrientes, alterações extremas de temperatura, desidratação, radiação e exposição a agentes químicos antibióticos (Cano & Borucki, 1995).

O processo de endoesporulação é desencadeado por estresses ambientais onde a bactéria entra em um estado de dormência e forma uma estrutura altamente resistente conhecida como endósporo (Atrih & Foster, 2002). Isso auxilia não apenas na sobrevivência das bactérias como também na dispersão bacteriana pelos diversos ecossistemas, se propagando até encontrarem condições favoráveis (Atrih & Foster, 2002; Zeigler & Nicholson, 2017).

Durante o desenvolvimento de endósporos ocorre uma reconfiguração completa do envelope celular e do metabolismo bacteriano. O processo abrange uma sequência de eventos morfológicos e bioquímicos que dura várias horas, requerendo uma coordenação espacial e temporal altamente regulada para a expressão de centenas de genes estrutural e funcionalmente diversos, como os relacionados a: sistemas de transdução de sinal, fatores sigma, reguladores transcricionais, enzimas metabólicas e proteínas estruturais do revestimento dos endósporos (Hilbert & Piggot, 2004). Essa reconfiguração envolve a formação de várias camadas protetoras, incluindo o exospório, que é um revestimento delgado externo; a capa do endósporo, que possui características bioquímicas que conferem proteção a várias moléculas tóxicas e agentes físicos, além de possivelmente conter enzimas envolvidas na germinação; e o córtex, que é formado por camadas de peptidoglicano e confere proteção térmica. Todas estas camadas atuam como barreiras físicas e químicas, para proteger os componentes bacterianos essenciais contra danos (Hoch, 1976). Enquanto isso, o núcleo do endósporo desenvolve mecanismos de estabilização do código genético, como a desidratação extrema que interrompe atividades metabólicas e confere proteção contra desnaturação térmica e oxidação (Atrih & Foster, 2002).

Outro aspecto essencial vinculado a endoesporulação é o processo de germinação, em que a bactéria retorna a um estado metabolicamente ativo. Isso ocorre quando o endósporo detecta sinais de condições ambientais favoráveis, como a presença de nutrientes essenciais, temperatura e pH adequados (Atrih & Foster, 2002). Nesse momento a permeabilidade da parede do esporo aumenta permitindo a entrada de água e enzimas hidrolíticas começam a degradar a camada externa do esporo, iniciando sua reativação metabólica (Moir, 2006).

A resistência dos endósporos é uma área de estudo extremamente relevante para aplicações humanas, principalmente na área da saúde e alimentícia. No contexto clínico, os endósporos podem sobreviver aos processos de descontaminação em ambientes hospitalares e causar doenças severas em humanos. Por exemplo, endósporos de *Clostridium difficile*, *Clostridium tetani* e *Bacillus anthracis* são agentes infecciosos resistentes que podem sobreviver a processos de descontaminação em hospitais e voltar a contaminar humanos (Tijerina-Rodriguez *et al.*, 2018; Tijerina-Rodriguez *et al.*, 2019; Romero-Rodríguez *et al.*, 2023). Na indústria alimentícia endósporos de *Bacillus cereus* e *Clostridium botulinum*, por exemplo, podem causar intoxicações alimentares graves devido à produção de toxinas e diminuir drasticamente a viabilidade dos alimentos (Romero-Rodríguez *et al.*, 2023; Griffiths & Schraft, 2017; Schmid *et al.*, 2024). Os endósporos são um dos principais desafios nos processos de esterilização, requerendo tratamentos específicos, como altas temperaturas, irradiação ou agentes químicos potentes (Romero-Rodríguez *et al.*, 2023).

Por outro lado, a endosporulação se trata de uma habilidade muito útil na área biotecnológica para produção de probióticos, pesticidas biológicos e outros produtos bacterianos devido à sua durabilidade para o armazenamento e transporte. Estes mecanismos foram estudados principalmente nas bactérias dos gêneros *Bacillus* e *Clostridium* devido à sua relevância em contextos médicos, industriais e ambientais. Grande parte do conhecimento existente na literatura sobre endósporos e seus mecanismos é referente a esses gêneros.

Em suma, os mecanismos genéticos envolvidos no processo de endoesporulação são complexos e seu estudo tem implicações importantes para o controle de infecções, a melhoria de processos industriais e a compreensão da persistência microbiana em ambientes extremos.

5.1. Informações de espécies bacterianas e sua capacidade de endoesporulação

Os dados para capacidade de endoesporulação em espécies de bactérias foram coletados do catálogo compilado por Madin *et al.* (2020), totalizando registros referentes a 5752 bactérias. Desses registros, o NCBI possui genomas completos para 1450 espécies. O conjunto de dados obtido apresenta uma diversidade bacteriana de 92 ordens, 186 famílias e 489 gêneros. Na Figura 23 está representada a distribuição percentual das ordens presentes e a quantidade de representantes em cada ordem. A ordem com maior quantidade de representantes é a *Bacillales*, a ordem contendo as do filo Firmicutes, as bactérias mais estudadas no tópico de endoesporulação. As ordens com no máximo 10 representantes são apresentadas em detalhes na tabela presente no Apêndice E.

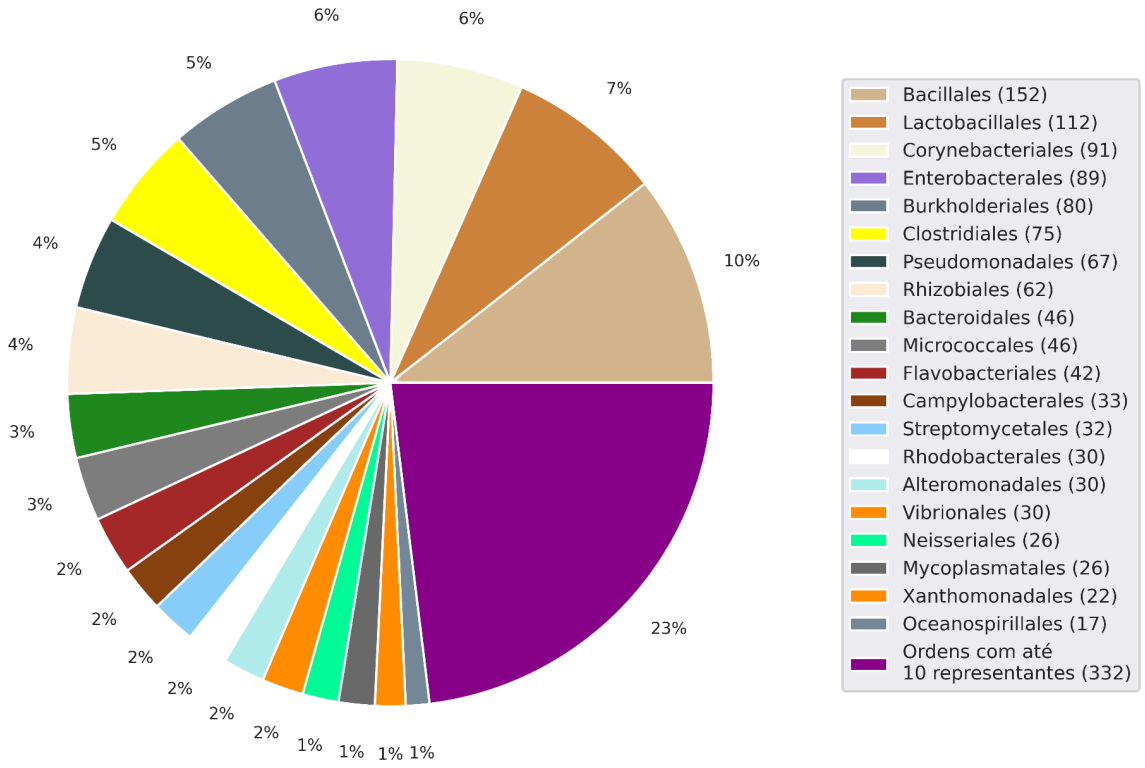

Figura 23: Distribuição de ordens no conjunto de dados para capacidade de endoesporulação.

A Figura 24 apresenta a distribuição fenotípica de capacidade de endoesporulação para os dados coletados, se tratando de apenas 16,5% do total de dados, um desbalanço representativo do encontrado na natureza

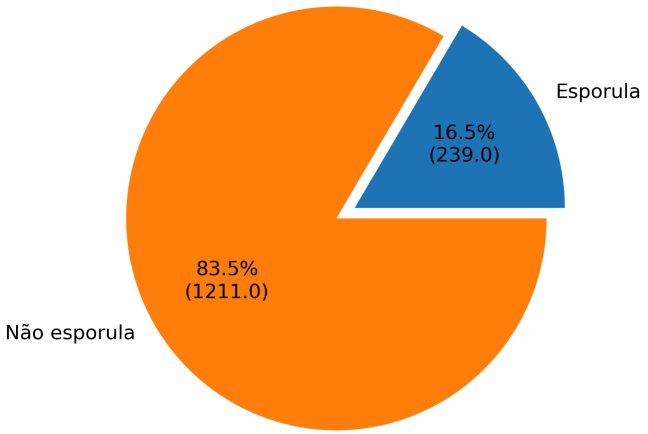

Figura 24: Quantidades de genomas completos no conjunto de dados para capacidade de endoesporulação.

Na Figura 25 estão os gráficos de EMD das distâncias entre os genomas do conjunto de dados. Percebe-se uma distribuição homogênea dos genomas pelo espaço de parâmetros dos gráficos e não é possível distinguir uma separação entre os genomas das bactérias capazes de endoesporulação das incapazes.

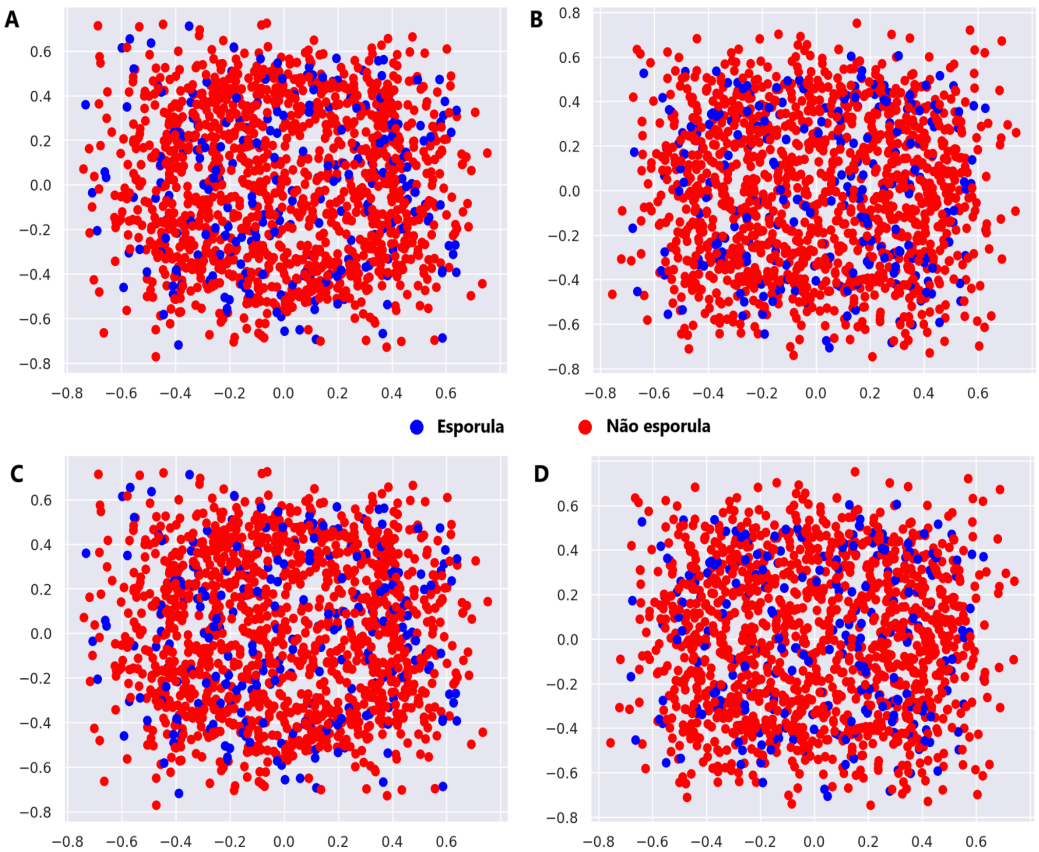

Figura 25: EMD dos genomas com distinção sobre capacidade de endoesporulação. Genomas utilizando representações: A) binária de COGs; B) binária de COGs e eNOGs; C) quantificação de COGs; D) quantificação de COGs e eNOGs;

5.2. Resultados da PheMaLe-bacteria

O conjunto de dados, por meio de seleção randômica, foi separado em 3 grupos: 64% dos exemplares do conjunto de dados, ou seja 928 espécies foram selecionadas para formar o grupo de dados de treino; 16%, ou seja 232 espécies selecionadas como exemplos para validação de treino e 20% do conjunto de dados, 290 espécies, foi selecionado para formar o conjunto de dados de teste. Os COGs identificados totalizam 4647, enquanto os eNOGs identificados totalizam 122321.

