UNIVERSIDADE FEDERAL DE MINAS GERAIS
Escola de Engenharia
Programa de Pós-Graduação em Engenharia Elétrica

Rafael Lopes Almeida

Estimação de Atividade e Propriedades Físico-Químicas de Compostos Antibacterianos Utilizando Aprendizado Profundo

Belo Horizonte

2023

Rafael Lopes Almeida

Estimação de Atividade e Propriedades Físico-Químicas de Compostos Antibacterianos Utilizando Aprendizado Profundo

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Frederico Gualberto Ferreira Coelho
Coorientador: Prof. Dr. Vinícius Gonçalves Maltarollo

Belo Horizonte
2023

Almeida, Rafael Lopes.
A447e Estimação de atividade e propriedades físico-químicas de compostos antibacterianos utilizando aprendizado profundo [recurso eletrônico] / Rafael Lopes Almeida. - 2023.
1 recurso online (51 f. : il., color.) : pdf.

Orientador: Frederico Gualberto Ferreira Coelho.
Coorientador: Vinícius Gonçalves Maltarollo.

Dissertação (mestrado) - Universidade Federal de Minas Gerais, Escola de Engenharia.

Bibliografia: f. 18-32.
Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia elétrica - Teses. 2. Redes neurais (Computação) - Teses. 3. Fármacos - Teses. 4. Aprendizado profundo - Teses. 5. Bactérias - Teses. I. Coelho, Frederico Gualberto Ferreira. II. Maltarollo, Vinícius Gonçalves. III. Universidade Federal de Minas Gerais. Escola de Engenharia. IV. Título.

CDU: 621.3(043)

Ficha catalográfica elaborada pela bibliotecária Ângela Cristina Silva CRB/6 2361
Biblioteca Prof. Mário Werneck, Escola de Engenharia da UFMG

UNIVERSIDADE FEDERAL DE MINAS GERAIS
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

FOLHA DE APROVAÇÃO

"ESTIMAÇÃO DE ATIVIDADE E PROPRIEDADES FÍSICO-QUÍMICAS DE COMPOSTOS ANTIBACTERIANOS UTILIZANDO APRENDIZADO PROFUNDO"

RAFAEL LOPES ALMEIDA

Dissertação de Mestrado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Mestre em Engenharia Elétrica. Aprovada em 29 de setembro de 2023. Por:

Prof. Dr. Frederico Gualberto Ferreira Coelho
DELT (UFMG) - Orientador

Prof. Dr. Vinícius Gonçalves Maltarollo
PFA (UFMG) - Coorientador

Prof. Dr. João Paulo Ataide Martins
DQ (UFMG)

Prof. Dr. Antônio de Pádua Braga
DELT (UFMG)

Documento assinado eletronicamente por **Frederico Gualberto Ferreira Coelho**, **Professor do Magistério Superior**, em 29/09/2023, às 10:39, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Antonio de Padua Braga**, **Membro**, em 29/09/2023, às 10:43, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Vinicius Gonçalves Maltarollo**, **Usuário Externo**, em 29/09/2023, às 11:24, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **João Paulo Ataide Martins**, **Professor do Magistério Superior**, em 29/09/2023, às 16:52, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2663142** e o código CRC **B5C459DE**.

AGRADECIMENTOS

Em primeiro lugar, gostaria de expressar minha profunda gratidão à minha mãe por seu incentivo e por tornar possível a minha educação. Quero agradecer sinceramente por todos os sacrifícios que você fez para me proporcionar oportunidades educacionais significativas. À minha irmã, agradeço pelo seu constante companheirismo e apoio nos momentos difíceis.

Além disso, gostaria de agradecer meus professores da graduação, que foram fundamentais ao me encorajarem a buscar um curso de pós-graduação. Em especial, gostaria de mencionar o Prof. Luiz Melk de Carvalho, o Prof. Rogério Rodrigues Lima e a Profa. Vanessa Cristina Lopes Santos por sua influência inspiradora.

Não posso deixar de expressar minha gratidão a todos os meus amigos que estiveram ao meu lado ao longo desta jornada e torceram pelo meu sucesso. Quero destacar meu amigo de laboratório, Gabriel Corrêa Veríssimo, pelo apoio incansável em todos os processos, trabalhos e apresentações durante o curso.

Agradeço ao meu orientador, o Prof. Frederico Gualberto Ferreira Coelho, e ao meu coorientador, o Prof. Vinícius Gonçalves Maltarollo, pelo seu comprometimento e paciência inabaláveis. Além disso, quero agradecer a toda a equipe do LITC por criar um ambiente de aprendizado enriquecedor e à equipe do MMLab por compartilhar conhecimentos valiosos que expandiram os horizontes da minha compreensão.

Por fim, agradeço a toda a equipe do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais, por me proporcionar a valiosa oportunidade de estudar em uma das instituições de ensino mais conceituadas do país.

"It is not knowledge, but the act of learning, not the possession of but the act of getting there, which grants the greatest enjoyment."

Johann Carl Friedrich Gauss

RESUMO

Uma das partes vitais para a manutenção da saúde e bem-estar das pessoas é a pesquisa e desenvolvimento de novos medicamentos. A busca por compostos inovadores e o reposicionamento de compostos existentes possibilita o tratamento de doenças e melhoram a qualidade de vida. Contudo, o desenvolvimento de novos fármacos é um processo demorado (levando até 10 anos) e de custo elevado (custando até 3 bilhões de dólares). Nesse contexto, uma classe de fármacos que exige uma demanda urgente de novos desenvolvimentos são os antibacterianos, devido ao grande crescimento de resistência a antibióticos por parte das bactérias. Infecções por bactérias resistentes causam maiores custos médicos, internações prolongadas e aumento da mortalidade. Ferramentas computacionais compreendem abordagens que, além de acelerar e diminuir os custos do processo, mitigam o avanço de doenças, incluindo infecções causadas por bactérias resistentes. Esse auxílio computacional é empregado de modo a automatizar testes e reduzir o número de compostos necessários nos testes pré-clínicos e nas fases clínicas iniciais (fases de maiores índices de descontinuação), focando os recursos nas amostras mais promissoras. Assim, o objetivo deste trabalho é propor modelos que utilizem aprendizado profundo, para estimar atividades antibacterianas e diversos parâmetros físico-químicos de substâncias com o intuito de descobrir potenciais antibacterianos. Foram coletados dados da atividade biológica de fármacos em quatro bactérias Gram-negativas (*Escherichia coli*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* e *Salmonella typhimurium*) e de três propriedades físico-químicas (solubilidade em água, solubilidade em dimetilsulfóxido e lipofilicidade). Foi utilizada a arquitetura de Rede Neural de Grafos para abordar tarefas de classificação e regressão. Os modelos passaram pelo processo de otimização dos hiperparâmetros dos modelos. Dentre outras métricas de validação avaliadas dos modelos, os resultados alcançados demonstraram um coeficiente de correlação de Matthews acima de 0,20, para modelos de classificação, e um coeficiente de determinação acima de 0,85, para os modelos de regressão. Os compostos com atividade antibacteriana e propriedades físico-químicas mais promissores poderão ser avaliados experimentalmente como potenciais antibacterianos.

Palavras-chaves: Planejamento de Fármacos; Resistência Bacteriana; Aprendizado Profundo; Rede Neural de Grafos.

ABSTRACT

One of the essential elements for maintaining people's health and well-being is the research and development of new drugs. Pursuing innovative compounds and repositioning existing compounds enable the treatment of diseases and improve the quality of life. However, developing new drugs is time-consuming (up to 10 years) and expensive (costing up to 3 billion dollars). In this context, a class of drugs that requires an urgent demand for new developments are antibacterials due to bacteria's significant growth of resistance to antibiotics. Resistant bacterial infections cause higher medical costs, prolonged hospital stays, and increased mortality. Computational tools include approaches that, in addition to accelerating and reducing process costs, mitigate the spread of diseases, including infections caused by resistant bacteria. This computational assistance is used to automate tests and reduce the number of compounds needed in preclinical tests and the initial clinical phases (phases with higher discontinuation rates), focusing resources on the most promising samples. Thus, this study aims to propose models that employ deep learning to predict antibacterial activities and various physicochemical parameters of substances to uncover potential antibacterial agents. Biological activity data of drugs against four Gram-negative bacteria (*Escherichia coli*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa*, and *Salmonella typhimurium*) were collected, along with three physicochemical properties (water solubility, Dimethyl sulfoxide solubility, and lipophilicity). Graph Neural Network architecture was employed to address classification and regression tasks. The models underwent a process of hyperparameter optimization. Among other validation metrics evaluated for the models, the results demonstrated a Matthews correlation coefficient exceeding 0.20 for classification models and a coefficient of determination above 0.85 for regression models. Compounds with antibacterial activity and more promising physicochemical properties may be experimentally evaluated as potential antibacterials.

Keywords: Drug Design; Antibacterial Resistance; Deep Learning; Graph Neural Network.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

DMSO	Dimetilsulfóxido
EUA	Estados Unidos da América
FDA	*Food and Drug Administration*
FN	Falsos Negativos
FP	Falsos Positivos
GNN	*Graph Neural Network* ou Redes neurais de grafos
LITC	Laboratório de Inteligência Computacional
logP	Logaritmo do coeficiente de partição octanol-água
MCC	*Matthews correlation coefficient* ou Coeficiente de correlação de Matthews
MMLab	Laboratório de Modelagem Molecular
MSE	*Mean square error* ou Erro quadrático médio
nM	Nano molar
RMSE	*Root mean square error* ou Raiz quadrada do erro quadrático médio
VN	Verdadeiros Negativos
VP	Verdadeiros Positivos

LISTA DE SÍMBOLOS

y_i	Valor observado
$\overline{y}$	Média dos valores observados
$\hat{y}_i$	Valor predito
$\overline{\hat{y}}$	Média dos valores preditos
$\pm$	Desvio padrão

SUMÁRIO

1 INTRODUÇÃO

A inteligência artificial e o aprendizado de máquina têm se tornado cada vez mais presentes em nossa vida cotidiana. Podemos interagir com nossos dispositivos usando comandos de voz, tornando tarefas simples mais ágeis e práticas. Além disso, essas tecnologias estão por trás das recomendações personalizadas em serviços de *streaming* e redes sociais, nos ajudando a descobrir novos conteúdos e produtos com base em nossas preferências.

Existem também aplicações no campo farmacêutico e um dos principais empregos é no planejamento de fármacos. Com a análise de grandes volumes de dados sobre moléculas, proteínas e doenças, os algoritmos de aprendizado de máquina podem identificar padrões e relações complexas que os cientistas talvez não conseguissem identificar facilmente sem o auxílio de ferramentas apropriadas. A utilização de inteligência artificial acelera o processo de pesquisa e pode levar à modelagem mais rápida e precisa de fármacos e medicamentos inovadores.

O desenvolvimento de fármacos é de vital importância para a sociedade. Essas substâncias são cruciais para tratar e prevenir doenças. A busca por compostos inovadores promove um futuro mais saudável para a sociedade, de forma que a evolução das técnicas existentes leva novos compostos que combatem enfermidades antes consideradas incuráveis e ajuda a mitigar os problemas associados com a resistência a medicamentos, como as bactérias resistentes a antibióticos.

1.1 Motivação

O processo para aprovar novos fármacos é um processo complexo, às vezes necessitando de buscas extensivas para obter compostos que apresentam boas propriedades

O composto selecionado inicialmente precisa passar por vários processos e testes, fazendo com que o tempo médio de desenvolvimento de um novo fármaco seja aproximadamente de 10 anos e tendo um custo que pode chegar de 2 a 3 bilhões de dólares (DAS et al., 2021).

Apesar dos grandes avanços tecnológicos e novas técnicas para auxiliar no processo de descoberta de novos fármacos, de acordo com Stokes et al. (2020) a produtividade do desenvolvimento de novos fármacos diminuiu, de forma que a quantidade de novas entidades químicas aprovadas pela *Food and Drug Administration* (FDA) não aumentou rapidamente acompanhando as inovações tecnológicas (KADURIN et al., 2017).

Doenças bacterianas podem afetar a população de diversas formas. Elas podem ser

transmitidas por meio do contato físico, consumo de alimentos e água contaminados, por contato com animais ou ambientes infectados e equipamentos médicos inadequadamente esterilizados também podem ser uma fonte de infecção.

O desenvolvimento de resistência a antibióticos das bactérias leva a outro fator de grande impacto na sociedade. Geralmente, infecções por bactérias resistentes causam maiores custos médicos, internações prolongadas e aumento da mortalidade. Além dos problemas ocasionados diretamente pela bactéria resistente, os pacientes infectados com essas bactérias são mais susceptíveis a outras doenças infecciosas do pulmão como gripe e síndrome respiratória aguda (CDC, 2019; DAS et al., 2021).

De acordo com CDC (2019), foi estimado que em 2019 nos EUA mais de 2,8 milhões de infecções causadas por bactérias resistentes ocorreram, e mais de 35 mil pessoas morreram por esta causa. Globalmente, esse número pode chegar a 700 mil mortes. É estimado que, no ano de 2050, as mortes por bactérias resistentes aumentarão para cerca de 10 milhões (DAS et al., 2021).

Neste contexto, é essencial o desenvolvimento contínuo de novos antibacterianos para enfrentar a crescente ameaça das bactérias resistentes, garantindo assim opções eficazes de tratamento para combater infecções.

Durante o desenvolvimento de novos fármacos, além de demonstrar atividade contra alvos específicos, como doenças ou bactérias, outros parâmetros desempenham um papel fundamental na garantia da eficácia desses medicamentos quando utilizados em seres humanos. A solubilidade em água e a lipofilicidade são dois parâmetros cruciais na formulação de fármacos, pois influenciam diretamente a absorção, distribuição, metabolismo e eliminação do composto. É de extrema importância que essas propriedades sejam bem compreendidas, garantindo que a molécula apresente características favoráveis para possibilitar a formulação eficaz do medicamento.

1.2 Objetivo

O objetivo deste trabalho é propor modelos que utilizam aprendizado profundo para estimar diversos parâmetros físico-químicos de compostos e sua atividade contra bactérias Gram-negativas. Os parâmetros físico-químicos são: solubilidade em água, solubilidade em dimetilsulfóxido, lipofilicidade.

1.3 Organização do Texto

O restante dos capítulos são organizados da seguinte forma:

O capítulo 2 apresenta conceitos relacionados com o processo de planejamento de

fármacos e suas especificidades, como os desafios enfrentados, ferramentas computacionais, os alvos biológicos abordados neste trabalho e os parâmetros físico-químicas de interesse.

O capítulo 3 apresenta os conceitos relacionados com aprendizado de máquina, como o tipo de tarefa envolvida no aprendizado supervisionado, aprendizado profundo, grafos e a utilização de rede neural de grafos, tratamento de dados desbalanceados, explicabilidade de modelos de aprendizado de máquina e as métricas utilizadas para a avaliação dos modelos.

O capítulo 4 descreve os materiais e métodos propostos, bem como os dados utilizados, o pré-processamento utilizado nos dados e a definição dos experimentos a serem realizados.

O capítulo 5 apresenta os resultados obtidos após a realização dos experimentos descritos no capítulo 4 e por fim o capítulo 6 fornece as considerações conclusivas acerca do trabalho juntamente com algumas limitações encontradas e passos futuros.

2 PLANEJAMENTO DE FÁRMACOS

O planejamento de fármacos é um processo importante para a manutenção da saúde, identificando e modelando novos fármacos, substâncias químicas com uso terapêutico, para combater doenças. O principal objetivo desta área é encontrar fármacos seguros e eficazes que possam atuar nesses alvos, proporcionando benefícios terapêuticos aos pacientes. Esse processo engloba principalmente uma abordagem sistemática para identificar alvos biológicos específicos associados a doenças e procurar agentes químicos ou biológicos que possam interagir com eles (DOYTCHINOVA, 2022).

Compostos com características promissoras para utilização como fármacos são encontrados em ecossistemas diversos da natureza, como em animais, plantas e fungos. Inicialmente, eles eram descobertos acidentalmente, como no caso da penicilina, mas hoje em dia eles são selecionados por critérios mais racionais. Durante o processo de planejamento de fármacos tradicional, são realizados testes na tentativa e erro, que é um processo impreciso, demorado e poucos compostos são descobertos (LI et al., 2021).

2.1 Planejamento de fármacos auxiliado por computador

As estratégias de planejamento de fármacos auxiliado por computador complementam as técnicas experimentais tradicionais de planejamento de fármacos, acelerando o processo por reduzir o número de compostos a serem sintetizados e testados. Estas estratégias podem ser divididas em dois grandes grupos: Planejamento de Fármacos Baseado na Estrutura do Alvo Molecular e Planejamento de Fármacos Baseado em Ligantes (VEMULA et al., 2023; LO et al., 2018).

Conforme Macalino et al. (2015), o Planejamento de Fármacos Baseado em Estrutura corresponde ao emprego de conhecimentos prévios obtidos sobre a estrutura tridimensional do sítio de ligação de uma determinada proteína. Um dos métodos mais comuns deste grupo é o *docking* molecular, que prevê como um composto pode se ligar a um sítio específico e estima sua afinidade de acordo com sua conformação e interação com resíduos do sítio de ligação.

Já o Planejamento de Fármacos Baseado em Ligantes, segundo Ece (2023), se baseia apenas nas propriedades de compostos conhecidos e o efeito que possuem sobre um determinado alvo molecular, teste celular (por ex.: atividade contra bactérias, atividade citotóxica contra células cancerosas) ou atividades em modelos *in vivo*. Ao estudar a estrutura e o efeito desses compostos podemos identificar algumas características e desenvolver novos compostos, que possuem características semelhantes. Uma das técnicas

mais utilizadas é a relação quantitativa estrutura-atividade, que utiliza métodos estatísticos ou algoritmos para identificar e quantificar relações entre a estrutura molecular e o efeito biológico.

A utilização de técnicas de aprendizado de máquina no planejamento de fármacos auxiliado por computador tem o potencial de otimizar o planejamento de novos fármacos reduzindo custos e a quantidade de compostos em testes. Essas reduções se dão por meio da identificação de compostos promissores, estimativa da atividade biológica de compostos e a análise de grande quantidade de dados rapidamente. As técnicas de aprendizado de máquina podem ser aplicadas, por exemplo, para identificar novos alvos para fármacos, prever a atividade biológica e a toxicidade de medicamentos (VAMATHEVAN et al., 2019; SCHNEIDER et al., 2020; CHEN et al., 2018).

Um exemplo de aplicação de técnicas de aprendizado de máquina é a triagem virtual. De acordo com Junior et al. (2019), é um processo computacional fundamental na pesquisa de novos medicamentos. Ao invés de testar experimentalmente inúmeros compostos para verificar propriedades físico-químicas e atividades biológicas, a triagem virtual permite que sejam realizadas previsões computacionais sobre estes compostos por meio das predições realizadas.

2.2 Infecções bacterianas

As bactérias Gram-negativas são microrganismos que possuem uma membrana externa composta por lipopolissacarídeos. Elas podem ser benéficas, atuando em funções do corpo humano, por exemplo, ou maléficas causando infecções do trato urinário, infecções abdominais, pneumonia e meningite (ALFEI; SCHITO, 2020).

A resistência a antibióticos ocorre quando a bactéria muda em resposta ao uso desse medicamento, mitigando a ação do fármaco. Ao longo dos anos foi observado o desenvolvimento da resistência aos novos compostos antibióticos que chegavam ao mercado em um curto período de tempo (SERAFIM et al., 2020).

Essa resistência é um grande problema, pois representa uma ameaça significativa à saúde pública ao reduzir a eficácia dos medicamentos utilizados, que são essenciais para o tratamento de doenças infecciosas, resultando em internações mais longas, gastos com saúde maiores e elevação das taxas de mortalidade (NADEEM et al., 2020).

Existe uma portaria emitida pelo Ministério da Saúde do Brasil que classifica os riscos associados a agentes biológicos, as bactérias são classificadas entre classe 1 até classe 4, sendo o nível 4 a que demonstra maior risco para o indivíduo e para sociedade. As seguintes bactérias foram selecionadas desta portaria para o desenvolvimento do trabalho: *Escherichia coli*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* e *Salmonella typhimurium*. Todas

essas bactérias são da Classe de risco 2, que significa que o risco individual é moderado e o risco para a comunidade é limitado (BRASIL, 2021).

2.3 Propriedades de interesse

Neste trabalho foram investigados dois tipos de propriedades de interesse. A primeira são propriedades de interesse físico-químico e a segunda é a atividade biológica.

As propriedades físico-químicas são características dos compostos que descrevem como eles interagem com o ambiente físico e químico ao seu redor. Essas propriedades são fundamentais para entender o comportamento das substâncias em várias situações, incluindo reações químicas, processos de dissolução, transporte em sistemas biológicos, entre outros (GUO et al., 2021).

No planejamento de fármacos a otimização dessas propriedades físico-químicas é um desafio importante. Os pesquisadores buscam desenvolver compostos que possuam o equilíbrio de todas propriedades de interesse para que os medicamentos sejam eficazes, seguros e viáveis para desenvolvimento clínico. A predição dessas propriedades pode ajudar no processo de desenvolvimento acelerando testes (PANTALEãO et al., 2022).

Já a propriedade de atividade biológica de uma substância é essencial para garantir que ela seja direcionada ao alvo correto e que tenha o efeito farmacológico desejado.

2.3.1 Solubilidade em água

A solubilidade em água refere-se à capacidade de uma substância se dissolver em água. Essa propriedade é de grande importância pois influencia diretamente na absorção, distribuição e na eficácia de um composto no organismo, uma vez que muitos processos biológicos ocorrem em meios aquosos (WALDEN et al., 2021).

Conforme Walden et al. (2021), a solubilidade em água é um fator crítico na formulação/administração e na absorção de um fármaco, de forma que um composto precisa ser solúvel o suficiente em água para ser absorvido pela corrente sanguínea. Compostos pouco solúveis podem exigir formulações especiais, como suspensões ou formulações com coadjuvantes para melhorar a solubilidade e a absorção.

2.3.2 Solubilidade em Dimetilsulfóxido

Dimetilsulfóxido (DMSO) é um solvente orgânico amplamente utilizado em laboratórios e na indústria farmacêutica devido às suas propriedades, como a capacidade de dissolver muitos compostos orgânicos e inorgânicos (TETKO et al., 2013).

Em alguns casos, segundo Tetko et al. (2013), o DMSO pode ser usado como coadjuvante para melhorar a solubilidade de compostos que são pouco solúveis em água. Isso pode ser útil na criação de formulações farmacêuticas específicas. Além do mais, usualmente o DMSO é empregado como co-solvente nos testes pré-clínicos iniciais de substâncias candidatas a fármacos.

2.3.3 Lipofilicidade

A lipofilicidade é uma propriedade físico-química de suma importância para o planejamento de fármacos. Ela se refere à afinidade de uma molécula por substâncias não polares (lipídicas), como óleos e gorduras, em contraposição à afinidade por substâncias polares, como a água. Esta propriedade usualmente é expressa como o logaritmo do coeficiente de partição octanol-água, o logP. Compostos com um logP positivo tendem a ser mais solúveis em gorduras, enquanto aqueles com um logP negativo tendem a ser mais solúveis em água (LEESON; YOUNG, 2015).

De acordo com Leeson e Young (2015), a lipofilicidade influencia diretamente a capacidade de um composto ser absorvido e distribuído pelo corpo. Moléculas muito hidrofílicas podem ter dificuldade em atravessar as membranas celulares lipídicas, o que pode limitar sua biodisponibilidade. Por outro lado, compostos muito lipofílicos tendem a se acumular em tecidos gordurosos e/ou atravessar a barreira hematoencefálica. Neste último caso, pode ser responsável por explicar efeitos adversos indesejáveis.

2.3.4 Atividade antibacteriana

A atividade antibacteriana é a capacidade de uma substância química, como um composto candidato a medicamento, interagir com sistemas biológicos específicos de maneira a produzir um efeito desejado. Essa atividade é fundamental para determinar a utilidade potencial de um composto na prevenção, tratamento ou diagnóstico de doenças. No presente trabalho será avaliada a capacidade de um composto inibir ou não o crescimento de bactérias (SHI; ZHAO; WEI, 2018).

2.4 Domínio de aplicabilidade

Domínio de aplicabilidade desempenha um papel crucial na avaliação e aplicação de modelos computacionais no planejamento de fármacos. Os modelos são treinados com conjuntos de dados de treinamento específicos, e cada modelo tem limitações inerentes (ROY; KAR; AMBURE, 2015).

Segundo Sahigara et al. (2012), é essencial entender o domínio de aplicabilidade de um modelo antes de usá-lo para fazer previsões. O domínio pode ser influenciado por vários

fatores, como a estrutura química dos compostos, faixa de propriedades físico-químicas relevantes, mecanismo de ação e a disponibilidade de dados experimentais.

Um modelo treinado para prever a toxicidade de compostos orgânicos pode ser válido apenas para compostos com estruturas químicas semelhantes àquelas presentes no conjunto de treinamento do modelo ou o modelo pode ser aplicável apenas a uma faixa específica de valores de propriedades físico-químicas. Então, além de treinar e validar o modelo, é crucial assegurar que as novas predições estejam dentro do domínio de aplicabilidade (GADALETA et al., 2016).

3 APRENDIZADO DE MÁQUINA

Aprendizado de máquina pode ser descrito como algoritmos que se adaptam automaticamente para alcançar seu resultado sem que sejam especificamente programados para tal. Nos algoritmos de aprendizado de máquina, em grande maioria, isso ocorre mostrando dados de entrada para treinar o algoritmo de modo que o resultado possa ser generalizado para dados que o algoritmo não foi apresentado ainda (MAIER et al., 2019).