Na Figura 26 são apresentadas as métricas de avaliação F1, AUC-ROC e AUC-PR dos modelos preditivos da PheMaLe-bacteria no conjunto de dados com as 4 diferentes representações: A) representação binária de presença e ausência de COGs, B) representação binária de presença e ausência de COGs e eNOGs, C) representação de relevância e quantificação de COGs, e D) representação de relevância e quantificação de COGs e eNOGs.

É possível notar que os modelos treinados utilizando a representação binária de grupos ortólogos curados e hipotéticos apresentaram uma performance melhor em comparação com os outros grupos. O modelo treinado com LightGBM e usando a representação binária de COGs e eNOGs se mostrou o melhor em todas as métricas de avaliação, as quais podem ser verificadas na Figura 27, onde é apresentada sua Precisão, Recall e a matriz de confusão para o conjunto de teste. Suas curvas de AUC-ROC e AUC-PR dos treinos podem ser encontradas no Apêndice F.

Dos 290 exemplos do conjunto de dados de teste podemos ver na Figura 27 as quantidades de predições corretas e erradas realizadas pelo modelo. As predições corretas se enquadram entre os negativos verdadeiros, ou seja, bactérias incapazes de endoesporulação que foram preditas corretamente como incapazes, e os positivos verdadeiros, bactérias capazes de endoesporulação que foram preditas corretamente como capazes. Nas predições erradas se enquadram os falsos positivos, ou seja, bactérias incapazes que foram preditas erroneamente como capazes, e os falsos negativos, composto por bactérias capazes preditas erroneamente como incapazes.

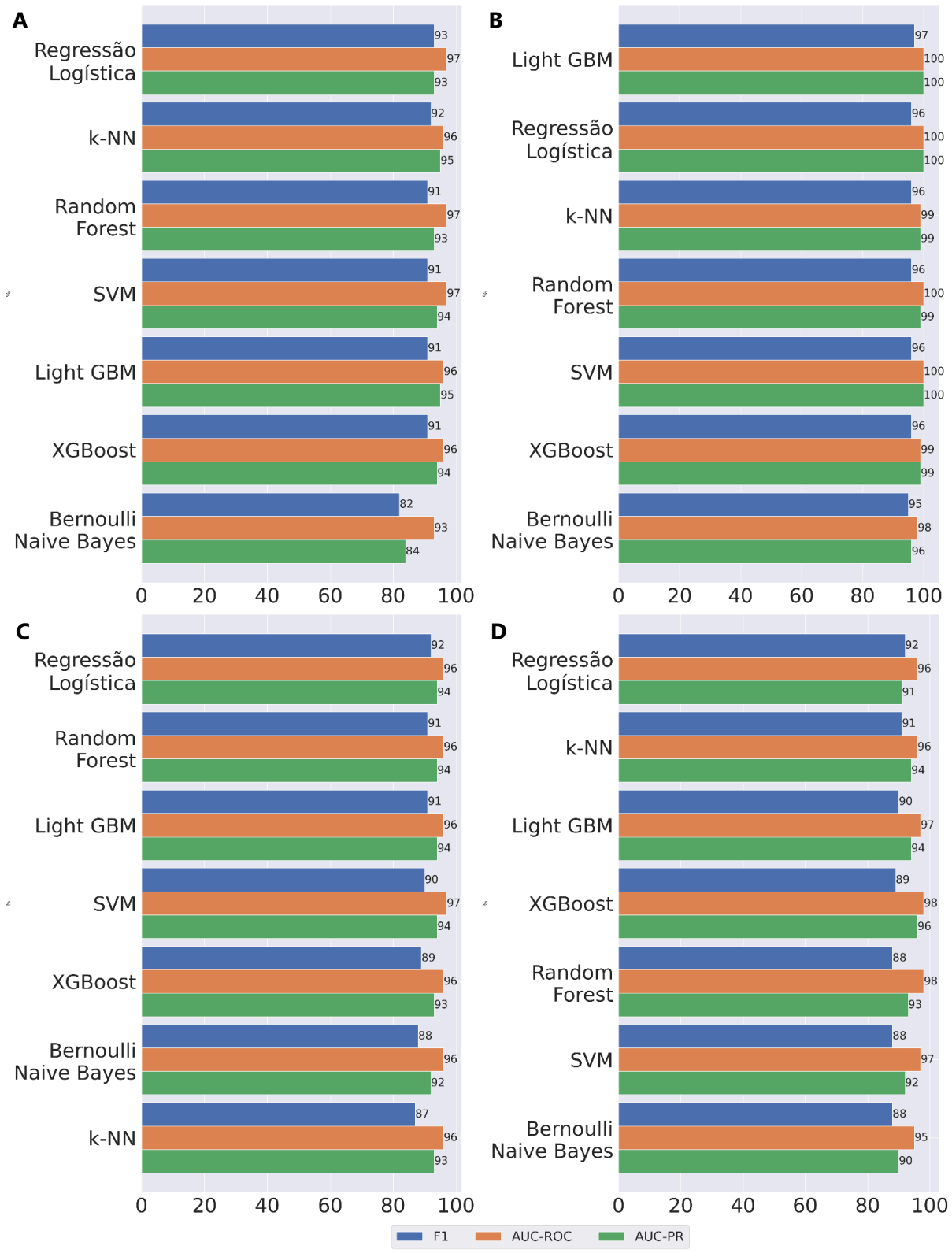

Figura 26: Métricas F1, AUC-ROC e AUC-PR dos modelos de predição de capacidade de endoesporulação sob o conjunto de dados teste. Modelos utilizando representação: A) binária de COGs; B) binária de COGs e eNOGs; C) quantitativa de COGs; D) quantitativa de COGs e eNOGs;

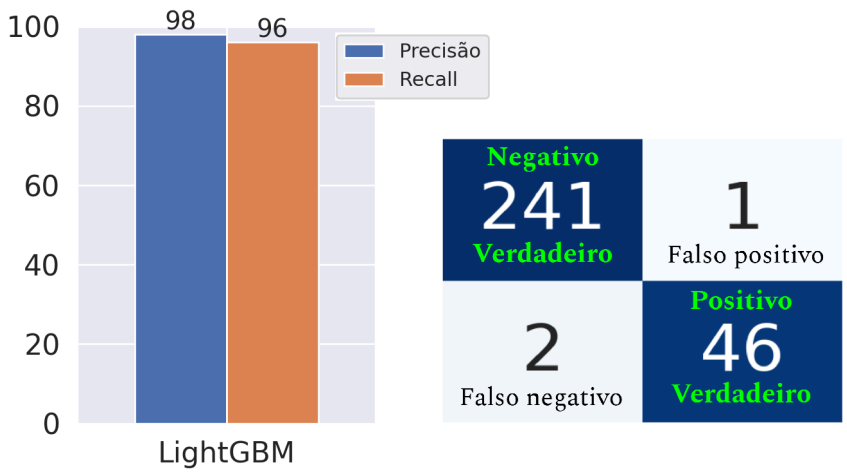

Figura 27: Precisão, Recall e F1 do modelo teinado com LightGBM e matriz de confusão com o conjunto teste de endoesporulação.

Na Tabela 2 estão as espécies do conjunto de testes que foram preditas erroneamente pelo modelo.

Tabela 2: Espécies preditas incorretamente para capacidade de endoesporulação pelo modelo treinado com LightGBM e representação binária de COGs. e eNOGs.

Espécie	Resultado	Observações
Coprococcus catus	Falso Positivo	Assumiu-se que os membros do gênero *Coprococcus* não formam esporos, no entanto essa afirmação foi contestada em estudos recentes (Browne *et al.*, 2016; Notting *et al.*, 2023). Mas ainda não há confirmação experimental de capacidade de esporulação na espécie em questão.
Thomasclavelia ramosa	Falso Negativo	Foi comprovado experimentalmente que a espécie é capaz de esporulação (Wasfy et al., 2024).
Flavonifractor plautii	Falso Negativo	Foi comprovado experimentalmente que a espécie é capaz de esporulação (Saito et al., 2024).

Os resultados destes Falsos Negativos podem ser corrigidos alterando o limiar padrão de 0.5 de decisão do modelo. Podemos ver pelos gráficos da curva de

precisão x recall e curva ROC apresentados na Figura 28 que há limiares que podem alcançar melhores resultados. No entanto a determinação exata deste limiar foge ao escopo deste trabalho, e fica a cargo de um trabalho futuro.

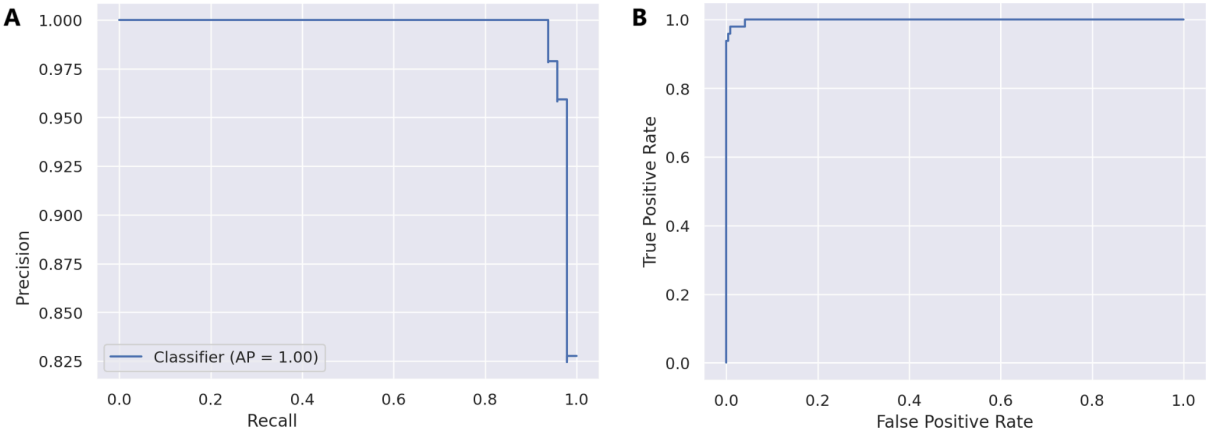

Figura 28: Métricas do treino do modelo treinado com LightGBM utilizando representação binária de COGs e eNOGs para classificação de capacidade de endoesporulação. A) curva de precisão x recall; B) curva ROC.

5.3. Análise de influência dos grupos ortólogos

Na Figura 29 são apresentados os 20 grupos ortólogos considerados mais relevantes e seus fatores de influência no modelo treinado com LightGBM e representação binária de COGs e eNOGs.