Os primeiros algoritmos de aprendizado de máquina foram inspirados pela capacidade que o cérebro possui para aprender as informações que recebe. Um algoritmo notável proposto na época por Rosenblatt (1958) foi o perceptron, com uma estrutura análoga ao neurônio. Ao longo do tempo, a pesquisa nesse campo teve avanços e pausas, chamados de "inverno da Inteligência Artificial", mas atualmente, com os avanços tecnológicos e a crescente disponibilidade de dados, tornou viável o desenvolvimento do aprendizado profundo, que possui aplicações complexas como reconhecimento de padrões, visão computacional, processamento de linguagem natural, etc (FRADKOV, 2020).

O aprendizado de máquina pode ser dividido em diferentes categorias, sendo elas: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. Neste trabalho serão empregadas técnicas de aprendizado supervisionado (YANG et al., 2019).

3.1 Aprendizado supervisionado

O aprendizado supervisionado é uma subcategoria de aprendizado de máquina. É uma abordagem na qual o algoritmo é treinado com um conjunto de dados rotulados para aprender a mapear os atributos de entrada para a saída, que é conhecida antecipadamente. Os algoritmos de aprendizado supervisionado são frequentemente aplicados em tarefas de classificação e regressão (LECUN; BENGIO; HINTON, 2015).

Na classificação, o objetivo é atribuir uma classe ou categoria a uma determinada entrada. Por exemplo, classificar compostos como ativo ou não ativo contra uma bactéria. Na regressão, o objetivo é prever um valor numérico contínuo, como a previsão da solubilidade de um fármaco com base em suas características (UDDIN et al., 2019). A Figura 1 ilustra os dois tipos de tarefas.

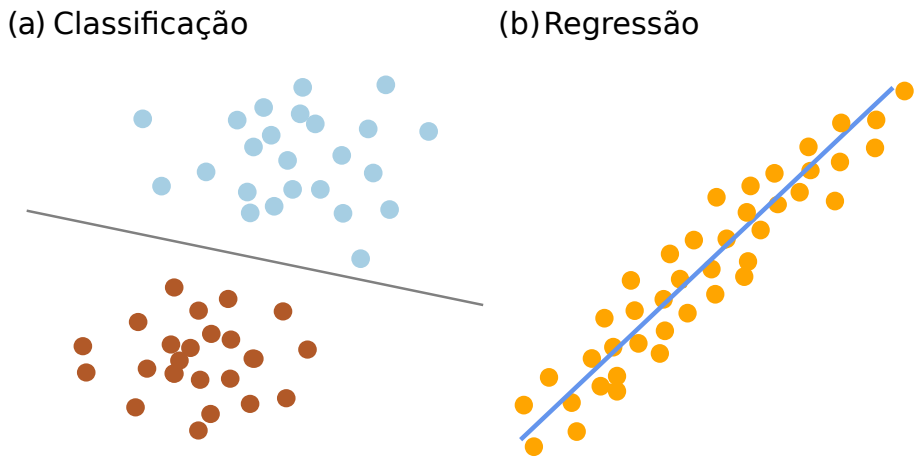

Figura 1 – Exemplo de tarefa de classificação e de regressão. Em (a), o objetivo do modelo de aprendizado de máquina é distinguir e separar as classes dos objetos (representados pelas cores diferentes) e em (b) o objetivo do modelo é ajustar uma equação que estima valores contínuos de acordo com os dados conhecidos.

3.2 Aprendizado profundo

O aprendizado profundo é uma técnica que emprega fortemente os conceitos de aprendizado de representação, possibilitando os algoritmos de receberem os dados de uma maneira mais pura e descobrir automaticamente as representações necessárias para realizar a tarefa (LECUN; BENGIO; HINTON, 2015).

De acordo com Abdel-Jaber et al. (2022), os modelos de aprendizado profundo são compostos por várias camadas que processam os dados de entrada e gradualmente extraem características abstratas e mais significativas à medida que avançam nas camadas. O aspecto principal do aprendizado profundo é que os parâmetros internos dos modelos não são especificados pelo usuário, mas são aprendidas com os dados apresentados para o algoritmo durante o processo de treinamento.

Para entender o funcionamento geral de um modelo de aprendizado profundo podemos analisar um dos primeiros modelos propostos que simulava o comportamento do neurônio humano: o Perceptron (ROSENBLATT, 1958), exemplificado na Figura 2.

Segundo Higham e Higham (2019), o Perceptron é uma função matemática, na qual os dados de entrada são multiplicados pelo peso de cada atributo, que são ajustados pela rede a cada iteração. Os resultados dessas operações são agregados por uma somatória e é aplicada uma função de ativação, que é uma operação não linear que geralmente restringe valores menores que zero, dessa maneira o neurônio só propaga a informação caso o somatório calculado dos dados de entrada exceda um determinado limite. O resultado obtido é comparado com o resultado esperado e caso seja diferente o erro é computado para ajustar os pesos.

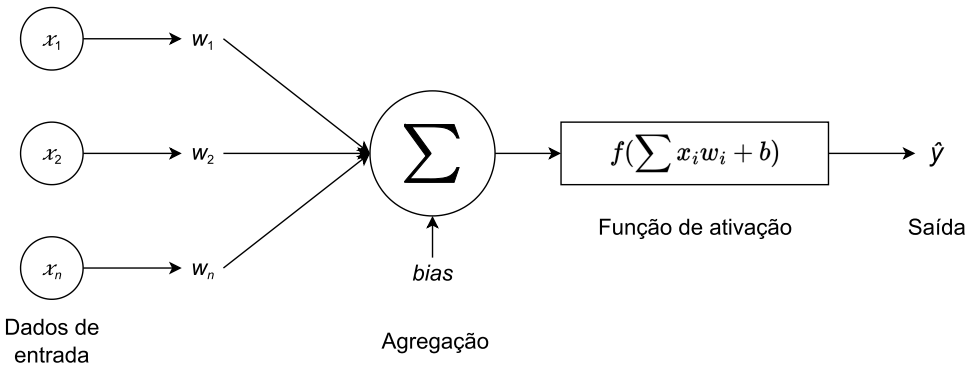

Figura 2 – Representação de um Perceptron.

No aprendizado profundo uma combinação de vários perceptrons divididos em camadas são ligados entre si para abordar problemas complexos. A Figura 3 mostra um exemplo de rede aprendizado profundo.

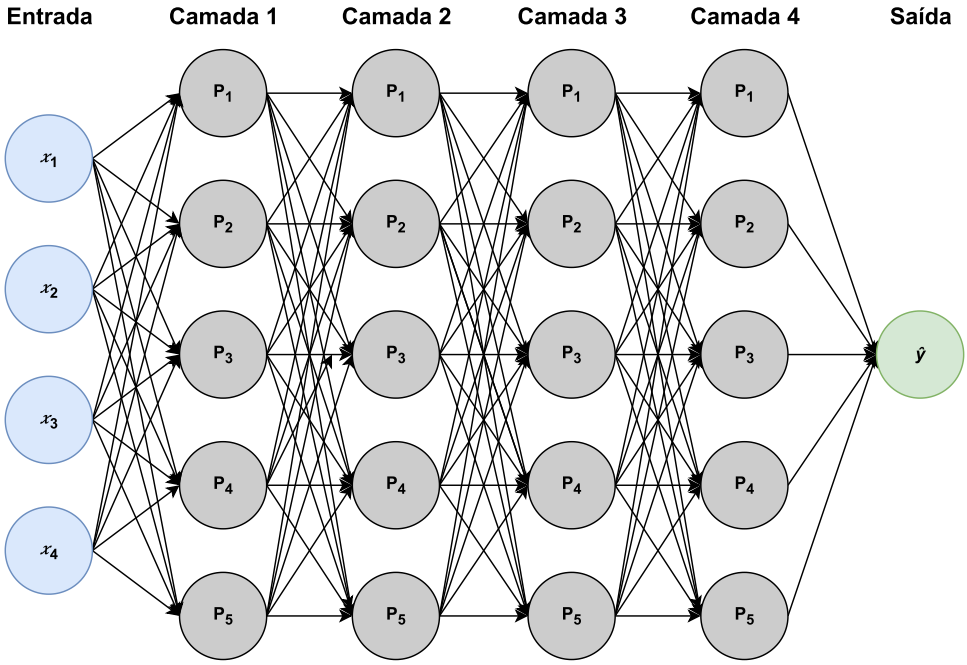

Figura 3 – Perceptron de múltiplas camadas.

Com o passar do tempo novas técnicas foram desenvolvidas e estruturas especializadas para cada tipo de problema abordado foram surgindo. Exemplos são as redes neurais convolucionais, redes neurais recorrentes ou as redes neurais de grafos, que são adaptadas para realizar operações diretamente em grafos (LECUN; BENGIO; HINTON, 2015; ZHOU et al., 2020).

3.3 Grafos

Grafos são representações matemáticas utilizadas para modelar a relação entre objetos. Estes objetos são compostos por vértices e as conexões entre eles são chamadas

de arestas. . Essas conexões podem ser não direcionadas, indicando uma ligação bilateral entre dois vértices, ou direcionadas, representando relações unilaterais mais complexas Hoang et al. (2023). A Figura 4 ilustra um exemplo de grafo.

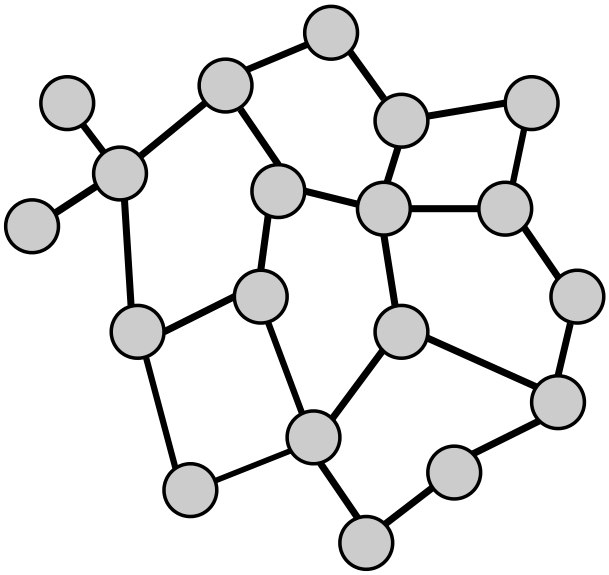

Figura 4 – Exemplo de estrutura de grafo. Em um contexto aplicado, o grafo pode representar um ciclo social onde os vértices são pessoas e as arestas relacionamentos entre elas.

Segundo Zhang, Cui e Zhu (2022), existem diversas metodologias de construção de grafos, cada uma explorando alguma relação ou propriedade de interesse. Esse tipo de representação é amplamente utilizado em diversas áreas como, análise de redes sociais, logística, sistemas de recomendação, estudo de moléculas e para modelar problemas que envolvem relações entre elementos distintos. Quando trabalhamos com moléculas, os átomos podem ser representados pelos vértices e as ligações químicas como as arestas.

Cada uma dessas duas representações (vértices ou arestas) pode ter suas características próprias, chamados de atributos. A Figura 5 mostra um exemplo de molécula representada como grafo e seus atributos.

A informação referente aos vértices de um grafo podem ser representadas por uma matriz de tamanho $N_{Vértice}$ x $N_{Atributo}$, onde $N_{Vértice}$ é o número de vértices e $N_{Atributo}$ é a quantidade de atributos relacionados. Já a informação das arestas pode ser representada por uma matriz de tamanho N_{Aresta} x $N_{Atributo}$, onde N_{Aresta} é o número de arestas e $N_{Atributo}$ é a quantidade de atributos relacionados. As informações relacionadas com a conectividade do grafo podem ser representadas utilizando uma matriz de adjacência de tamanho $N_{Vértice} \times N_{Vértice}$, onde $N_{Vértice}$ é o número de vértices. Por fim, podemos extrair uma informação do contexto global do grafo, utilizando um escalar ou um vetor Hoang et al. (2023).

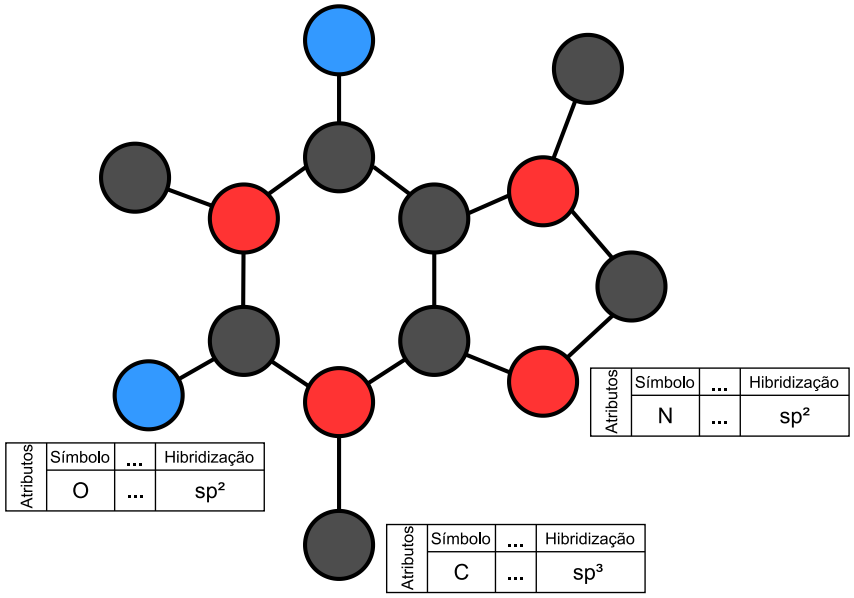

Figura 5 – Molécula de cafeína sendo representada como um grafo. Cada átomo e ligação possui seus atributos particulares.

3.4 Rede neural de grafos

Redes neurais de grafos, do inglês *Graph Neural Network* (GNN) são uma classe de modelos de aprendizado de máquina que foram projetados para lidar com dados na forma de grafos. O princípio é permitir que os vértices de um grafo capturem informações não apenas de seus vizinhos diretos, mas também de vizinhos de vizinhos e assim por diante, em um processo iterativo. Isso permite que a rede neural explore as relações complexas e padrões contidos nos dados de grafo (ASIF et al., 2021).

Uma maneira de explicar o funcionamento das redes neurais de grafos é utilizando a abordagem de rede neural de passagem de mensagem proposta por Gilmer et al. (2017). As informações são passadas entre os vértices de um grafo por meio de mensagens, de forma que cada vértice envia uma mensagem contendo seus atributos para seus vértices vizinhos e conforme essas mensagens são recebidas os vértices atualizam seus atributos incorporando as informações dos vizinhos. Esse procedimento ocorre em várias iterações, permitindo que informações de vizinhos de diferentes níveis de distância sejam consideradas. Podemos dividir esse processo em três fases: agregação, atualização e saída. A Figura 6 ilustra esse processo.

Conforme Gilmer et al. (2017), as mensagens contendo os atributos são enviadas para os vértices vizinhos e elas precisam ser agregadas de alguma forma para sumarizar a informação da vizinhança e manter o mesmo tamanho dos atributos originais. Esse processo também é importante para garantir a invariância de permutação das mensagens, pois grafos não possuem uma ordem inerente. Esse processo geralmente é realizado com uma operação de soma, máxima ou média nas mensagens.

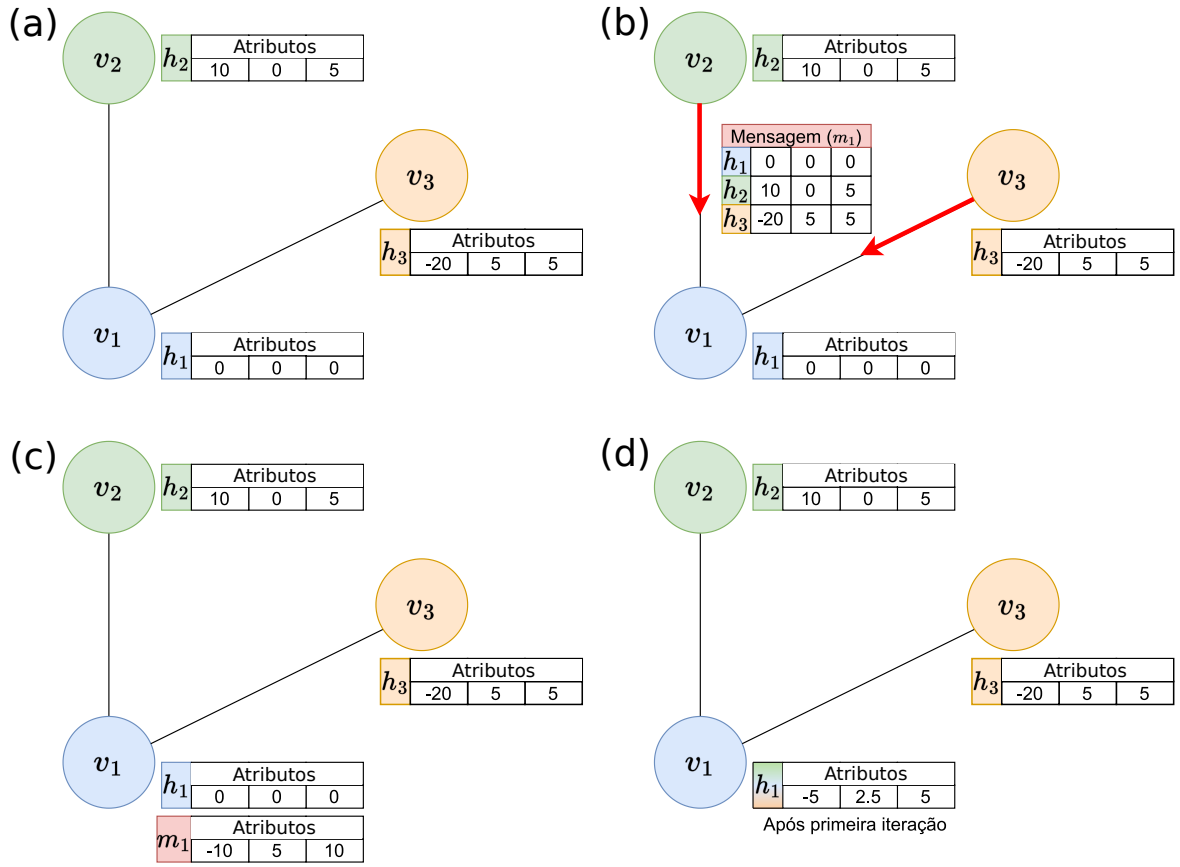

Figura 6 – Exemplo do funcionamento de uma rede neural de passagem de mensagem. Primeiramente, o vértice v_i é escolhido em (a) para ser analisado. Em (b) é realizada a passagem dos atributos dos vértices vizinhos e em (c) a mensagem é agregada com um somatório em m_1. Em (d) o valor de h_1 é atualizado utilizando a operação de média entre o valor inicial de h_1 e m_1.

Na fase de atualização o vértice atualiza seus atributos com base nas mensagens que foram agregadas. Essa atualização pode ser realizada usando operações lineares, não lineares ou outras operações relevantes. O objetivo é permitir que o vértice ajuste seus atributos com base nas informações do passo anterior.

A fase de saída é onde o modelo combina todas as informações do grafo para obter uma representação final ou uma resposta para todo o grafo. Isso é importante porque, em muitos casos, estamos interessados não apenas nas informações em cada vértice individual do grafo, mas também em entender o grafo como um todo. Para realizar esse processo podemos combinar as informações dos vértices e suas conexões com operadores (soma, média, etc) ou por meio de técnicas mais avançadas como agrupamento de grafos (do inglês, *graph pooling*).

3.5 Estratégias para dados desbalanceados

Conjunto de dados desbalanceados apresentam uma distribuição de classes não uniforme. Isso significa que uma classe possui consideravelmente mais dados do que a outra. Esse desequilíbrio na distribuição é problemático, pois pode prejudicar o aprendizado da classe minoritária (WANG et al., 2021). Existem técnicas que podemos empregar para mitigar esse problema. O ajuste de peso da função de perda e a transferência de aprendizado são algumas delas. A Figura 7 ilustra os dois casos.

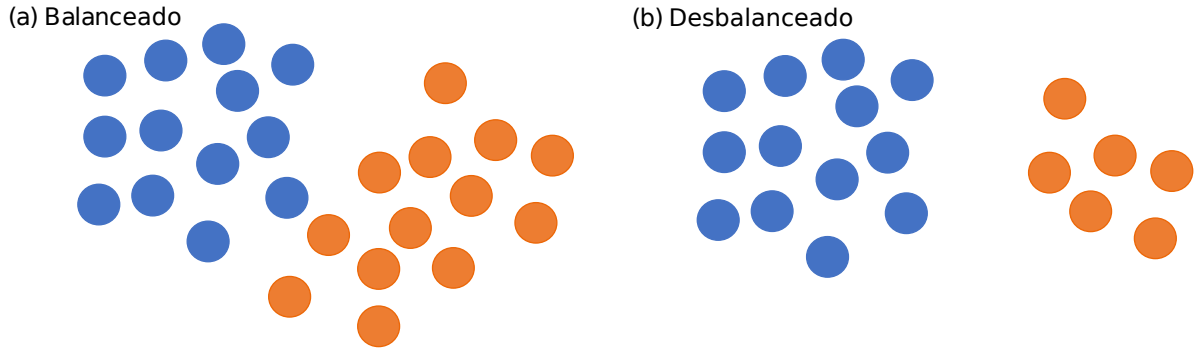

Figura 7 – Exemplos de conjunto de dados. Em (a) os dados são totalmente balanceados, e em (b) os dados estão desbalanceados, prejudicando a classe laranja.

O ajuste de pesos da função de perda consiste em atribuir pesos diferentes durante o treinamento do modelo, atribuindo maior importância à classe minoritária. Isso é feito para dar mais importância à classe minoritária e ajudar o modelo a aprender melhor a partir dos dados desbalanceados (JADON, 2020).

As técnicas de transferência de aprendizado e pré-treinamento também são eficazes para melhorar o desempenho dos modelos. Elas se baseiam no conhecimento prévio adquirido a partir de conjuntos de dados similares.

Segundo Simões et al. (2018), a transferência de aprendizado implica em reutilizar os pesos de um modelo já treinado. Em alguns casos, apenas as camadas finais do modelo são treinadas para a nova tarefa, mantendo as camadas iniciais fixas. O pré-treinamento, por outro lado, envolve o treinamento inicial em um conjunto de dados em grande escala. Posteriormente, ocorre o ajuste fino do modelo em um conjunto de dados menor e específico para a tarefa alvo.

3.6 Métricas utilizadas para validação da qualidade de predição

As métricas são fundamentais para avaliar os modelos preditivos em geral, incluindo os modelos baseados em aprendizado de máquina, fornecendo uma maneira objetiva e quantificável para medir o desempenho de um modelo em relação aos dados de teste.

3.6.1 Métricas para classificação

As métricas de acurácia, precisão, revocação, pontuação F1 são comumente empregadas para avaliar o desempenho de modelos para a tarefa de classificação e o coeficiente de correlação de Matthews, do inglês *Matthews Correlation Coefficient* (MCC) é recomendado para avaliar conjunto de dados desbalanceados. Todas estas métricas podem ser obtidas a partir da matriz de confusão (VALERO-CARRERAS; ALCARAZ; LANDETE, 2023).

De acordo com Valero-Carreras, Alcaraz e Landete (2023), a matriz de confusão é uma ferramenta visual utilizada para avaliar o desempenho de modelos de classificação. Ela apresenta uma representação resumida das previsões feitas pelo modelo em comparação com as classes verdadeiras dos dados de teste. A Figura 8 ilustra os elementos da matriz de confusão.

		Predito	
		Classe A	Classe B
Real	Classe A	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	Classe B	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Figura 8 – Exemplo de uma matriz de confusão.

A matriz de confusão é construída em torno de quatro elementos principais:

- **Verdadeiros Positivos (VP):** São os casos em que o modelo previu corretamente a classe A como classe A.
- **Falsos Positivos (FP):** São os casos em que o modelo previu incorretamente a classe B como classe A.
- **Verdadeiros Negativos (VN):** São os casos em que o modelo previu corretamente a classe B como classe B.
- **Falsos Negativos (FN):** São os casos em que o modelo previu incorretamente a classe A como classe B.

Podemos obter as métricas de acurácia, precisão, revocação, pontuação F1 e MCC por meio dos elementos principais apresentados.

A acurácia indica o desempenho geral do modelo de forma que: dentre todas as classificações realizadas, quantas o modelo classificou corretamente. Por levar em consideração todas as amostras, não é uma métrica confiável para avaliar conjuntos de dados desbalanceados. A Equação (3.1) mostra como a acurácia é calculada.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{3.1}$$

A precisão é a proporção de verdadeiros positivos em relação a todos os positivos previstos. Ajuda a entender o quão confiáveis são as previsões positivas do modelo. A Equação (3.2) demonstra como é calculada a precisão.

$$\text{Precisão} = \frac{VP}{VP + FP} \tag{3.2}$$

A revocação é a proporção de verdadeiros positivos em relação a todos os casos positivos reais e é responsável por medir a capacidade do modelo em identificar todos os casos positivos. A Equação (3.3) explica o cálculo da revocação.

$$\text{Revocação} = \frac{VP}{VP + FN} \tag{3.3}$$

A pontuação F1 é uma métrica que combina precisão e revocação em uma única medida, sendo útil quando é preciso equilibrar a importância de ambas as métricas. A Equação (3.4) esclarece como se determina a pontuação F1.

$$\text{F1} = \frac{2 \times \text{Precisão} \times \text{Revocação}}{\text{Precisão} + \text{Revocação}} \tag{3.4}$$

Conforme Chicco, Tötsch e Jurman (2021), o MCC é uma métrica que considera todos os aspectos da matriz de confusão e fornece uma medida geral do quão bem um modelo de classificação está realizando a tarefa, independentemente do desequilíbrio entre as classes ou do tamanho da amostra. O MCC varia entre -1 e $+1$, sendo quanto mais próximo de $+1$ o MCC, melhor o desempenho do modelo. A Equação (3.5) descreve o método para calcular o MCC.

$$\text{MCC} = \frac{VP \times VN - FP \times FN}{\sqrt{(VP + FP)(VP + FN)(VN + FP)(VN + FN)}} \tag{3.5}$$

3.6.2 Métricas para regressão

As métricas de raiz quadrada do erro quadrático médio (RMSE), do inglês *Root mean square error* (RMSE) e coeficiente de determinação (R^2) são comumente empregadas

para avaliar o desempenho de modelos para a tarefa de regressão. Outra métrica bastante útil é a Coeficiente de Correlação de Concordância (CCC).