Nota-se uma grande quantidade de grupos ortólogos hipotéticos entre os grupos mais relevantes ao modelo, compreendendo 10 dos 20 grupos . Deve-se prestar especial atenção aos 3 grupos ortólogos mais influentes no modelo, todos hipotéticos e apresentando fatores de influência maiores do que o dobro do valor dos outros grupos ortólogos, em particular o grupo 2ADZG, que apresenta o maior fator, acima de 2,5 vezes o valor do fator de influência do segundo grupo ortólogo mais relevante. Dos grupos ortólogos curados verificou-se que a maioria são diretamente relacionados à atividade de endoesporulação, com mais detalhes após a Figura 29.

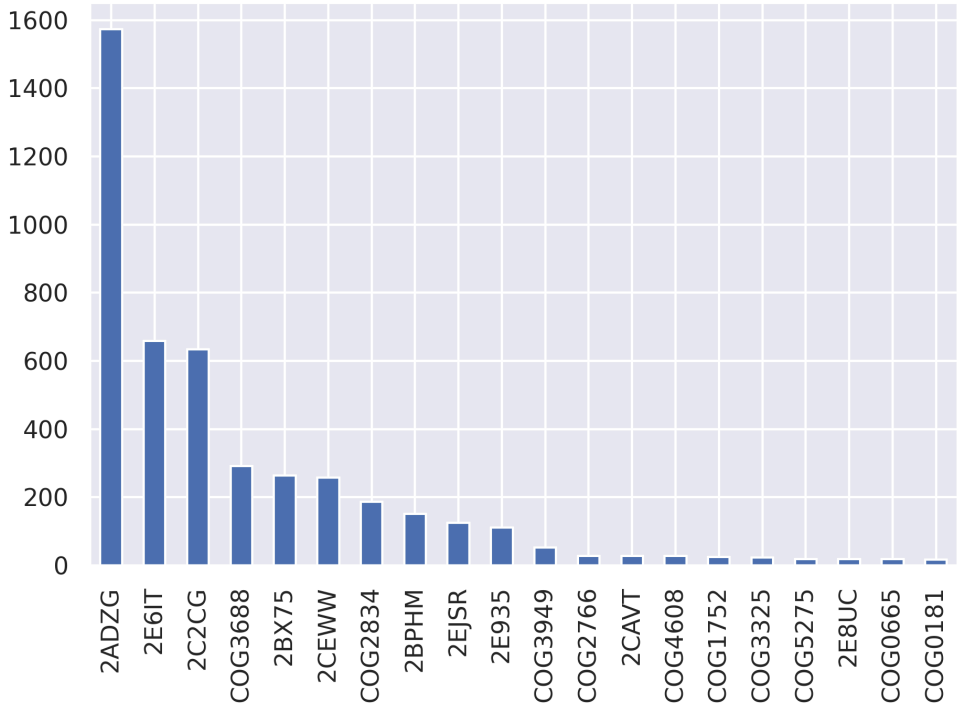

Figura 29: 20 grupos ortólogos mais influentes para predição de capacidade de endoesporulação no modelo treinado com LightGBM utilizando representação binária de COGs e eNOGs para predição de capacidade de endoesporulação.

- COG3688 (COG3688 - NCBI, s.d.) EndoRNase envolvida no decaimento do mRNA. 8 RNAs regulados positivamente são membros dos regulons AbrB e/ou SigW. O AbrB codifica um repressor envolvido na regulação de processos induzidos pela fome, como casos de endoesporulação, enquanto SigW está envolvido na resposta a vários tipos de estresse do envelope em *B. subtilis* (Leroy *et al.*, 2017).

- COG2834 (COG2834 - NCBI, s.d.) responsável pela proteína LolA que realiza transporte de lipoproteínas através do periplasma de bactérias Gram-negativas durante a montagem da membrana externa. Foi verificado que uma proteína com homologia estrutural com LolA, a GerS, está relacionada à endoesporulação (Feliciano *et al.*, 2021). Não se encontraram estudos relacionando diretamente a proteína LolA e endoesporulação.

- COG3949 (COG3949 - NCBI, s.d.) proteína de membrana não caracterizada YkvI. Verificou-se que esta proteína tem relação direta com endoesporulação

na espécie *Bacillus subtilis* (Meeske *et al.*, 2016) e diversos outros membros do filo *Firmicutes* (Galperin *et al.*, 2022).

- COG2766 (COG2766 - NCBI, s.d.) recentemente descoberto, se trata de uma protease dependente de ATP, envolvida naregulação da endoesporulação em *Bacillus subtilis* (Zhang *et al.*, 2022).
- COG4608 (COG4608 - NCBI, s.d.) proteína componente de oligopeptídeo permease, participante de um sistema de transporte dependente de proteína de ligação essencial para endoesporulação (Koide & Hoch, 1994).
- COG1752 (COG1752 - NCBI, s.d.) responsável por acilesterase/fosfolipase.
- COG3325 (COG3325 - NCBI, s.d.) responsável pela quitinase que influencia a síntese de toxinas cristalinas e a endoesporulação de bactérias (Barboza-Corona *et al.*, 2009).

5.4. Conclusão

O PheMaLe-bacteria obteve altas métricas de avaliação para predição de capacidade de endoesporulação, indicando um bom aprendizado dos múltiplos padrões genéticos relacionados ao fenótipo, mesmo com um conjunto de dados altamente desbalanceado.

Os modelos treinados utilizando a representação binária de COGs e eNOGs apresentaram os melhores resultados nas métricas de avaliação, e diversos eNOGs figuram entre os grupos ortólogos mais influentes nos modelos treinados, indicando a existência de grupos ortólogos hipotéticos com considerável influência sobre a capacidade de endoesporulação. Com uma precisão de 98% e recall de 96%, o modelo treinado com algoritmo LightGBM e representação binária de COGs e eNOGs errou na classificação de apenas 3 bactérias em um total de 290 bactérias do conjunto de dados de teste. Uma predição considerada errada está possivelmente correta, a depender de confirmação experimental.

Diversos dos grupos ortólogos curados considerados relevantes pelo modelo foram identificados pela literatura como positivamente relacionados a mecanismos envolvidos na endoesporulação. Além disso, múltiplos grupos ortólogos hipotéticos

também foram considerados pertinentes, indicando uma possibilidade de envolvimento biológico com o fenótipo em questão.

Por motivos desconhecidos, os modelos treinados com representação quantitativa de COGs e eNOGs não apresentaram métricas de avaliação boas, o que demonstra que uma representação binária de presença e ausência de grupos ortólogos é mais significativa para os algoritmos de aprendizado.

6. Conclusão

Este trabalho teve por objetivo testar a capacidade de aprendizado de múltiplos algoritmos de aprendizado de máquina na identificação de padrões genéticos relacionados a fenótipos bacterianos geneticamente complexos. Foi desenvolvida uma ferramenta denominada PheMaLe-bacteria que engloba 21 algoritmos de aprendizado de máquina diferentes para o treinamento de modelos preditores de fenótipos e realiza uma análise de quais parâmetros genéticos foram identificados como correlacionados aos fenótipos treinados. Para avaliar e validar esta ferramenta, foram feitos testes com conjuntos de dados públicos de espécies bacterianas e informações de seus fenótipos. Foram testados os fenótipos de capacidade de fixação de nitrogênio, faixa de pH tolerada e capacidade de endoesporulação, todos considerados de caráter genético complexo.

Muitos trabalhos da literatura com este mesmo tópico de aplicação de aprendizado de máquina à predição de fenótipos bacterianos expuseram as informações gênicas como uma lista de presença ou ausência de grupos funcionais nos genomas, ou seja, utilizaram apenas uma representação binária dos grupos presentes em cada genoma. O PheMaLe-bacteria envidenciou o impacto de diferentes representações simbólicas gênicas, binárias e quantitativas, no aprendizado. O modelo de predição de capacidade de fixação de nitrogênio que utilizou a representação quantitativa obteve resultados melhores que o modelo que utilizou representação binária, e ainda foi capaz de identificar fenótipos erroneamente associados a espécies bacterianas nos bancos de dados utilizados, enquanto o modelo que utilizou representação binária demonstrou-se inapto para detectar esses mesmos erros.

Os modelos treinados também foram capazes de identificar diversos grupos funcionais relevantes, cuja relação com os fenótipos de interesse foi corroborada por evidências da literatura científica. Esse resultado destaca a utilidade da ferramenta não apenas para a predição de fenótipos, mas também para a identificação de possíveis associações entre grupos funcionais e determinadas características biológicas, incluindo aqueles ainda não plenamente investigados. Além disso, a inclusão de grupos funcionais hipotéticos nos conjuntos de parâmetros revelou que múltiplos modelos preditivos identificaram esses grupos como altamente

correlacionados aos fenótipos analisados. Isso sugere a possibilidade de que genes pouco caracterizados desempenhem papéis importantes nesses processos, contribuindo para a ampliação do conhecimento sobre suas funções biológicas.

A aplicação do PheMaLe-bacteria em diferentes conjuntos de dados genômicos, incluindo aqueles com alto grau de desequilíbrio na quantidade de exemplares, evidenciou a capacidade da ferramenta de se adaptar a diversas espécies bacterianas. Essa flexibilidade torna a abordagem promissora para o estudo de microbiomas complexos e para o desenvolvimento de estratégias direcionadas na biotecnologia, saúde, agricultura e pesquisa científica.

A predição para faixa de pH tolerada ainda se mantém um desafio, com os modelos treinados apresentando uma baixa capacidade de predição. Todavia, os resultados obtidos ainda representam um avanço no estado da arte em relação à predição da faixa de pH tolerado. Os resultados obtidos na aplicação a predição de capacidade de fixação de nitrogênio e de endoesporulação consolidam o avanço neste tópico. Entretanto, desafios permanecem, como a necessidade de bases de dados mais completas e diversificadas, bem como o contínuo aprimoramento dos modelos de aprendizado de máquina para lidar com dados ruidosos e desequilibrados. A inclusão de informações adicionais, como dados fenotípicos experimentais e metadados ecológicos, podem ampliar a precisão e a aplicabilidade da ferramenta.

Em resumo, este trabalho reforça a relevância do aprendizado de máquina como uma abordagem útil para decifrar a complexidade genômica e prever características bacterianas com alto grau de confiabilidade. Avanços futuros nesta linha de pesquisa têm o potencial de transformar nossa compreensão sobre a diversidade bacteriana e suas aplicações em múltiplas áreas, promovendo inovações que beneficiem tanto a ciência quanto a sociedade.