O RMSE é a raiz quadrada do erro quadrático médio, do inglês *Mean square error* (MSE), que é a média dos quadrados das diferenças entre as previsões do modelo e os valores reais. O RMSE fornece uma medida do erro na unidade original da variável resposta, facilitando a interpretação. Quanto menor o RMSE, melhor o desempenho Hodson (2022). A Equação (3.6) detalha como se calcula o RMSE.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.6}$$

Onde n é a quantidade de dados, y_i é o valor observado e $\hat{y}_i$ o valor predito.

O coeficiente de determinação mede a proporção da variabilidade na variável de resposta que é explicada pelo modelo. Ele varia de 0 a 1, onde 0 significa que o modelo não explica nenhuma variabilidade e 1 significa que o modelo explica toda a variabilidade. Quanto mais próximo de 1, melhor o modelo Chicco, Warrens e Jurman (2021). A Equação (3.7) indica a forma de calcular o $\mathbf{R^2}$.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2} \tag{3.7}$$

Onde y_i é o valor observado, $\hat{y}_i$ o valor predito e $\overline{y}$ a média dos valores observados.

O CCC é uma métrica utilizada para avaliar a concordância entre duas medidas. Ela mede a distância das predições até a linha de regressão e a que distância a linha de regressão se desvia de uma linha que passa até a origem. Seu resultado varia de -1 a $+1$, sendo quanto mais próximo de $+1$ o CCC, melhor o desempenho do modelo (GRAMATICA; SANGION, 2016). A Equação (3.8) esclarece como se determina o CCC.

$$\text{CCC} = \frac{2\sum_i^n (y_i - \overline{y})(\hat{y}_i - \overline{\hat{y}})}{\sum_i^n (y_i - \overline{y})^2 + \sum_i^n (\hat{y}_i - \overline{\hat{y}})^2 + n(\overline{y} - \overline{\hat{y}})} \tag{3.8}$$

Onde n é a quantidade de dados a serem preditos, y_i é o valor observado, $\hat{y}_i$ o valor predito, $\overline{y}$ a média dos valores observados e $\overline{\hat{y}}$ a média dos valores preditos.

4 MATERIAL E MÉTODOS

A abordagem metodológica utilizada neste trabalho pode ser separada em estágios, apresentados no fluxograma apresentado na Figura 9.

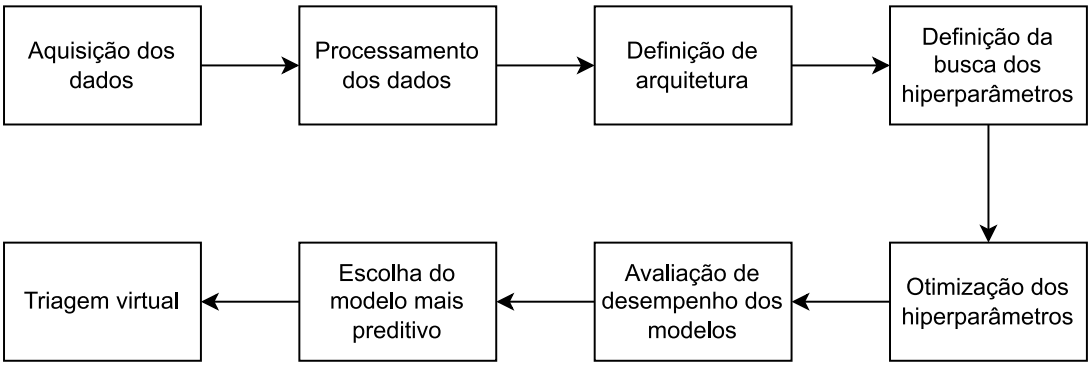

Figura 9 – Fluxograma das abordagens realizadas neste trabalho.

4.1 Aquisição dos dados

A aquisição de conjuntos de dados é um passo crítico no desenvolvimento de modelos de aprendizado de máquina com bom desempenho. A qualidade dos dados utilizados têm um impacto direto na capacidade do modelo de generalizar padrões e fazer previsões assertivas.

Os dados relacionados a propriedades físico-químicas utilizados foram obtidos de artigos científicos e os dados relacionados com a atividade biológica foram retirados do *ChEMBL* (MENDEZ et al., 2019), que é um banco de dados de substâncias químicas e atividades biológicas.

A Tabela 1 apresenta os conjuntos de dados utilizados, sua finalidade, quantidade de compostos disponíveis e sua fonte.

Tabela 1 – Descrição dos conjuntos de dados utilizados no trabalho, quantidade de compostos e fonte.

Conjunto de dados	**Finalidade**	**Tarefa**	**Quantidade de compostos**	**Fonte**
aqSolDB	Solubilidade em água	Regressão	9982	Sorkun, Khetan e Er (2019)
DMSO	Solubilidade em 10mM de DMSO	Classificação	50620	Tetko et al. (2013)
LogP	Lipofilicidade	Regressão	13688	Lukashina et al. (2020)
Acinetobacter baumannii	Atividade biológica	Classificação	11014	Mendez et al. (2019)
Escherichia coli	Atividade biológica	Classificação	70475	Mendez et al. (2019)
Pseudomonas aeruginosa	Atividade biológica	Classificação	43538	Mendez et al. (2019)
Salmonella Typhimurium	Atividade biológica	Classificação	4318	Mendez et al. (2019)
BraCoLi	Triagem virtual	-	1171	Veríssimo et al. (2022a)

4.2 Processamento de dados

Após a aquisição, geralmente é necessário um passo de limpeza e processamento para assegurar que os dados sejam precisos, confiáveis e representativos. Esse processo inclui lidar com valores ausentes, remover duplicatas, padronizar valores e transformar formatos para que os dados estejam prontos para serem utilizados. O fluxograma descrito na Figura 10 mostra os passos utilizados.

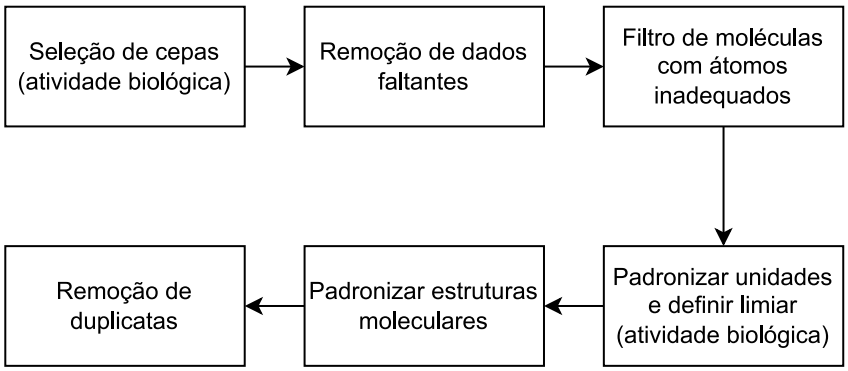

Figura 10 – Fluxograma dos passos de tratamento e limpeza de dados utilizados.

O primeiro passo foi escolher e selecionar as cepas das bactérias a serem utilizadas. Para a *Acinetobacter baumannii* foi escolhida a cepa ATCC19606, para a *Escherichia coli* a cepa ATCC25922, para a *Pseudomonas aeruginosa* a cepa ATCC27853 ou ATCC15442 e para a *Salmonella typhimurium* a cepa ATCC14028. Estas cepas foram escolhidas por serem consideradas cepas padrão e por não apresentarem modificações genéticas.

Após esse passo foi aplicado um filtro que remove as moléculas que possuem átomos diferentes de Carbono, Oxigênio, Nitrogênio, Enxofre, Fósforo, Flúor, Iodo, Bromo e Cloro. Essa remoção é realizada pois muitos programas que realizam os cálculos de descritores moleculares não conseguem processar moléculas inorgânicas (FOURCHES; MURATOV; TROPSHA, 2010).

Para os dados de atividade biológica foi necessário um passo adicional de padronização das unidades e seleção do limiar de classificação. Os dados reportados podem vir com unidades diferentes e por isso todas foram padronizadas para nano molar (nM).

Seguindo a metodologia de Shi, Zhao e Wei (2018) foi definido o limiar de 10000 nM para definir se o composto é ativo ou não. Compostos com uma atividade inferior a esse limiar foram consideradas ativas e compostos com uma atividade superior como inativas.

O próximo passo realizado foi a padronização das moléculas, pois uma mesma estrutura e grupos funcionais podem estar representados de forma diferente. Essa padronização foi realizada utilizando a biblioteca *MolVS* (SWAIN; MEYERS, 2018) que é uma biblioteca de padronização e validação de estruturas moleculares. A padronização

das estruturas contribui com o treinamento do modelo, pois reduz o ruído introduzido por diferentes representações moleculares, permitindo que o modelo foque nos padrões e características relevantes das estruturas moleculares.

E por fim, foi realizada a remoção de dados duplicados para evitar os efeitos de sobreajuste no modelo. Os dados relacionados com atividade biológica passam por um processo extra de comparação onde é verificado o rótulo dos dados para garantir a integridade das informações.

Caso exista duplicatas com rótulos diferentes para classificação, ambas são excluídas e caso possuam o mesmo rótulo é considerado o dado do pior caso, que é o de maior valor. Já para regressão é realizada uma média dos rótulos dos dados duplicados. A quantidade de dados obtidos após o processamento dos dados está descrita na Tabela 2.

Tabela 2 – Informações sobre a quantidade de compostos iniciais e quantidade de compostos após o processamento dos conjuntos de dados.

Conjunto de dados	**Quantidade inicial de compostos**	**Quantidade final de compostos**
Solubilidade em água	9982	8880
Solubilidade em DMSO	50620	50600
Lipofilicidade	13688	13610
Acinetobacter baumannii	11014	305
Escherichia coli	70475	1487
Pseudomonas aeruginosa	43538	35
Salmonella Typhimurium	4318	52
BraCoLi	1171	1161

Após esse processo, os dados precisam ser divididos em conjuntos de treinamento, validação e teste. A importância de uma boa distribuição entre os conjuntos é fundamental para garantir que o modelo seja capaz de generalizar bem para dados que não foram vistos anteriormente. Isso afeta diretamente a capacidade do modelo de realizar previsões confiáveis em situações do mundo real.

Para garantir a distribuição adequada das moléculas entre os conjuntos foi utilizado o algoritmo *MASSA* (VERíSSIMO et al., 2022b), que separa os compostos em agrupamentos de características similares. Isso permite um bom balanceamento de estruturas moleculares, propriedades moleculares e atividades biológicas, evitando que o modelo não aprenda um determinado grupo funcional devido a separações aleatórias. Os dados de cada conjunto foram divididos em 80% treino (utilizado para treinar o modelo), 10 teste (utilizado para verificar o andamento do treino) e 10% validação (utilizado como simulação de dados reais externos).

4.3 Definição dos experimentos

O desenvolvimento do presente trabalho foi realizado utilizando *Python* e a implementação das GNN foi realizada com o *PyTorch Geometric* (FEY; LENSSEN, 2019), que é uma biblioteca baseada no *PyTorch* (PASZKE et al., 2019) com vários métodos de aprendizado de máquina para grafos. Outra biblioteca de suma importância utilizada foi o *Ray Tune* (LIAW et al., 2018), que além de otimizar os hiperparâmetros possibilita a execução paralelizável de experimentos. A máquina utilizada para o treinamento e teste dos modelos possui um processador *AMD Ryzen* 7 3700X, 64 GB de memória RAM e uma placa de vídeo *NVIDIA* RTX A4000.

O processo de escolha de arquitetura do modelo, treinamento e otimização dos hiperparâmetros seguiram várias sugestões do trabalho de Godbole et al. (2023), que é um documento que sintetiza boas práticas para aplicações de aprendizado profundo.

A arquitetura escolhida foi a *Attentive FP*, que é uma GNN especializada para a representação molecular. Ela usa os princípios de mecanismos de atenção separadamente para os átomos e para as ligações, permitindo o aprendizado de propriedades locais e não-locais da estrutura química. A arquitetura *Attentive FP* foi escolhida pois demonstrou um bom resultado preditivo para representações moleculares e pelo fato de ser uma arquitetura já bem estabelecida.

O processo de treinamento e otimização dos hiperparâmetros foram realizados em duas etapas. A primeira etapa consistiu do treinamento de 500 modelos únicos para cada conjunto de dados, otimizando os hiperparâmetros da arquitetura em si. Utilizando o modelo que obteve o maior valor de MCC ou RMSE na primeira etapa, foram treinados mais 500 modelos otimizando agora os hiperparâmetros do otimizador.

A quantidade de *embeddings*, que pode ser definida como a dimensão de um espaço de representação, variou de 32 a 256; o número de camadas da rede neural variou entre 2 e 4; a taxa de *dropout*, que é uma técnica de regularização utilizada para prevenir o sobreajuste, variou de 0,2 a 0,4. Além disso, foram testados os otimizadores da rede neural SGD e Adam; taxa de aprendizado, que define o tamanho dos passos que o otimizador dá ao ajustar os pesos durante o treinamento, entre 0,00001 a 0,01.

4.4 Pré-treinamento

Nos conjuntos de dados relacionados com a atividade biológica bacteriana foi realizado um pré-treinamento treinando um modelo com dados das bactérias Gram-negativas alvo deste trabalho que não possuíam especificação de cepa.

Esses dados não foram utilizados efetivamente no treinamento direto dos modelos

por não fornecer a cepa bacteriana utilizada e consequentemente se há alguma modificação no gene da bactéria. Isso pode inserir dados ruidosos para o modelo e prejudicar o aprendizado.

Após a seleção, processamento e limpeza de todos os dados relacionados a bactérias que não possuem uma cepa especificada, foi obtido um conjunto de dados contendo 28908 compostos. O modelo treinado com estes dados foi utilizado como ponto de partida para o treinamento dos modelos específicos para cada bactéria alvo.

4.5 Avaliação dos experimentos e triagem virtual

Após o treinamento, foi selecionado o modelo de cada conjunto de dados que obteve o maior valor de MCC (para classificação) ou CCC (para regressão), e foi obtido a média e o desvio padrão das métricas avaliadas para vinte execuções.

Os melhores modelos foram utilizados para realizar uma triagem virtual em uma biblioteca de compostos. A biblioteca escolhida foi a *Brazilian Compound Library* (BraCoLi) (VERíSSIMO et al., 2022a), que é um conjunto de dados desenvolvido por grupos de pesquisa do Brasil. Os dados da triagem virtual passaram pelo processo de limpeza para assegurar a padronização das estruturas moleculares.

Foi realizada uma análise do domínio de aplicabilidade dos modelos treinados para avaliar e assegurar que as predições são feitas com base nas informações vistas no conjunto de treino. Foi utilizada a metodologia descrita por Fernandes et al. (2021) que realiza uma análise de componentes principais em um descritor vetorial da molécula e em seguida analisa a distância entre os pontos de treino, teste e validação. Foi utilizado a distância do Cosseno, distância Euclidiana, distância de Manhattan, e a distância de Wasserstein para calcular a distância entre os pontos (amostras do conjunto de teste, de validação e da quimioteca utilizada na triagem virtual em comparação com as amostras do conjunto de treinamento) e o composto será considerado fora do domínio de aplicabilidade somente se o consenso das quatro métricas indicarem.

5 RESULTADOS E DISCUSSÃO

Como mencionado, o processo de otimização de hiperparâmetros ocorreu em duas etapas, a primeira sendo responsável pela otimização dos hiperparâmetros particulares da arquitetura *Attentive FP* e a segunda etapa pelos hiperparâmetros do otimizador. Foi escolhido o modelo com o maior valor de MCC, para classificação, e CCC, para regressão.

5.1 Modelos treinados

Primeiro foram treinados os modelos da tarefa de regressão, começando com solubilidade em água. Os hiperparâmetros otimizados de arquitetura possuem as seguintes características: duas camadas, *timestep* único, um *dropout* de 0,208 e 236 *embeddings*. Já para os hiperparâmetros do otimizador temos as seguintes características: otimizador Adam com os valores de beta em 0,524 e 0,994, taxa de aprendizado de 0,006.

Já para lipofilicidade os hiperparâmetros encontrados possuem as seguintes características: duas camadas, *timestep* único, um *dropout* de 0,202 e 248 *embeddings*. Já para os hiperparâmetros do otimizador temos as seguintes características: otimizador Adam com os valores de beta em 0,627 e 0,744, taxa de aprendizado de 0,0033.

A Tabela 3 mostra o desempenho das métricas avaliadas de regressão, a Figura 11a mostra o resultado da regressão da solubilidade em água e a Figura 11b mostra o resultado da regressão de lipofilicidade.

Os modelos de classificação foram treinados a seguir, começando pelo conjunto de dados de solubilidade de DMSO. Os hiperparâmetros possuem as seguintes características: duas camadas, *timestep* único, um *dropout* de 0,218 e 244 *embeddings*. Já para os hiperparâmetros do otimizador temos as seguintes características: otimizador Adam com os valores de beta em 0,841 e 0,931 taxa de aprendizado de 0,0071.

Para os modelos de atividade biológica, primeiro foi realizado o pré-treinamento de um modelo geral com informações das bactérias Gram-negativas alvo deste trabalho para fornecer o ponto inicial para o treinamento específico para os modelos de cada bactéria.

Os hiperparâmetros do modelo geral possuem as seguintes características: duas camadas, *timestep* único, um *dropout* de 0,226 e 248 *embeddings*. Já para os hiperparâmetros do otimizador temos as seguintes características: otimizador Adam com os valores de beta em 0,688 e 0,622, taxa de aprendizado de 0,0013.

A Tabela 3 mostra o desempenho das métricas avaliadas de classificação, a Figura 11c mostra a matriz de confusão do modelo de solubilidade em DMSO e a Figura 11d

mostra a matriz de confusão do modelo geral de atividade biológica.

Tabela 3 – Métricas dos modelos preditivos de solubilidade em água, lipofilicidade, solubilidade em DMSO e atividade biológica contra bactérias Gram-negativas.

	Regressão		**Classificação**	
	Solubilidade em água	**Lipofilicidade**	**Solubilidade em DMSO**	**Atividade biológica**
CCC *Teste*	0.907 ± 0.002	0.957 ± 0.001	**MCC** *Teste*	0.174 ± 0.004	0.312 ± 0.009
CCC *Validação*	0.923 ± 0.000	0.955 ± 0.000	**MCC** *Validação*	0.204 ± 0.000	0.321 ± 0.000
R² *Teste*	0.829 ± 0.003	0.918 ± 0.002	**Precisão** *Teste*	0.976 ± 0.001	0.940 ± 0.005
R² *Validação*	0.857 ± 0.000	0.913 ± 0.000	**Precisão** *Validação*	0.976 ± 0.000	0.939 ± 0.000
RMSE *Teste*	0.966 ± 0.008	0.533 ± 0.007	**Revocação** *Teste*	0.695 ± 0.010	0.581 ± 0.011
RMSE *Validação*	0.877 ± 0.000	0.555 ± 0.000	**Revocação** *Validação*	0.765 ± 0.000	0.602 ± 0.000
			F1 *Teste*	0.812 ± 0.007	0.718 ± 0.007
			F1 *Validação*	0.858 ± 0.000	0.733 ± 0.000

5.2 Modelos com transferência de aprendizado

O modelo geral de atividade biológica foi utilizado como ponto de partida para o treinamento dos modelos específicos para as bactérias alvo deste trabalho. A Figura 12a mostra a matriz de confusão para a *Acinetobacter baumannii*, a Figura 12b mostra a matriz de confusão para a *Escherichia coli*, a Figura 12c mostra a matriz de confusão para a *Pseudomonas aeruginosa* e a Figura 12d mostra a matriz de confusão para a *Salmonella typhimurium*.

Os modelos possuem os mesmos hiperparâmetros que o modelo geral e a Tabela 4 mostra o desempenho das métricas avaliadas de classificação.

5.3 Triagem virtual

Os modelos escolhidos foram utilizados para realizar predições nos dados da biblioteca BraCoLi. Foram preditas as propriedades de solubilidade em água, solubilidade em DMSO e lipofilicidade e também a atividade antibacteriana contra *Acinetobacter baumannii*, *Escherichia coli*, *Pseudomonas aeruginosa* e *Salmonella typhimurium*.

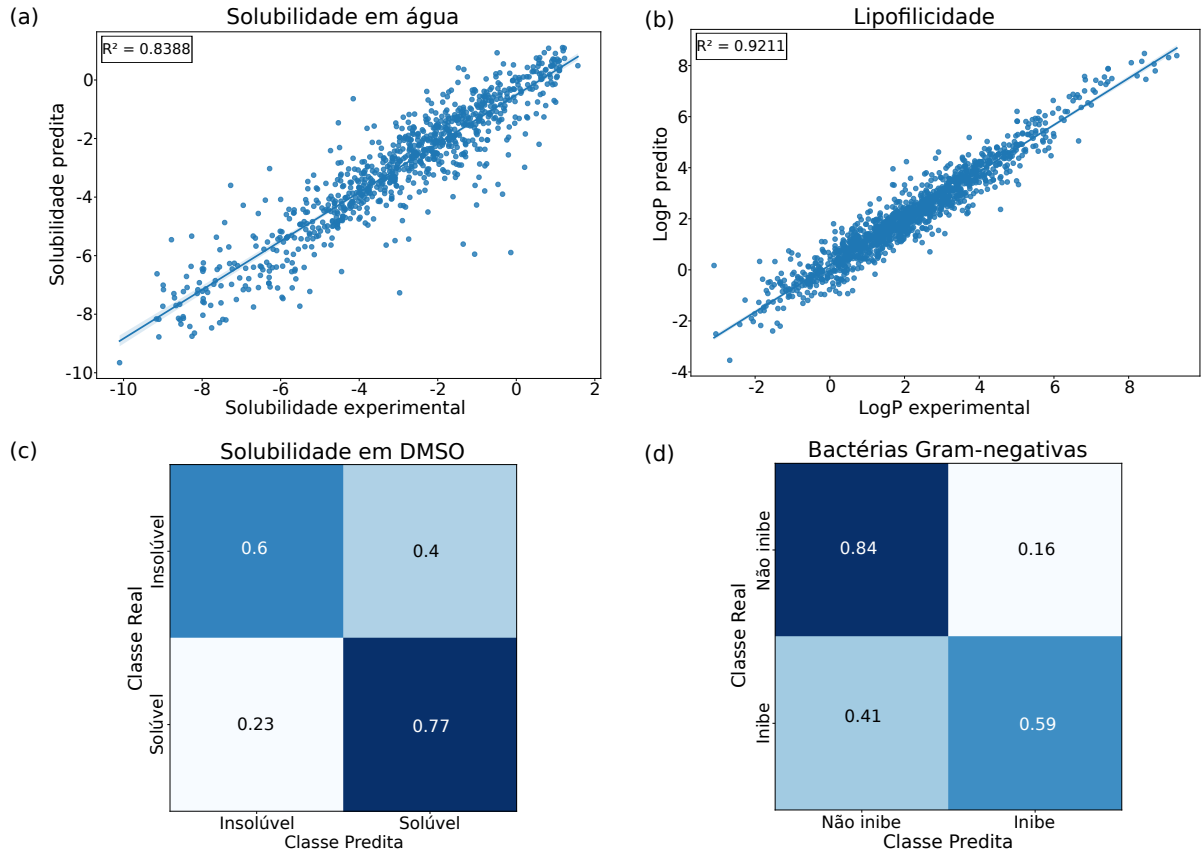

Figura 11 – Resultado do treinamento dos modelos. Em (a) está descrito a reta de regressão para o modelo preditivo de solubilidade em água, Em (b) a reta de regressão para o modelo preditivo de lipofilicidade, em (c) está exposta a matriz de confusão do modelo preditivo de solubilidade em DMSO e em (d) está exposta a matriz de confusão do modelo preditivo de atividade biológica contra bactérias Gram-negativas.

Foram identificados 792 compostos com atividade antibacteriana contra *Acinetobacter baumannii*, 803 compostos com atividade antibacteriana contra *Escherichia coli*, 971 compostos com atividade antibacteriana contra *Pseudomonas aeruginosa*, 464 compostos com atividade antibacteriana contra *Salmonella typhimurium*. Também foram identificados 302 compostos que possuem atividade antibacteriana contra todas as bactérias testadas, indicando um possível antibacteriano de amplo espectro.

Os compostos que apresentaram atividade antibacteriana contra todas as bactérias testadas foram classificados em relação às propriedades preditas. A Figura 13 mostra a distribuição dos valores preditos de solubilidade em água e lipofilicidade.