7. Referências

Albertsen, M., Hugenholtz, P., Skarshewski, A., Kåre Lehmann Nielsen, Tyson, G. W., & Per Halkjær Nielsen. (2013). Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes. *Nature Biotechnology*, *31*(6), 533–538. https://doi.org/10.1038/nbt.2579

Ambrosio, R., Ortiz-Marquez, J. C. F., & Curatti, L. (2017). Metabolic engineering of a diazotrophic bacterium improves ammonium release and biofertilization of plants and microalgae. *Metabolic Engineering*, *40*, 59–68. https://doi.org/10.1016/j.ymben.2017.01.002

Ankisettypalli, K., Cheng, J. J., Baker, E. N., & Bashiri, G. (2016). PdxH proteins of mycobacteria are typical members of the classical pyridoxine/pyridoxamine 5′-phosphate oxidase family. *FEBS Letters*, *590*(4), 453–460. https://doi.org/10.1002/1873-3468.12080

Atrih, A., & Foster, S. J. (2002). Bacterial endospores the ultimate survivors. *International Dairy Journal*, 12(2-3), 217–223. https://doi.org/10.1016/s0958-6946(01)00157-1

Baldani, J. I., Pot, B., Kirchhof, G., Falsen, E., Baldani, V. L. D., Olivares, F. L., Hoste, B., Kersters, K., Hartmann, A., Gillis, M., & Döbereiner, J. (1996). Emended Description of *Herbaspirillum*; Inclusion of [*Pseudomonas*] *rubrisubalbicans*, a Mild Plant Pathogen, as *Herbaspirillum rubrisubalbicans comb. nov.*; and Classification of a Group of Clinical Isolates (EF Group 1) as *Herbaspirillum* Species 3. *International Journal of Systematic Bacteriology*, 46(3), 802–810. https://doi.org/10.1099/00207713-46-3-802

Barboza-Corona, J. E., Ortiz-Rodríguez, T., de la Fuente-Salcido, N., Bideshi, D. K., Ibarra, J. E., & Salcedo-Hernández, R. (2009). Hyperproduction of chitinase influences crystal toxin synthesis and sporulation of *Bacillus thuringiensis*. Antonie van Leeuwenhoek, 96(1), 31–42. https://doi.org/10.1007/s10482-009-9332-9

Bennett, E. M., Murray, J. W., & Isalan, M. (2023). Engineering Nitrogenases for Synthetic Nitrogen Fixation: From Pathway Engineering to Directed Evolution. *BioDesign Research*, *5*. https://doi.org/10.34133/bdr.0005

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42. https://doi.org/10.1093/nar/gks1195

Berrabah, F., Ratet, P., & Gourion, B. (2015). Multiple steps control immunity during the intracellular accommodation of *rhizobia*. *Journal of Experimental Botany*, *66*(7), 1977–1985. https://doi.org/10.1093/jxb/eru545

Boyd, E. S., & Peters, J. W. (2013). New insights into the evolutionary history of biological nitrogen fixation. *Frontiers in Microbiology*, 4. https://doi.org/10.3389/fmicb.2013.00201

Browne, H. P., Forster, S. C., Anonye, B. O., Kumar, N., Neville, B. A., Stares, M. D., … Lawley, T. D. (2016). Culturing of "unculturable" Human Microbiota Reveals Novel Taxa and Extensive Sporulation. *Nature*, 533(7604), 543–546. https://doi.org/10.1038/nature17645

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics*, *10*(1), 421.

Cano, R. J., & Borucki, M. K. (1995). Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. *Science* (New York, N.Y.), 268(5213), 1060–1064. https://doi.org/10.1126/science.7538699

Capestany, C. A., Tribble, G. D., Maeda, K., Demuth, D. R., & Lamont, R. J. (2008). Role of the Clp System in Stress Tolerance, Biofilm Formation, and Intracellular Invasion in *Porphyromonas gingivalis*. *Journal of Bacteriology*, *190*(4), 1436–1446. https://doi.org/10.1128/jb.01632-07

Chamrád, I., Simerský, R., Lenobel, R., & Novák, O. (2024). Exploring affinity chromatography in proteomics: A comprehensive review. *Analytica Chimica Acta*, *1306*, 342513–342513. https://doi.org/10.1016/j.aca.2024.342513

Chen, J.-S. (2006). Nitrogen Fixation in the *Clostridia*. *Kluwer Academic Publishers EBooks*, 53–64. https://doi.org/10.1007/1-4020-2179-8_3

Chen, Y., Lin, Y., Zhu, J., Zhou, J., Lin, H., Fu, Y., & Zhou, Y. (2024). Transcriptomic analysis of nitrogen metabolism pathways in *Klebsiella aerogenes* under nitrogen-rich conditions. *Frontiers in Microbiology*, *15*. https://doi.org/10.3389/fmicb.2024.1323160

Choudhuri, S. (2014). Sequence Alignment and Similarity Searching in Genomic Databases: BLAST and FASTA. In Bioinformatics for beginners : genes, genomes, molecular evolution, databases and analytical tools (pp. 133–155). London, Uk ; Waltham, Ma, Usa ; San Diego, Va, Usa: Academic Press Is An Imprint Of Elsevier. Retrieved from https://www.sciencedirect.com/book/9780124104716/bioinformatics-for-beginners

Cort, J. R., Yee, A., Edwards, A. M., Arrowsmith, C. H., & Kennedy, M. A. (2000). NMR structure determination and structure-based functional characterization of conserved hypothetical protein MTH1175 from *Methanobacterium thermoautotrophicum*. *Journal of Structural and Functional Genomics*, *1*(1), 15–25. https://doi.org/10.1023/a:1011348803324

Cortez, D., Neira, G., González, C., Vergara, E., & Holmes, D. S. (2022). A Large-Scale Genome-Based Survey of Acidophilic Bacteria Suggests That Genome Streamlining Is an Adaption for Life at Low pH. *Frontiers in Microbiology*, *13*. https://doi.org/10.3389/fmicb.2022.803241

Curatti, L., Brown, C. S., Ludden, P. W., & Rubio, L. M. (2005). Genes required for rapid expression of nitrogenase activity in *Azotobacter vinelandii*. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(18), 6291–6296. https://doi.org/10.1073/pnas.0501216102

da Silva, K., Florentino, L. A., da Silva, K. B., de Brandt, E., Vandamme, P., & de Souza Moreira, F. M. (2012). *Cupriavidus necator* isolates are able to fix nitrogen in symbiosis with different legume species. *Systematic and Applied Microbiology*, *35*(3), 175–182. https://doi.org/10.1016/j.syapm.2011.10.005

Demtröder, L., Pfänder, Y., Schäkermann, S., Bandow, J. E., & Masepohl, B. (2019). *NifA* is the master regulator of both nitrogenase systems in *Rhodobacter capsulatus*. *MicrobiologyOpen*, *8*(12). https://doi.org/10.1002/mbo3.921

DiASPora project. (s.d.). Acesso em 7 de maio de 2024, endereço digital: https://diaspora-project.de/

Ding, W., Baumdicker, F., & Neher, R. A. (2017). panX: pan-genome analysis and exploration. *Nucleic Acids Research*, *46*(1), e5–e5. https://doi.org/10.1093/nar/gkx977

Dixon, R., & Kahn, D. (2004). Genetic regulation of biological nitrogen fixation. *Nature Reviews Microbiology*, *2*(8), 621–631. https://doi.org/10.1038/nrmicro954

Doranga, S., & Conway, T. (2024). Nitrogen assimilation by *E. coli* in the mammalian intestine. *MBio*, *15*(3). https://doi.org/10.1128/mbio.00025-24

Dos Santos, P. C., Fang, Z., Mason, S. W., Setubal, J. C., & Dixon, R. (2012). Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics*, *13*(1). https://doi.org/10.1186/1471-2164-13-162

Dräger, A., & Planatscher, H. (2013). Metabolic Networks. In W. Dubitzky, O. Wolkenhauer, K.-H. Cho, & H. Yokota (Eds.), *Encyclopedia of Systems Biology* (pp. 1249–1251). New York, Ny: Springer New York.

Eddy, S. R. (2020). HMMER: biosequence analysis using profile hidden Markov models. Acesso em 12 de dezembro de 2024, endereço digital: http://hmmer.org/

EggNOG Database. (s.d.). Acesso em 12 de Agosto de 2023, endereço digital: http://eggnog6.embl.de

Elston, R., Mulligan, C., & Thomas, G. H. (2023). Flipping the switch: dynamic modulation of membrane transporter activity in bacteria. *Microbiology*, *169*(11). https://doi.org/10.1099/mic.0.001412

Feldbauer, R., Schulz, F., Horn, M., & Rattei, T. (2015). Prediction of microbial phenotypes based on comparative genomics. *BMC Bioinformatics*, *16*(S14). https://doi.org/10.1186/1471-2105-16-s14-s1

Feliciano, C. A., Eckenroth, B. E., Diaz, O. R., Doublié, S., & Shen, A. (2021). A lipoprotein allosterically activates the *CwlD* amidase during *Clostridioides difficile* spore formation. *PLoS Genetics*, 17(9), e1009791–e1009791. https://doi.org/10.1371/journal.pgen.1009791

Frees, D., Chastanet, A., Qazi, S., Sørensen, K., Hill, P., Msadek, T., & Ingmer, H. (2004). Clp ATPases are required for stress tolerance, intracellular replication and biofilm formation in *Staphylococcus aureus*. *Molecular Microbiology*, *54*(5), 1445–1462. https://doi.org/10.1111/j.1365-2958.2004.04368.x

Fu, J., Xie, X., Zhang, S., Kang, N., Zong, G., Zhang, P., & Cao, G. (2023). Rich Organic Nitrogen Impacts Clavulanic Acid Biosynthesis through the Arginine Metabolic Pathway in *Streptomyces clavuligerus* F613-1. *Microbiol Spectr*, *11*(1). https://doi.org/10.1128/spectrum.02017-22

Furtak, K., Grządziel, J., Gałązka, A., & Niedźwiecki, J. (2020). Prevalence of unclassified bacteria in the soil bacterial community from floodplain meadows (fluvisols) under simulated flood conditions revealed by a metataxonomic approachss. *CATENA*, *188*, 104448. https://doi.org/10.1016/j.catena.2019.104448

Gaby, J. C., & Buckley, D. H. (2014). A comprehensive aligned *NifH* gene database: a multipurpose tool for studies of nitrogen-fixing bacteria. *Database*, *2014*. https://doi.org/10.1093/database/bau001

Galloway, J. N., Aber, J. D., Erisman, J. W., Seitzinger, S. P., Howarth, R. W., Cowling, E. B., & Cosby, J. B. (2003). The Nitrogen Cascade. *BioScience*, 53(4), 341. https://doi.org/10.1641/0006-3568(2003)053[0341:tnc]2.0.co;2

Galperin, M. Y., Gaidenko, T. A., Mulkidjanian, A. Y., Nakano, M., & Price, C. W. (2001). *MHYT*, a new integral membrane sensor domain. *FEMS Microbiology Letters*, 205(1), 17–23. https://doi.org/10.1111/j.1574-6968.2001.tb10919.x

Galperin, M. Y., Yutin, N., Wolf, Y. I., Vera Alvarez, R., & Koonin, E. V. (2022). Conservation and Evolution of the Sporulation Gene Set in Diverse Members of the *Firmicutes*. *Journal of Bacteriology*, 204(6). https://doi.org/10.1128/jb.00079-22

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, *49*(D1), D274–D281. https://doi.org/10.1093/nar/gkaa1018

Griffiths, M. W., & Schraft, H. (2017). *Bacillus cereus* Food Poisoning. In C. E. R. Dodd, T. Aldsworth, R. A. Stein, D. O. Cliver, & H. P. Riemann (Eds.), Foodborne Diseases (pp. 395–405). *Academic Press*. https://doi.org/10.1016/b978-0-12-385007-2.00020-6

Gubler, M., Zürcher, T., & Hennecke, H. (1989). The *Bradyrhizobium japonicum* fixBCX operon: identification of fixX and of a 5' mRNA region affecting the level of the fixBCX transcript. *Molecular Microbiology*, *3*(2), 141–148. https://doi.org/10.1111/j.1365-2958.1989.tb01803.x

Heidrich, C., Templin, M. F., Ursinus, A., Merdanovic, M., Berger, J., Schwarz, H., ... Höltje, J.-V. (2001). Involvement of N-acetylmuramyl-l-alanine amidases in cell separation and antibiotic-induced autolysis of *Escherichia coli*. *Molecular Microbiology*, *41*(1), 167–178. https://doi.org/10.1046/j.1365-2958.2001.02499.x

Hilbert, D. W., & Piggot, P. J. (2004). Compartmentalization of Gene Expression during *Bacillus subtilis* Spore Formation. *Microbiology and Molecular Biology Reviews*, 68(2), 234–262. https://doi.org/10.1128/mmbr.68.2.234-262.2004

Hoch, J. A. (1976). Genetics of Bacterial Sporulation. *Advances in Genetics*, 18, 69–98. https://doi.org/10.1016/s0065-2660(08)60437-x

Holmquist, M. (2000). Alpha Beta-Hydrolase Fold Enzymes Structures, Functions and Mechanisms. *Current Protein and Peptide Science*, 1(2), 209–235. https://doi.org/10.2174/1389203003381405

Hou, S., Makarova, K. S., Saw, J. H., Senin, P., Ly, B. V., Zhou, Z., ... Wang, L. (2008). Complete genome sequence of the extremely acidophilic methanotroph isolate V4, *Methylacidiphilum infernorum*, a representative of the bacterial phylum *Verrucomicrobia*. *Biology Direct*, *3*(26), 26. https://doi.org/10.1186/1745-6150-3-26

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., ... Bork, P. (2018). eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Huynh, T. N., & Stewart, V. (2023). Purine catabolism by *enterobacteria*. In R. K. Poole & D. J. Kelly (Eds.), *Advances in Microbial Physiology - Volume 82* (pp. 1–321). Academic Press.

Igarashi, R. Y., & Seefeldt, L. C. (2003). Nitrogen Fixation: The Mechanism of the Mo-Dependent Nitrogenase. *Critical Reviews in Biochemistry and Molecular Biology*, *38*(4), 351–384. https://doi.org/10.1080/10409230391036766

Imran, A., Hakim, S., Tariq, M., Nawaz, M. S., Laraib, I., Gulzar, U., ... Ahmad, M. (2021). Diazotrophs for Lowering Nitrogen Pollution Crises: Looking Deep Into the Roots. *Frontiers in Microbiology*, *12*. https://doi.org/10.3389/fmicb.2021.637815

Jeruzalmi, D., Yurieva, O., Zhao, Y., Young, M., Stewart, J., Hingorani, M., ... Kuriyan, J. (2001). Mechanism of Processivity Clamp Opening by the Delta Subunit Wrench of the Clamp Loader Complex of *E. coli* DNA Polymerase III. *Cell*, *106*(4), 417–428. https://doi.org/10.1016/s0092-8674(01)00462-7

Jim, K., Parmar, K., Singh, M., & Saeed Tavazoie. (2004). A Cross-Genomic Approach for Systematic Mapping of Phenotypic Traits to Genes. Genome Research, 14(1), 109–115. https://doi.org/10.1101/gr.1586704

Kanamaru, K., Stephenson, S., & Perego, M. (2002). Overexpression of the *PepF* Oligopeptidase Inhibits Sporulation Initiation in *Bacillus subtilis*. *Journal of Bacteriology*, *184*(1), 43–50. https://doi.org/10.1128/jb.184.1.43-50.2002

Kaschuk, G., & Hungria, M. (2017). Diversity and Importance of Diazotrophic Bacteria to Agricultural Sustainability in the Tropics. *Springer EBooks*, 269–292. https://doi.org/10.1007/978-3-319-55804-2_12

KEGG ORTHOLOGY. (s.d.). Acesso em 11 de dezembro de 2024, endereço digital: https://www.genome.jp

Kennedy, C., & Dean, D. (1992). The *NifU*, *NifS* and *NifV* gene products are required for activity of all three nitrogenases of *Azotobacter vinelandii*. *MGG Molecular & General Genetics*, *231*(3), 494–498. https://doi.org/10.1007/bf00292722

Kerff, F., Petrella, S., Mercier, F., Sauvage, E., Herman, R., Pennartz, A., ... Charlier, P. (2009). Specific Structural Features of the N-Acetylmuramoyl-l-Alanine Amidase AmiD from *Escherichia coli* and Mechanistic Implications for Enzymes of This Family. *Journal of Molecular Biology*, *397*(1), 249–259. https://doi.org/10.1016/j.jmb.2009.12.038

Koide, A., & Hoch, J. A. (1994). Identification of a second oligopeptide transport system in *Bacillus subtilis* and determination of its role in sporulation. Molecular Microbiology, 13(3), 417–426. https://doi.org/10.1111/j.1365-2958.1994.tb00436.x

Konopka, J. B. (2012). N-Acetylglucosamine Functions in Cell Signaling. *Scientifica*, *2012*, 1–15. https://doi.org/10.6064/2012/489208

Koropatkin, N. M., Pakrasi, H. B., & Smith, T. J. (2006). Atomic structure of a nitrate-binding protein crucial for photosynthetic productivity. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(26), 9820–9825. https://doi.org/10.1073/pnas.0602517103

Krizsan, A., Knappe, D., & Hoffmann, R. (2015). Influence of the *yjiL-mdtM* Gene Cluster on the Antibacterial Activity of Proline-Rich Antimicrobial Peptides Overcoming *Escherichia coli* Resistance Induced by the Missing SbmA Transporter System. *Antimicrobial Agents and Chemotherapy*, *59*(10), 5992–5998. https://doi.org/10.1128/aac.01307-15

Krulwich, T. A., Sachs, G., & Padan, E. (2011). Molecular aspects of bacterial pH sensing and homeostasis. *Nature Reviews Microbiology*, *9*(5), 330–343. https://doi.org/10.1038/nrmicro2549

Kucho, K., Tamari, D., Matsuyama, S., Nabekura, T., & Tisa, L. S. (2017). Nitrogen Fixation Mutants of the *Actinobacterium Frankia Casuarinae CcI3*. *Microbes and Environments*, 32(4), 344–351. https://doi.org/10.1264/jsme2.me17099

Kumarevel, T., Fujimoto, Z., Karthe, P., Oda, M., Mizuno, H., & Kumar, P. K. R. (2004). Crystal Structure of Activated HutP: An RNA Binding Protein that Regulates Transcription of the hut Operon in *Bacillus subtilis*. *Structure*, *12*(7), 1269–1280. https://doi.org/10.1016/j.str.2004.05.005

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M. R., Ahn, T.-H., … Ussery, D. W. (2015). Insights from 20 years of bacterial genome sequencing. *Functional & Integrative Genomics*, *15*(2), 141–161. https://doi.org/10.1007/s10142-015-0433-4

Leroy, M., Piton, J., Gilet, L., Pellegrini, O., Proux, C., Coppée, J., Figaro, S., & Condon, C. (2017). Rae1/YacP, a new endoribonuclease involved in ribosome‐dependent mRNA decay in *Bacillus subtilis*. *The EMBO Journal*, 36(9), 1167–1181. https://doi.org/10.15252/embj.201796540

LeVeen, H. H., LeVeen, E. G., & LeVeen, R. F. (1994). Awakenings to the pathogenicity of urease and the requirement for continuous long term therapy. *Biomedicine & Pharmacotherapy*, 48(3-4), 157–166. https://doi.org/10.1016/0753-3322(94)90104-x

Li, Q., Zhang, H., Zhang, L., & Chen, S. (2021). Functional analysis of multiple *NifB* genes of *Paenibacillus* strains in synthesis of Mo-, Fe- and V-nitrogenases. *Microbial Cell Factories*, *20*(1). https://doi.org/10.1186/s12934-021-01629-9

Lie, T. J., Kuo, Y. P., Leite, M., Costa, K. C., Harwood, C. S., & Leigh, J. A. (2022). A Genetic Study of *Nif*-Associated Genes in a Hyperthermophilic Methanogen. *Microbiology Spectrum*, *10*(1). https://doi.org/10.1128/spectrum.02093-21

Lin, D. X., Tang, H., Wang, E. T., & Chen, W. X. (2009). An ABC transporter is required for alkaline stress and potassium transport regulation in *Sinorhizobium meliloti*. *FEMS Microbiology Letters*, *293*(1), 35–41. https://doi.org/10.1111/j.1574-6968.2009.01500.x

Lin, Z., Xue, H., Malakhov, M. M., Knutson, K. A., & Pan, W. (2022). Accounting for nonlinear effects of gene expression identifies additional associated genes in transcriptome-wide association studies. *Human Molecular Genetics*, *31*(14), 2462–2470. https://doi.org/10.1093/hmg/ddac015

Liu, S., Wu, W., Zhao, Q., Liang, H., Che, S., Zhang, H., … Bartlam, M. (2022). Structural characterization of the urease accessory protein UreF from *Klebsiella pneumoniae*. *Acta Crystallographica. Section F, Structural Biology Communications*, *78*(2), 75–80. https://doi.org/10.1107/s2053230x22000474

Louca, S., Mazel, F., Doebeli, M., & Parfrey, L. W. (2019). A census-based estimate of Earth's bacterial and archaeal diversity. *PLOS Biology*, *17*(2), e3000106. https://doi.org/10.1371/journal.pbio.3000106

Lowe, S. E., Pankratz, H. S., & Zeikus, J. G. (1989). Influence of pH extremes on sporulation and ultrastructure of *Sarcina ventriculi*. *Journal of Bacteriology*, *171*(7), 3775–3781. https://doi.org/10.1128/jb.171.7.3775-3781.1989

Lüftinger, L., Hyden, P., & Feldbauer, R. (s.d.). Phenotrex 0.6.0 documentation. Acesso em 8 de maio de 2024, endereço digital: https://phenotrex.readthedocs.io/

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4765–4774). Red Hook, NY, USA: Curran Associates Inc. Endereço digital: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

MacDonald, N. J., & Beiko, R. G. (2010). Efficient learning of microbial genotype–phenotype association rules. *Bioinformatics*, *26*(15), 1834–1840. https://doi.org/10.1093/bioinformatics/btq305

Madin, J. S., Nielsen, D. A., Brbic, M., Corkrey, R., Danko, D., Edwards, K., … Rocha, E. P. C. (2020). A synthesis of bacterial and archaeal phenotypic trait data. *Scientific Data*, *7*(1). https://doi.org/10.1038/s41597-020-0497-4

Magnusson, W. E., & Mourão, G. (2005). Estatística sem matemática: a ligação entre questões e análise. *Editora Planta*.

Mantione, K. J., Kream, R. M., Kuzelova, H., Ptacek, R., Raboch, J., Samuel, J. M., & Stefano, G. B. (2014). Comparing Bioinformatic Gene Expression Profiling Methods: Microarray and RNA-Seq. *Medical Science Monitor Basic Research*, *20*(23), 138–141. https://doi.org/10.12659/msmbr.892101

Markowitz, V. M., Chen, I.-M. . A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., … Kyrpides, N. C. (2011). IMG: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, *40*(D1), D115–D122. https://doi.org/10.1093/nar/gkr1044

Masrati, G., Dwivedi, M., Rimon, A., Gluck-Margolin, Y., Kessel, A., Ashkenazy, H., … Ben-Tal, N. (2018). Broad phylogenetic analysis of cation/proton antiporters reveals transport determinants. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-06770-5

Meeske, A. J., Rodrigues, C. D. A., Brady, J., Lim, H. C., Bernhardt, T. G., & Rudner, D. Z. (2016). High-Throughput Genetic Screens Identify a Large and Diverse Collection of New Sporulation Genes in *Bacillus subtilis*. *PLOS Biology*, 14(1), e1002341. https://doi.org/10.1371/journal.pbio.1002341

Mira, A., Ochman, H., & Moran, N. A. (2001). Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, *17*(10), 589–596. https://doi.org/10.1016/s0168-9525(01)02447-7

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G., Sonnhammer, E. L. L., … Bateman, A. (2020). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. https://doi.org/10.1093/nar/gkaa913

Moir, A. (2006). How do spores germinate? Journal of Applied Microbiology, 101(3), 526–530. https://doi.org/10.1111/j.1365-2672.2006.02885.x

Montenegro, J. D. (2022). Gene Co-expression Network Analysis. Methods in Molecular Biology (Clifton, N.J.), 2443, 387–404. https://doi.org/10.1007/978-1-0716-2067-0_19

Morawska, L. P., Detert, R. G. J. O. W., Frenzel, E., Dirkzwager, M., Hoffmann, T., Bremer, E., & Kuipers, O. P. (2022). Stress-induced activation of the proline biosynthetic pathway in *Bacillus subtilis* : a population-wide and single-cell study of the osmotically controlled *proHJ* promoter. *Microbial Biotechnology*, *15*(9). https://doi.org/10.1111/1751-7915.14073

Mu, H., Li, W., Hu, K., Zhu, H., Ren, S., Peng, Y., … Quan, H. (2022). Characterization of *Achromobacter denitrificans* QHR-5 for heterotrophic nitrification-aerobic denitrification with iron oxidation function isolated from BSIS：Nitrogen removal performance and enhanced SND capability of BSIS. *Biochemical Engineering Journal*, *191*, 108759. https://doi.org/10.1016/j.bej.2022.108759

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., … Bork, P. (2009). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Research*, *38*(suppl_1), D190–D195. https://doi.org/10.1093/nar/gkp951

NCBI. (s.d.). Acesso em 13 de agosto de 2024, endereço digital: https://www.ncbi.nlm.nih.gov/research/cog/cog

Nelson, D. L., & Cox, M. M. (2021). Lehninger Principles Of Biochemistry. (8th ed.). W H Freeman.