Para exemplificar, foram separados 6 compostos com características desejáveis, sendo eles: BR010508, BR010105, BR010681, BR020355, BR010167 e BR020108. A Figura 14 mostra a estrutura molecular dos compostos e a Tabela 5 mostra o resultado das predições.

Tabela 4 – Métricas dos modelos preditivos de atividade biológica contra *Acinetobacter baumannii*, *Escherichia coli*, *Pseudomonas aeruginosa* e *Salmonella typhimurium*.

	Acinetobacter baumannii	***Escherichia coli***	***Pseudomonas aeruginosa***	***Salmonella typhimurium***
MCC *Teste*	0.622 ± 0.021	0.869 ± 0.016	1 ± 0.000	1 ± 0.000
MCC *Validação*	0.451 ± 0.000	0.798 ± 0.000	1 ± 0.000	1 ± 0.000
Precisão *Teste*	0.950 ± 0.001	1 ± 0.000	1 ± 0.000	1 ± 0.000
Precisão *Validação*	0.920 ± 0.000	0.964 ± 0.000	1 ± 0.000	1 ± 0.000
Revocação *Teste*	0.833 ± 0.016	0.932 ± 0.009	1 ± 0.000	1 ± 0.000
Revocação *Validação*	0.885 ± 0.000	0.939 ± 0.000	1 ± 0.000	1 ± 0.000
F1 *Teste*	0.888 ± 0.009	0.965 ± 0.005	1 ± 0.000	1 ± 0.000
F1 *Validação*	0.902 ± 0.000	0.952 ± 0.000	1 ± 0.000	1 ± 0.000

5.4 Domínio de aplicabilidade

O domínio de aplicabilidade dos modelos foi verificado para as separações de treino, teste, validação e triagem virtual. A Tabela 6 mostra a quantidade de compostos avaliados e a quantidade que foi calculada como fora do domínio de aplicabilidade.

Com os resultados obtidos da análise foi possível plotar um gráfico bidimensional com as informações da análise de componentes principais e um gráfico tridimensional com as informações de cada amostra. A Figura 15 mostra o gráfico bidimensional e a Figura 16 mostra o gráfico tridimensional dos modelos treinados. A Figura 17 mostra o gráfico bidimensional e a Figura 18 mostra o gráfico tridimensional dos modelos com transferência

Tabela 5 – Valores preditos para as 6 substancias selecionadas

Composto	**Solubilidade em água**	**LogP**	**Solubilidade em DMSO**	***Acinetobacter baumannii***	***Escherichia coli***	***Pseudomonas aeruginosa***	***Salmonella typhimurium***
BR010508	-4.630	2.542	1	1	1	1	1
BR010105	-3.361	2.565	1	1	1	1	1
BR010681	-6.12	2.225	1	1	1	1	1
BR020355	-3.926	2.756	1	1	1	1	1
BR010167	-4.316	2.501	1	1	1	1	1
BR020108	-4.524	2.646	1	1	1	1	1

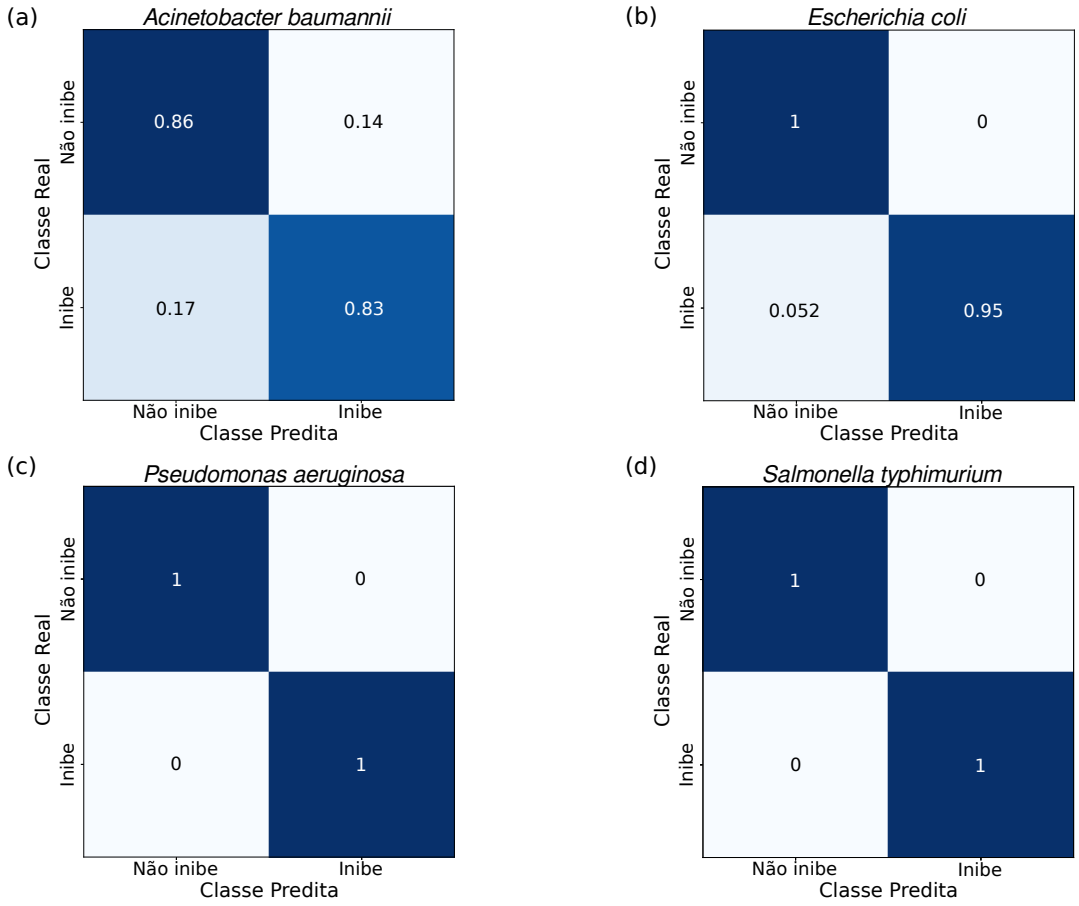

Figura 12 – Resultado do treinamento dos modelos com transferência de aprendizado. Em (a) é apresentada a matriz de confusão para o modelo preditivo de atividade biológica contra *Acinetobacter baumannii*, em (b) é apresentada a matriz de confusão para o modelo preditivo de atividade biológica contra *Escherichia coli*, em (c) é mostrada a matriz de confusão para o modelo preditivo de atividade biológica contra *Pseudomonas aeruginosa* e em (d) é mostrada a matriz de confusão para o modelo preditivo de atividade biológica contra *Salmonella typhimurium*.

de aprendizado.

5.5 Discussão

As métricas apresentadas pelos modelos específicos de bactérias Gram-negativas demonstram que ocorreu uma grande melhoria devido ao processo de transferência de aprendizado. O modelo preditivo contra *Acinetobacter baumannii* alcançou um MCC de 0,451 e o contra *Escherichia coli* um MCC de 0,798.

É importante ressaltar que mesmo os modelos contra *Pseudomonas aeruginosa* e *Salmonella Typhimurium* terem um MCC de 1, estes conjuntos de dados sofrem de uma extrema escassez de dados, diminuindo a confiabilidade do modelo.

Essa incerteza sobre o modelo fica nítida ao observar a Tabela 6, pois com

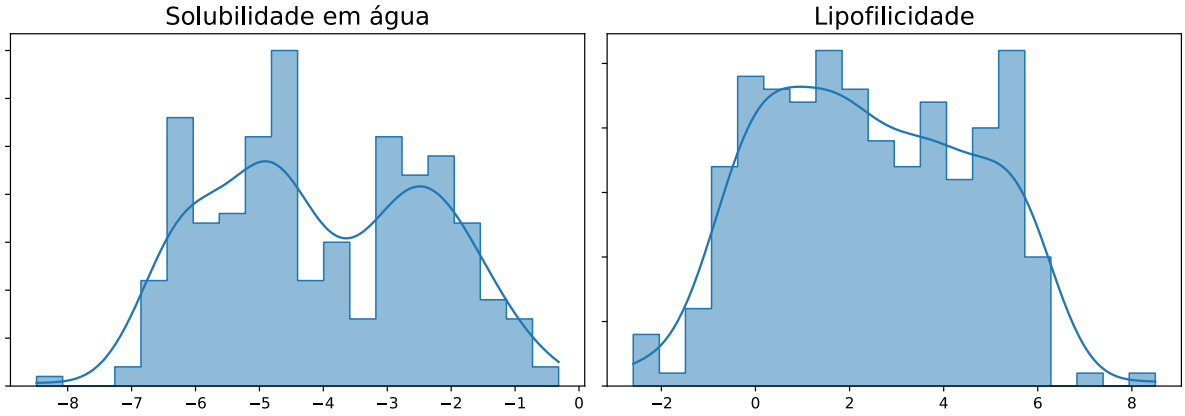

Figura 13 – Distribuição dos valores preditos de solubilidade em água e lipofilicidade para os compostos que tem atividade biológica contra as bactérias utilizadas neste trabalho

menos dados o espaço do domínio de aplicabilidade é reduzido. O modelo preditivo contra *Salmonella Typhimurium* teve 151 compostos calculados como fora do domínio de aplicabilidade, que é cerca de 13% dos dados externos da triagem virtual.

Os modelos de regressão apresentaram um excelente desempenho, atingindo um R^2 de 0,857 para solubilidade em água e 0,913 para lipofilicidade. O trabalho de Gramatica e Sangion (2016) discute que um valor aceitável de R^2 para se realizar predições com modelos computacionais (de regressão) é 0,65.

Um ponto a ser considerado sobre o domínio de aplicabilidade é que enquanto os

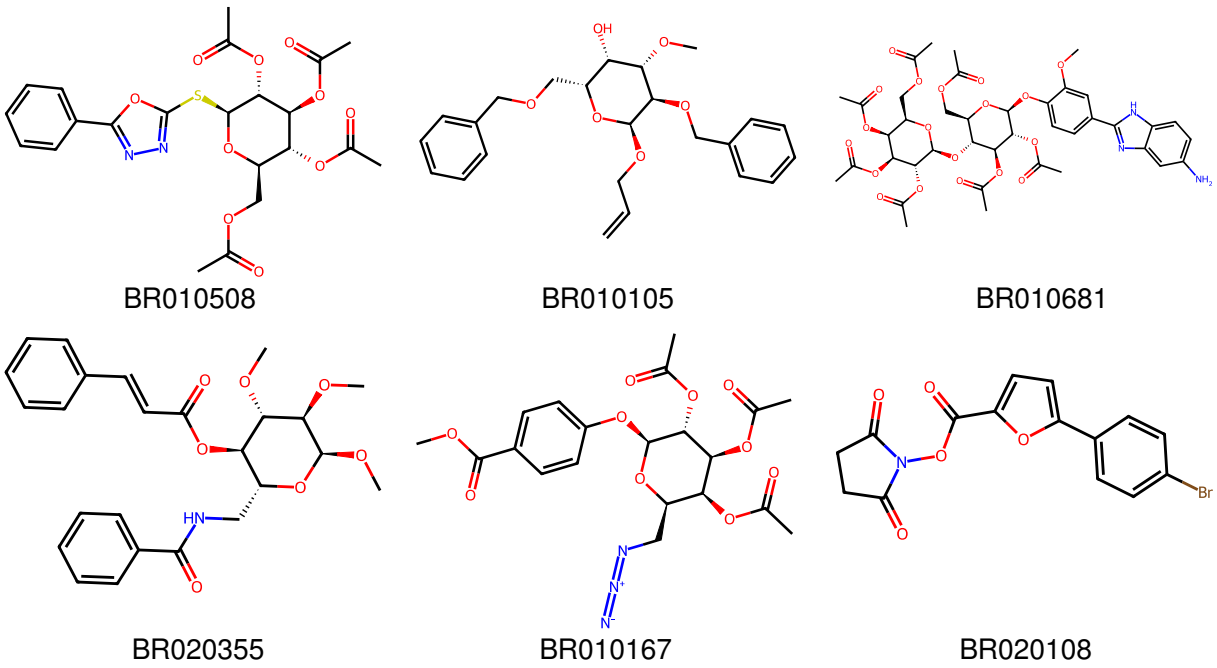

Figura 14 – Estrutura molecular de 6 compostos com características desejáveis para realização de testes experimentais

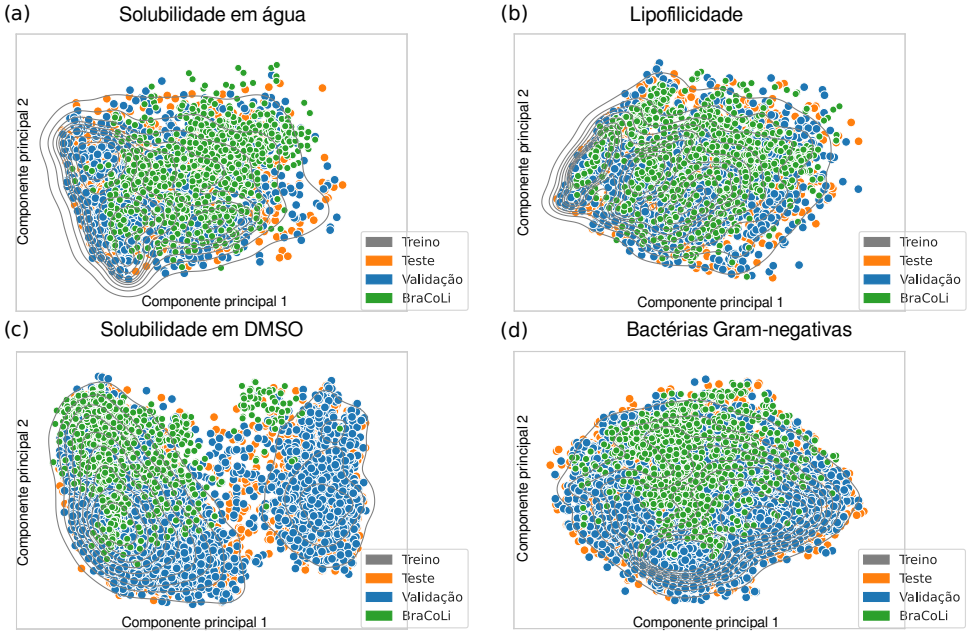

Figura 15 – Gráficos bidimensionais do domínio de aplicabilidade. Em (a): solubilidade em água, em (b): lipofilicidade, em (d) solubilidade em DMSO e em (d) atividade contra bactérias Gram-negativas.

modelos foram treinados com as moléculas representadas por grafos, o cálculo de domínio de aplicabilidade foi realizado utilizando uma representação vetorial da molécula. É de suma importância que os dados e a estimativa do domínio sejam realizados utilizando os mesmos princípios para verdadeiramente assegurar que os compostos são similares.