Nielsen, J., Tillegreen, C. B., & Petranovic, D. (2022). Innovation trends in industrial biotechnology. *Trends in Biotechnology*, *40*(10). https://doi.org/10.1016/j.tibtech.2022.03.007

Notting, F., Pirovano, W., Sybesma, W., & Kort, R. (2023). The butyrate-producing and spore-forming bacterial genus *Coprococcus* as a potential biomarker for neurological disorders. *Gut Microbiome*, 4(e16). https://doi.org/10.1017/gmb.2023.14

Nouwen, N., Arrighi, J.-F., Cartieaux, F., Chaintreuil, C., Gully, D., Klopp, C., & Giraud, E. (2017). The role of rhizobial (*NifV*) and plant (*FEN1*) homocitrate synthases in *Aeschynomene*/photosynthetic *Bradyrhizobium* symbiosis. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-00559-0

Ogawa, T., Tanaka, A., Kawamoto, J., & Kurihara, T. (2018). Purification and characterization of 1-acyl-sn-glycerol-3-phosphate acyltransferase with a substrate preference for polyunsaturated fatty acyl donors from the eicosapentaenoic acid-producing bacterium *Shewanella livingstonensis* Ac10. *The Journal of Biochemistry*, *164*(1), 33–39. https://doi.org/10.1093/jb/mvy025

Pearson, W. R. (2013). An Introduction to Sequence Similarity ("Homology") Searching. *Current Protocols in Bioinformatics*, *42*(1), 3.1.1–3.1.8. https://doi.org/10.1002/0471250953.bi0301s42

Penuelas, J., Coello, F., & Sardans, J. (2023). A better use of fertilizers is needed for global food security and environmental sustainability. *Agriculture & Food Security*, *12*(1). https://doi.org/10.1186/s40066-023-00409-5

Petersohn, A., Brigulla, M., Haas, S. A., Hoheisel, J. D., Völker, U., & Hecker, M. (2001). Global Analysis of the General Stress Response of *Bacillus subtilis*. *Journal of Bacteriology*, 183(19), 5617–5631. https://doi.org/10.1128/jb.183.19.5617-5631.2001

Pi, H.-W., Lin, J.-J., Chen, C.-A., Wang, P.-H., Chiang, Y.-R., Huang, C.-C., … Li, W.-H. (2022). Origin and Evolution of Nitrogen Fixation in Prokaryotes. *Molecular Biology and Evolution*, *39*(9). https://doi.org/10.1093/molbev/msac181

Poolman, B. (2023). Physicochemical homeostasis in bacteria. *Fems Microbiology Reviews*, *47*(4). https://doi.org/10.1093/femsre/fuad033

Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., … Bork, P. (2013). eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, *42*(D1), D231–D239. https://doi.org/10.1093/nar/gkt1253

Qi, Y., Liu, H., Chen, X., & Liu, L. (2019). Engineering microbial membranes to increase stress tolerance of industrial strains. *Metabolic Engineering*, *53*, 24–34. https://doi.org/10.1016/j.ymben.2018.12.010

Quatrini, R., Appia-Ayme, C., Denis, Y., Jedlicki, E., Holmes, D. S., & Bonnefoy, V. (2009). Extending the models for iron and sulfur oxidation in the extreme Acidophile *Acidithiobacillus ferrooxidans*. *BMC Genomics*, *10*(1), 394. https://doi.org/10.1186/1471-2164-10-394

Queraltó, C., Álvarez, R., Ortega, C., Díaz-Yáñez, F., Paredes-Sabja, D., & Gil, F. (2023). Role and Regulation of *Clp* Proteases: A Target against Gram-Positive Bacteria. *Bacteria*, *2*(1), 21–36. https://doi.org/10.3390/bacteria2010002

Ragot, S. A., Kertesz, M. A., & Bünemann, E. K. (2015). *phoD* Alkaline Phosphatase Gene Diversity in Soil. *Applied and Environmental Microbiology*, 81(20), 7281–7289. https://doi.org/10.1128/aem.01823-15

Ramoneda, J., Stallard-Olivera, E., Hoffert, M., Winfrey, C. C., Stadler, M., Niño-García, J. P., & Fierer, N. (2023). Building a genome-based understanding of bacterial pH preferences. *Science Advances*, *9*(17). https://doi.org/10.1126/sciadv.adf8998

Read, T. D., & Massey, R. C. (2014). Characterizing the genetic basis of bacterial phenotypes using genome-wide association studies: a new direction for bacteriology. *Genome Medicine*, 6(11). https://doi.org/10.1186/s13073-014-0109-z

Reimer, L. C., Sardà Carbasse, J., Koblitz, J., Ebeling, C., Podstawka, A., & Overmann, J. (2021). BacDive in 2022: the knowledge base for standardized bacterial and archaeal data. *Nucleic Acids Research*, *50*(D1), D741–D746. https://doi.org/10.1093/nar/gkab961

Rivera-Araya, J., Pollender, A., Huynh, D., Schlömann, M., Chávez, R., & Levicán, G. (2019). Osmotic Imbalance, Cytoplasm Acidification and Oxidative Stress Induction Support the High Toxicity of Chloride in Acidophilic Bacteria. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.02455

Rocha, E. P. C., Danchin, A., & Alain Viari. (1999). Functional and evolutionary roles of long repeats in prokaryotes. *Research in Microbiology*, *150*(9-10), 725–733. https://doi.org/10.1016/s0923-2508(99)00120-5

Romero-Rodríguez, A., Ruiz-Villafán, B., Martínez-de la Peña, C. F., & Sánchez, S. (2023). Targeting the Impossible: A Review of New Strategies against Endospores. Antibiotics, 12(2), 248. https://doi.org/10.3390/antibiotics12020248

Rosman, A. C., Clara, M., Thiebaut, F., Forero, G., Alves, E., & Adriana Silva Hemerly. (2024). Dual RNA-seq of maize and *H. seropedicae* ZAE94 association, in different doses of nitrate, reveals novel insights into Plant-PGPB-environment relationship. *Frontiers in Plant Science*, 15. https://doi.org/10.3389/fpls.2024.1346523

Saber, M. M., & Shapiro, B. J. (2020). Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microbial Genomics*, 6(3). https://doi.org/10.1099/mgen.0.000337

Saito, S., Baba, S., Nikai, H., Fujisawa, R., & Fujiwara, T. (2024). *Flavonifractor plautii* Bacteremia With Generalized Peritonitis: A Case Report and Literature Review. *Cureus*, 16(7). https://doi.org/10.7759/cureus.63890

Schmid, P. J., Forstner, P., & Kittinger, C. (2024). Sliding motility of *Bacillus cereus* mediates vancomycin pseudo-resistance during antimicrobial susceptibility testing. *Journal of Antimicrobial Chemotherapy*, 79(7), 1628–1636. https://doi.org/10.1093/jac/dkae156

Schmidt, R., Zahn, R., Bernd Bukau, & Mogk, A. (2009). *ClpS* is the recognition component for *Escherichia coli* substrates of the N-end rule degradation pathway. *Molecular Microbiology*, *72*(2), 506–517. https://doi.org/10.1111/j.1365-2958.2009.06666.x

Shoda, S., Uyama, H., Kadokawa, J., Kimura, S., & Kobayashi, S. (2016). Enzymes as Green Catalysts for Precision Macromolecular Synthesis. *Chemical Reviews*, 116(4), 2307–2413. https://doi.org/10.1021/acs.chemrev.5b00472

Silva, P. R. A. da, Simões-Araújo, J. L., Vidal, M. S., Cruz, L. M., Souza, E. M. de, & Baldani, J. I. (2018). Draft genome sequence of *Paraburkholderia tropica* Ppe8 strain, a sugarcane endophytic diazotrophic bacterium. *Brazilian Journal of Microbiology : [Publication of the Brazilian Society for Microbiology]*, *49*(2), 210–211. https://doi.org/10.1016/j.bjm.2017.07.005

Simon, H. M., Homer, M. J., & Roberts, G. P. (1996). Perturbation of *NifT* expression in *Klebsiella pneumoniae* has limited effect on nitrogen fixation. *Journal of Bacteriology*, *178*(10), 2975–2977. https://doi.org/10.1128/jb.178.10.2975-2977.1996

Singh, A. K., Singh, R., Tomar, D., Pandya, C. D., & Singh, R. (2012). The leucine aminopeptidase of Staphylococcus aureus is secreted and contributes to biofilm formation. *International Journal of Infectious Diseases*, *16*(5), e375–e381. https://doi.org/10.1016/j.ijid.2012.01.009

Slonczewski, J. L., Fujisawa, M., Dopson, M., & Krulwich, T. A. (2009). Cytoplasmic pH Measurement and Homeostasis in Bacteria and Archaea. *Advances in Microbial Physiology*, *55*(317), 1–317. https://doi.org/10.1016/s0065-2911(09)05501-5

Smits, W. K., Kuipers, O. P., & Veening, J.-W. (2006). Phenotypic variation in bacteria: the role of feedback regulation. *Nature Reviews Microbiology*, *4*(4), 259–271. https://doi.org/10.1038/nrmicro1381

Sogin, M. L., Morrison, H. G., Huber, J. A., Welch, D. M., Huse, S. M., Neal, P. R., … Herndl, G. J. (2006). Microbial diversity in the deep sea and the underexplored "rare biosphere." *Proceedings of the National Academy of Sciences*, *103*(32), 12115–12120. https://doi.org/10.1073/pnas.0605127103

Sperotto, R. A., Gross, J., Vedoy, C., Passaglia, L. M. P., & Schrank, I. S. (2004). The Electron Transfer Flavoprotein fixABCX Gene Products from *Azospirillum brasilense* Show a *NifA*-Dependent Promoter Regulation. *Current Microbiology*, *49*(4), 267–273. https://doi.org/10.1007/s00284-004-4318-3

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., Iyer, R., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big Data: Astronomical or Genomical? PLOS Biology, 13(7), e1002195. https://doi.org/10.1371/journal.pbio.1002195

Stricker, O., Masepohl, B., Klipp, W., & Böhme, H. (1997). Identification and characterization of the *NifV-NifZ-NifT* gene region from the filamentous cyanobacterium *Anabaena sp*. strain PCC 7120. *Journal of Bacteriology*, *179*(9), 2930–2937. https://doi.org/10.1128/jb.179.9.2930-2937.1997

Sullivan, J. T., Brown, S. D., & Ronson, C. W. (2013). The *NifA-RpoN* Regulon of *Mesorhizobium loti* Strain R7A and Its Symbiotic Activation by a Novel *LacI/GalR*-Family Regulator. *PLOS ONE*, *8*(1), e53762–e53762. https://doi.org/10.1371/journal.pone.0053762

Sutandy, F. X. R., Qian, J., Chen, C.-S., & Zhu, H. (2013). Overview of Protein Microarrays. *Current Protocols in Protein Science*, *72*(1), 27.1.1–27.1.16. https://doi.org/10.1002/0471140864.ps2701s72

Tadrowski, A. C., Evans, M. R., & Waclaw, B. (2018). Phenotypic Switching Can Speed up Microbial Evolution. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-27095-9

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A Genomic Perspective on Protein Families. *Science*, *278*(5338), 631–637. https://doi.org/10.1126/science.278.5338.631

Tijerina-Rodriguez, L. E., Martínez-Meléndez, A., Villarreal-Treviño, L., Morfín-Otero, R., Camacho-Ortiz, A., Baines, S. D., Rodríguez-Noriega, E., & Garza-González, E. (2018). Enhanced Sporulation and Vancomycin Resistance Associated With *Clostridium difficile* From Recurrent Infections. *Open Forum Infectious Diseases*, 5 (suppl_1), S186–S186. https://doi.org/10.1093/ofid/ofy210.511

Tijerina-Rodríguez, L. E., Villarreal-Treviño, L., Baines, S. D., Morfín-Otero, R., Camacho-Ortíz, A., Flores-Treviño, S., Maldonado-Garza, H., Rodríguez-Noriega, E., & Garza-González, E. (2019). High sporulation and overexpression of virulence factors in biofilms and reduced susceptibility to vancomycin and linezolid in recurrent *Clostridium [Clostridioides] difficile* infection isolates. *PLOS ONE*, 14(7), e0220671–e0220671. https://doi.org/10.1371/journal.pone.0220671

Tripura, C., Sudhakar Reddy, P., Reddy, M. K., Sashidhar, B., & Podile, A. R. (2007). Glucose dehydrogenase of a rhizobacterial strain of *Enterobacter asburiae* involved in mineral phosphate solubilization shares properties and sequence homology with other members of *Enterobacteriaceae*. *Indian Journal of Microbiology*, 47(2), 126–131. https://doi.org/10.1007/s12088-007-0025-7

Tsoy, O., Ravcheev, D. A., & Arcady Mushegian. (2009). Comparative Genomics of Ethanolamine Utilization. *Journal of Bacteriology*, *191*(23), 7157–7164. https://doi.org/10.1128/jb.00838-09

Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, *5*(1), 170203. https://doi.org/10.1038/sdata.2017.203

UniProt. (s.d.). Acesso em 15 de dezembro de 2024, endereço digital: https://www.uniprot.org/uniprotkb

Vliegenthart, J. S., P A Ketelaar-van Gaalen, & van. (1991). Nucleotide sequence of the *AacC3* gene, a gentamicin resistance determinant encoding aminoglycoside-(3)-N-acetyltransferase III expressed in Pseudomonas aeruginosa but not in *Escherichia coli*. *Antimicrobial Agents and Chemotherapy*, *35*(5), 892–897. https://doi.org/10.1128/aac.35.5.892

Wang, J. Y., Wang, R., Zhang, Y. M., Liu, H. C., Chen, W. F., Wang, E. T., … Chen, W. X. (2013). *Bradyrhizobium daqingense sp. nov.*, isolated from soybean nodules. *International Journal of Systematic and Evolutionary Microbiology*, *63*(Pt_2), 616–624. https://doi.org/10.1099/ijs.0.034280-0

Wasfy, R. M., Abdoulaye, A., Borentain, P., Mbaye, B., Alou, M. T., Caputo, A., Andrieu, C., Mottola, G., Levasseur, A., Million, M., & Gerolami, R. (2024). *Thomasclavelia ramosais* a Signature of Gut Dysbiosis associated with Alcohol-Related Hepatocellular Carcinoma: A First Microbial Culturomics Study. *MedRxiv* (Cold Spring Harbor Laboratory). https://doi.org/10.1101/2024.08.19.24312231

Weimann, A., Mooren, K., Frank, J., Pope, P. B., Bremges, A., & McHardy, A. C. (2016). From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. *MSystems*, *1*(6). https://doi.org/10.1128/msystems.00101-16

Wen, Z.-L., Yang, M.-K., Fazal, A., Liao, Y.-H., Cheng, L.-R., Hua, X.-M., … Yang, Y.-H. (2020). Differential Impacts on Bacterial Composition and Abundance in Rhizosphere Compartments between Al-Tolerant and Al-Sensitive Soybean Genotypes in Acidic Soil. *Journal of Microbiology and Biotechnology*, *30*(8), 1169–1179. https://doi.org/10.4014/jmb.2003.03018

Wetzel, D., & McBride, S. M. (2020). The Impact of pH on *Clostridioides difficile* Sporulation and Physiology. *Applied and Environmental Microbiology*, *86*(4). https://doi.org/10.1128/AEM.02706-19

Winkelmann, I., Povilas Uzdavinys, Kenney, I. M., Brock, J., Meier, P. F., Wagner, L.-M., … Drew, D. (2022). Crystal structure of the Na+/H+ antiporter *NhaA* at active pH reveals the mechanistic basis for pH sensing. *Nature Communications*, *13*(1). https://doi.org/10.1038/s41467-022-34120-z

Wright, P. A. (1995). Nitrogen excretion: three end products, many physiological roles. *Journal of Experimental Biology*, *198*(2), 273–281. https://doi.org/10.1242/jeb.198.2.273

Xiao, S., Li, S., Yuan, Z., & Zhou, L. (2020). Pyrroline-5-carboxylate reductase 1 (PYCR1) upregulation contributes to gastric cancer progression and indicates poor survival outcome. *Annals of Translational Medicine*, *8*(15), 937–937. https://doi.org/10.21037/atm-19-4402

Yang, M., Wen, Z., Hao, C., Fazal, A., Liao, Y., Luo, F., … Yang, Y. (2021). Differential Assembly and Shifts of the Rhizosphere Bacterial Community by a Dual Transgenic Glyphosate-Tolerant Soybean Line with and without Glyphosate Application. *Horticulturae*, *7*(10), 374. https://doi.org/10.3390/horticulturae7100374

Yang, N., Li, Z., Wu, Z., Liu, X., Zhang, Y., Sun, T., … Tong, Y. (2023). Differential effects of nitrate and ammonium on the growth of algae and microcystin production by nitrogen-fixing *Nostoc sp*. and non-nitrogen-fixing *Microcystis aeruginosa*. *Water Science & Technology*, *88*(1), 136–150. https://doi.org/10.2166/wst.2023.205

Yeom, J., Gao, X., & Groisman, E. A. (2018). Reduction in adaptor amounts establishes degradation hierarchy among protease substrates. *Proceedings of the National Academy of Sciences of the United States of America*, 115(19). https://doi.org/10.1073/pnas.1722246115

Yue, D., Peng, Y., Yin, Q., & Xiao, L. (2014). Proteomic analysis of *Microcystis aeruginosa* in response to nitrogen and phosphorus starvation. *Journal of Applied Phycology*, *27*(3), 1195–1204. https://doi.org/10.1007/s10811-014-0405-4

Zaheer, R., Morton, R., Proudfoot, M., Yakunin, A., & Finan, T. M. (2009). Genetic and biochemical properties of an alkaline phosphatase PhoX family protein found in many bacteria. *Environmental Microbiology*, 11(6), 1572–1587. https://doi.org/10.1111/j.1462-2920.2009.01885.x

Zahran, H. H. (1999). *Rhizobium*-Legume Symbiosis and Nitrogen Fixation under Severe Conditions and in an Arid Climate. *Microbiology and Molecular Biology Reviews*, *63*(4), 968–989. https://doi.org/10.1128/mmbr.63.4.968-989.1999

Zeigler, D. R., & Nicholson, W. L. (2017). Experimental evolution of *Bacillus subtilis*. *Environmental Microbiology*, 19(9), 3415–3422. https://doi.org/10.1111/1462-2920.13831

Zeytuni, N., Flanagan, K. A., Worrall, L. J., Massoni, S. C., Camp, A. H., & Strynadka, N. C. J. (2017). Structural characterization of SpoIIIAB sporulation-essential protein in Bacillus subtilis. *Journal of Structural Biology*, *202*(2), 105–112. https://doi.org/10.1016/j.jsb.2017.12.009

Zhang, A., Lebrun, R., Espinosa, L., Galinier, A., & Pompeo, F. (2022). *PrkA* is an ATP-dependent protease that regulates sporulation in *Bacillus subtilis*. *Journal of Biological Chemistry*, 298(10). https://doi.org/10.1016/j.jbc.2022.102436

Zhang, H., Venkatesan, S., Ng, E., & Nan, B. (2023). Coordinated peptidoglycan synthases and hydrolases stabilize the bacterial cell wall. *Nature Communications*, *14*(1), 5357. https://doi.org/10.1038/s41467-023-41082-3

Zhao, Z., Cristian, A., & Rosen, G. (2020). Keeping up with the genomes: efficient learning of our increasing knowledge of the tree of life. *BMC Bioinformatics*, *21*(1). https://doi.org/10.1186/s12859-020-03744-7

Glossário e abreviações

Adaptive Boosting (AdaBoost) - técnica de aprendizado de máquina que constroi um modelo preditivo através da combinação de múltiplos modelos preditivos mais simples. Durante o treinamento são atribuídos pesos diferentes aos dados de treinamento, sendo dado mais peso aos dados considerados mais difíceis de prever.

Aprendizado de Máquina - algoritmos e técnicas que permitem aos computadores aprenderem padrões e fazerem previsões a partir de dados, sem serem explicitamente programados para isso. O aprendizado de máquina envolve a utilização de modelos matemáticos e estatísticos para identificar relações e padrões nos dados, permitindo que o computador faça previsões ou tome decisões com base nesses padrões identificados.

Árvore de Decisão - algoritmo de aprendizado de máquina que representa uma série de decisões e em forma de árvore que resultam em uma inferência. Cada nó interno da árvore representa um parâmetro dos dados a ser analisado, e cada ramo da árvore representa um possível resultado de uma decisão baseada nesse parâmetro.

AUC-ROC - métrica para quantificação da capacidade de um classificador binário em distinguir entre classes positivas e negativas. Um valor acima de 0,5 indica que um modelo classifica mais verdadeiros positivos e negativos do que falsos positivos e negativos. Mais útil quando não há desequilíbrio considerável entre as classes dos dados testados e o custo da classificação incorreta é consistente entre elas.

AUC-PR - métrica para a relação de equilíbrio entre precisão e recall para um classificador binário, tornando a uma métrica útil no caso de dados com classes de quantidades desequilibradas.

Bernoulli Naive Bayes - algoritmo de classificação probabilístico que se baseia no teorema de Bayes.

Cluster of Ortholog Genes - grupo de genes ortólogos catalogado no banco de dados COG (Galperin *et al.*, 2021). A atualização atual (2021) contém 4,981 grupos ortólogos provenientes dos genomas de 2.103 espécies bacterianas e 193 de arqueas.

Elastic Net - regressão linear com a introdução da penalidade de regularização L1 (penalidade de Lasso) e regularização L2 (penalidade de Ridge).

Escalonamento Multidimensional (EMD) - método que utiliza distâncias de similaridade entre indivíduos de um conjunto de dados para organiza-los em uma

representação dimensional mais baixa, como bidimensional ou tridimensional, para facilitar a visualização de indivíduos similares e dissimilares.

Espaço dos parâmetros - abstração de espaço dimensional onde objetos podem ser mapeados de acordo com os seus parâmetros. Por exemplo, no caso de genomas seus parâmetros podem ser os genes presentes ou os grupos funcionais presentes.

Extended Clusters of Orthologous Groups (eNOG) - banco de dados EggNOG composto por grupos ortólogos identificados por um processo computacional não supervisionado, e por esse motivo considerados hipotéticos.

F1 - média harmônica de precisão e recall para classificações binárias, sendo portanto um indicador unificado de ambas as métricas.

Genes homólogos - genes que possuem uma origem evolutiva comum mas que sofreram especiação. Eles podem estar presentes no mesmo organismo (genes parálogos) ou em organismos diferentes (genes ortólogos), descrevendo variações e conservações genéticas ao longo da evolução. Esses genes podem desempenhar funções semelhantes ou diferentes, dependendo das adaptações específicas de cada organismo.

Genes ortólogos - genes homólogos que apresentam funcionalidades biológicas semelhantes.

Gradient Boosting - técnica de aprendizado de máquina que constrói um modelo preditivo através da combinação de múltiplos modelos preditivos simples. Os modelos são treinados sequencialmente, dessa forma cada novo modelo é treinado tendo como objetivo corrigir os erros residuais dos modelos anteriores.

Grupo de genes ortólogos/Grupo ortólogo - grupos de genes que compartilham uma ancestralidade comum e preservam sua funcionalidade.

K-Nearest neighbors (KNN) - algoritmo de aprendizado de máquina baseado na ideia de que objetos semelhantes tendem a estar próximos no espaço de parâmetros. Realiza a separação do espaço de parâmetros em agrupamentos.

LightGBM - algoritmo baseado em Gradient Boosting aplicado a Árvores de Decisão desenvolvida pela Microsoft. Se destaca por utilizar algoritmos de otimização Gradient-based One-Side Sampling e Exclusive Feature Bundling para reduzir o tempo de treinamento.

Parâmetro - uma informação específica ou característica individual dentro de um conjunto que define determinado objeto. No contexto biológico, por exemplo, os genes

presentes em um genoma podem ser considerados seus parâmetros, pois representam atributos mensuráveis que contribuem para a definição do organismo.

Partial Least Squares (PLS) - técnica estatística que busca combinações lineares dos parâmetros que melhor explicam a variância nas predições, pela maximização da covariância entre os parâmetros e a predição.

Passive Aggressive Regressor (PAR) - algoritmo de aprendizado de máquina com aprendizado incremental e adaptativo pela minimização da perda localmente em cada exemplo de treinamento em vez da minimização global de uma função de custo.

Precisão - porcentagem de acertos das previsões positivas.

Quantile Regression - técnica estatística usada para modelar diferentes quartis da distribuição de uma variável, em vez de apenas a média, como na regressão linear tradicional.

Random Forest - algoritmo que consiste em uma coleção de árvores de decisão individuais. Cada árvore de decisão é treinada de forma independente, usando uma amostra aleatória dos dados de treinamento e uma seleção aleatória de parâmetros. Cada árvore de decisão contribui com sua própria previsão e a previsão final é determinada pela média ou votação das previsões individuais.

Recall/Sensibilidade - porcentagem de casos positivos preditos corretamente.

Ridge - regressão linear com a introdução de uma penalidade de regularização L2 (também conhecida como penalidade de Ridge).

Support Vector Machine (SVM) - algoritmo de aprendizado de máquina supervisionado que é usado tanto para classificação quanto para regressão. Ele trabalha encontrando o hiperplano de separação que melhor divide os dados de treinamento em diferentes classes.

Tweedie Regression - algoritmo de regressão usado para lidar com dados que exibem uma distribuição Tweedie, que é uma generalização da distribuição Poisson, da distribuição gama e distribuição Wald.

XGBoost - algoritmo baseado em Gradient Boosting aplicado a Árvores de Decisão. Se destaca por utilizar método de divisão de decisão baseado em histogramas, que realiza o agrupamento dos parâmetros de entrada e utiliza esses agrupamentos para decisão

Apêndice A

Tabela A: Ordens com no máximo 10 representantes no conjunto de dados utilizado para treino e avaliação do preditor de capacidade de fixação de nitrogênio.

Ordem	Quantidade de espécies
Propionibacteriales	10
Cellvibrionales	9
Xanthomonadales	9
Aeromonadales	9
Micromonosporales	9
Sphingomonadales	8
Pseudonocardiales	8
Desulfuromonadales	8
Thermoanaerobacterales	7
Desulfovibrionales	7
Spirochaetales	6
Thiotrichales	6
Actinomycetales	6
Rhodocyclales	6
Desulfobacterales	5
Streptosporangiales	5
Sphingobacteriales	5
Chromatiales	4
Bifidobacteriales	4
Cytophagales	4
Thermotogales	3
Chitinophagales	3
Myxococcales	3
Nitrosomonadales	3
Acidithiobacillales	3
Veillonellales	3

Nitrospirales	2
Bacteroidetes Order II. Incertae sedis	2
Frankiales	2
Deinococcales	2
Acidiferrobacterales	2
Methylococcales	2
Fusobacteriales	2
Methylacidiphilales	2
Selenomonadales	2
Candidatus Brocadiales	2
Thermales	2
Chlorobiales	2
Eggerthellales	2
Acidimicrobiales	1
Erysipelotrichales	1
Marinilabiliales	1
Nevskiales	1
Chrysiogenales	1
Petrotogales	1
Nautiliales	1
Anaerolineales	1
Caulobacterales	1
Ferrovales	1
Deferribacterales	1
Chroococcales	1
Dehalococcoidales	1
Rubrobacterales	1
Hydrogenophilales	1
Aquificales	1
Bacteriovoracales	1
Thermoflexales	1

Apêndice B

Aqui são apresentados os 20 grupos ortólogos mais influentes e seus fatores de influência para o modelo treinado utilizando SVM com representação binária de COGs e eNOGs (Figura B.1) e para o modelo treinado utilizando SVM com representação quantitativa de COGs e eNOGs (Figura B.2).

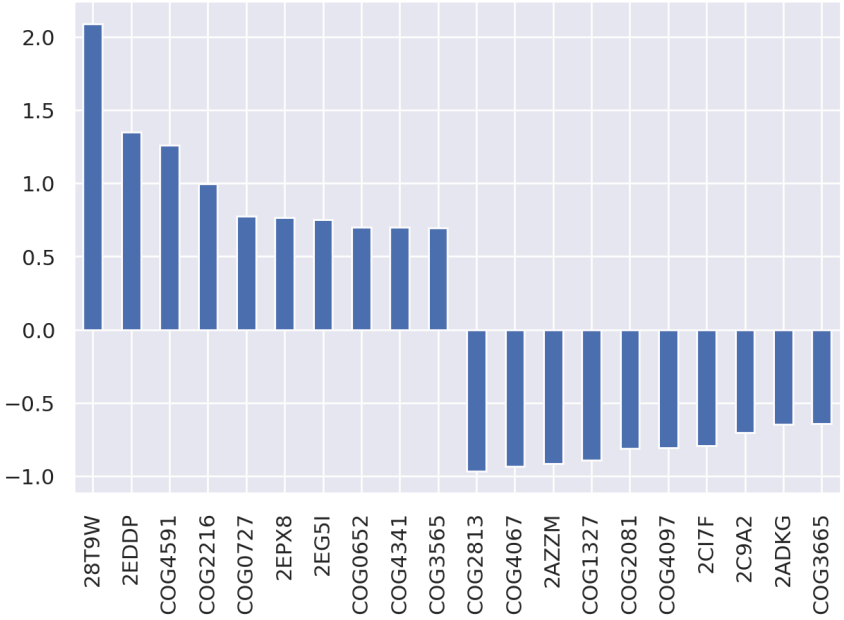

Figura B.1: 20 grupos ortólogos mais influentes para predição de fixação de nitrogênio no modelo treinado com SVM utilizando representação binária de COGs e eNOGs.

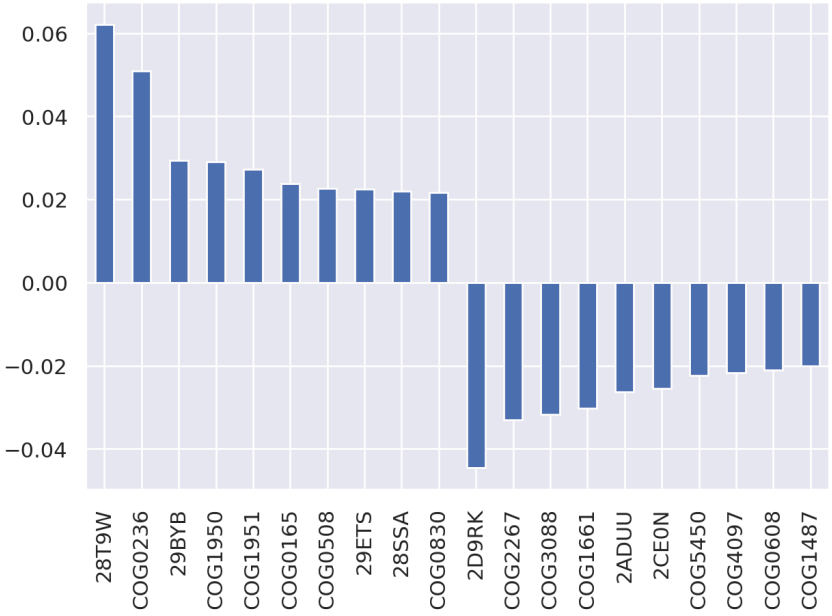

Figura B.2: 20 grupos ortólogos mais influentes para predição de fixação de nitrogênio no modelotreinado com SVM utilizando representação quantitativa de COGs e eNOGs.

Apêndice C

Tabela C: Ordens com no máximo 10 representantes no conjunto de dados utilizado para treino e avaliação do preditor de faixas de pH toleradas.

Ordem	Quantidade de espécies
Pirellulales	10
Deinococcales	10
Campylobacterales	9
Terriglobales	8
Pseudonocardiales	7
Aeromonadales	7
Rhodocyclales	6
Streptosporangiales	6
Chromatiales	6
Desulfobacterales	6
Spirochaetales	5
Desulfovibrionales	5
Methylococcales	4
Desulfuromonadales	4
Nitrosomonadales	4
Beggiatoales	4
Vibrionales	4
Dehalococcoidales	3
Actinomycetales	3
Aquificales	3
Thermoanaerobacterales	3
Synergistales	3
Nautiliales	3
Acidimicrobiales	3
Verrucomicrobiales	2
Thermotogales	2

Frankiales	2
Bifidobacteriales	2
Isosphaerales	2
Rubrobacterales	2
Ignavibacteriales	2
Erysipelotrichales	2
Kosmotogales	2
Orbales	2
Acidithiobacillales	2
Caulobacterales	2
Thermodesulfobacteriales	2
Deferribacterales	2
Emcibacterales	1
Nakamurellales	1
Acidiferrobacterales	1
Thermosediminibacterales	1
Nevskiales	1
Sphaerobacterales	1
Myxococcales	1
Rhodothermales	1
Chloroflexales	1
Caldisericales	1
Coriobacteriales	1
Epsilonproteobacteria	1
Catenulisporales	1
Dictyoglomerales	1
Mesoaciditogales	1
Veillonellales	1
Chthonomonadales	1
Magnetococcales	1
Vicinamibacterales	1

Ktedonobacterales	1
Puniceicoccales	1
Jiangellales	1
Petrotogales	1
Parvularculales	1
Chrysiogenales	1
Gemmatales	1
Thermosulfidibacterales	1

Apêndice D

Tabela D: Ordens com no máximo 10 representantes no conjunto de dados utilizado para treino e avaliação do preditor de capacidade de endoesporulação.

Ordem	Quantidade de espécies
Cytophagales	9
Rickettsiales	9
Cellvibrionales	8
Veillonellales	7
Desulfuromonadales	7
Erysipelotrichales	6
Caulobacterales	6
Desulfovibrionales	6
Streptosporangiales	6
Chlamydiales	6
Thiotrichales	6
Rhodocyclales	5
Thermoanaerobacterales	5
Deinococcales	5
Coriobacteriales	4
Thermotogales	4
Tissierellales	4
Acidithiobacillales	4
Myxococcales	4
Chitinophagales	3
Legionellales	3
Thermales	2
Methylacidiphilales	2
Nitrosomonadales	2
Verrucomicrobiales	2
Acidaminococcales	2
Frankiales	2
Eggerthellales	2

Selenomonadales	2
Rubrobacterales	2
Planctomycetales	2
Oscillatoriales	2
Bacteroidetes Order II. Incertae sedis	2
Desulfobacterales	2
Chlorobiales	2
Methylococcales	2
Chromatiales	2
Entomoplasmatales	2
Acholeplasmatales	2
Cardiobacteriales	1
Nevskiales	1
Brachyspirales	1
Kosmotogales	1
Synergistales	1
Petrotogales	1
Chrysiogenales	1
Marinilabiliales	1
Halanaerobiales	1
Acidimicrobiales	1
Acidobacteriales	1
Chroococcales	1
Nitrospirales	1
Candidatus Brocadiales	1
Saprospirales	1
Bdellovibrionales	1
Deferribacterales	1
Aquificales	1
Bacteriovoracales	1
Dehalococcoidales	1
Thermoflexales	1