A triagem virtual realizada demonstrou que existem em média 757 compostos com uma potencial atividade antibacteriana contra as bactérias utilizadas neste trabalho. Por se tratar de uma biblioteca de compostos pesquisados por brasileiros, o acesso a estas substâncias é mais acessível e é possível realizar testes experimentais mais rapidamente.

Tabela 6 – Número total de compostos avaliados no domínio de aplicabilidade em cada separação de dados e quantidade de compostos que ficaram fora dos limites do domínio.

Conjunto de dados	**Total *Treino***	**Total *Teste***	**Total *Validação***	**Total *Triagem***	**Fora *Teste***	**Fora *Validação***	**Fora *Triagem***
Solubilidade em água	7103	888	889	1161	0	2	0
Solubilidade em DMSO	40480	5060	5060	1161	3	5	2
Lipofilicidade	10887	1361	1362	1161	7	11	1
Acinetobacter baumannii	244	30	31	1161	0	0	6
Escherichia coli	1190	148	149	1161	1	1	2
Pseudomonas aeruginosa	28	3	4	1161	0	0	4
Salmonella typhimurium	41	5	6	1161	0	0	151

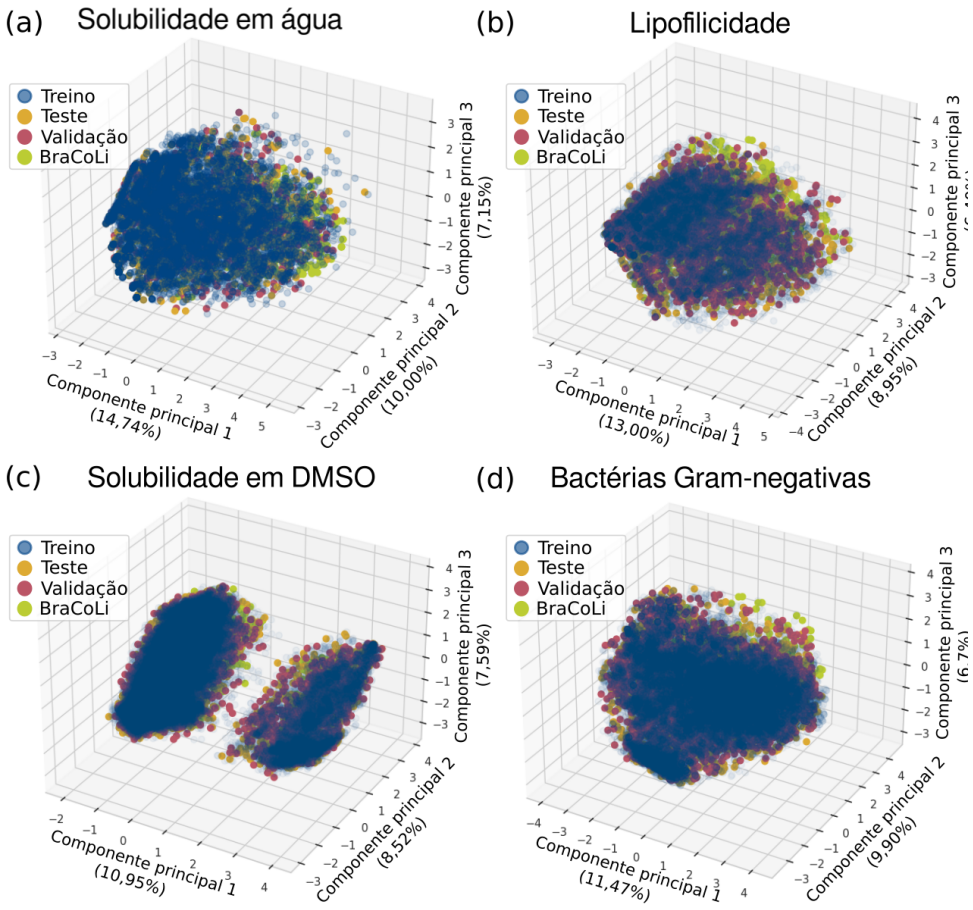

Figura 16 – Gráficos tridimensionais do domínio de aplicabilidade. Em (a): solubilidade em água, em (b): lipofilicidade, em (d) solubilidade em DMSO e em (d) atividade contra bactérias Gram-negativas. Os valores entre parenteses são a variância da respectivas componentes

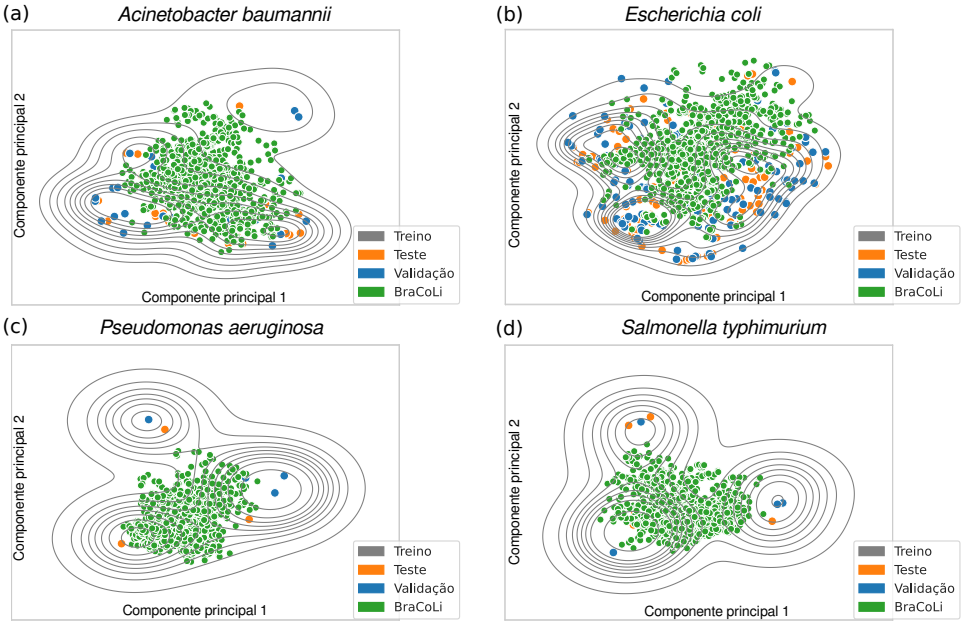

Figura 17 – Gráficos bidimensionais do domínio de aplicabilidade. Em (a): solubilidade em água, em (b): lipofilicidade, em (d) solubilidade em DMSO e em (d) atividade contra bactérias Gram-negativas.

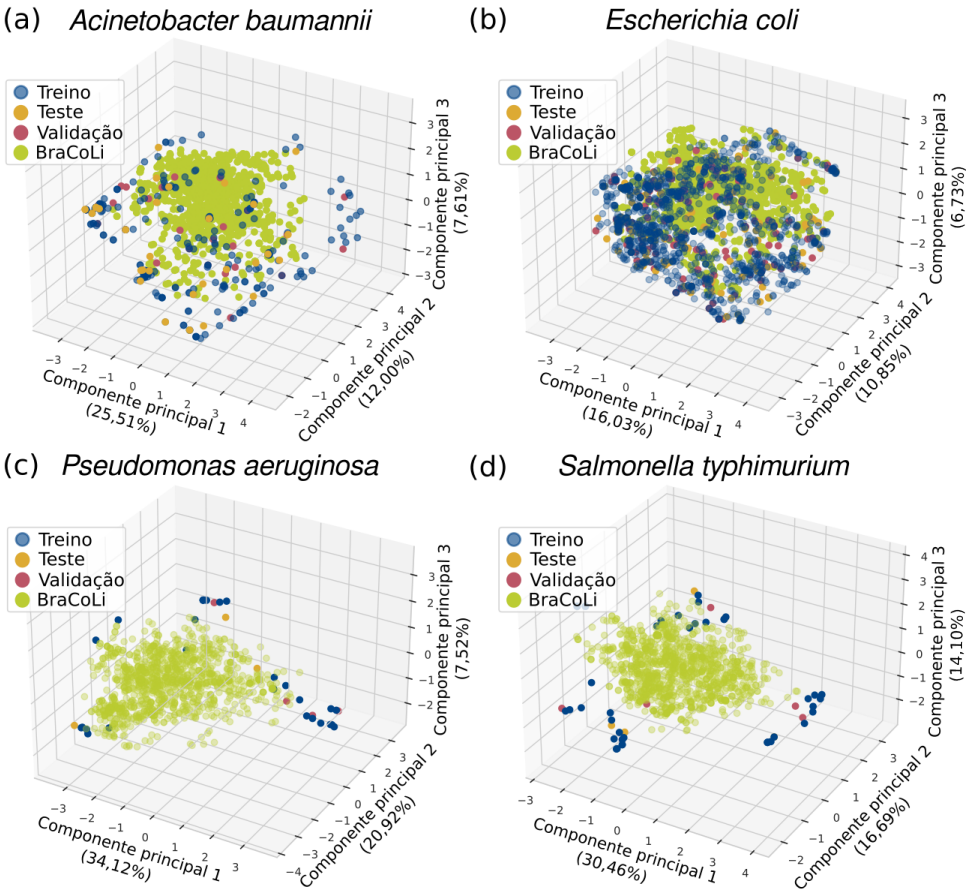

Figura 18 – Gráficos tridimensionais do domínio de aplicabilidade. Em (a): solubilidade em água, em (b): lipofilicidade, em (d) solubilidade em DMSO e em (d) atividade contra bactérias Gram-negativas. Os valores entre parenteses são a variância da respectivas componentes

6 CONCLUSÃO

Neste trabalho foi apresentada uma aplicação de aprendizado profundo com GNN para a área de planejamento de fármacos com o intuito de realizar predições acerca das propriedades físico-químicas (solubilidade em água, lipofilicidade e solubilidade em DMSO) e de atividade biológica contra 4 bactérias Gram-negativas (*Acinetobacter baumannii*, *Escherichia coli*, *Pseudomonas aeruginosa* e *Salmonella Typhimurium*).

Os modelos preditivos para as tarefas de classificação obtiveram um MCC acima de 0,20 em todos os casos, e os modelos de regressão obtiveram um R^2 superior a 0,85.

Foi realizada a análise do domínio de aplicabilidade dos modelos, parte crucial em um trabalho computacional de planejamento de fármacos e também uma triagem virtual em uma biblioteca de compostos onde foi identificado possíveis antibacterianos.

As GNNs podem ser aplicadas para outros âmbitos dentro de planejamento de fármacos, não só atividade bacteriana, sendo assim este trabalho apresenta uma contribuição para acelerar o desenvolvimentos de fármacos diversos.

6.1 Trabalhos futuros

Durante a execução do trabalho foram identificados alguns assuntos que precisam de continuidade e de uma atenção especial. Como continuação do trabalho, pretende-se realizar testes experimentais com os compostos identificados na triagem virtual com potencial antibacteriano. Assim, é possível verificar na prática o poder preditivo do modelo.

Um ponto identificado durante a execução do trabalho foi a falta de uma ferramenta já difundida para calcular o domínio de aplicabilidade para moléculas que estão representadas como grafos. Assim, um estudo será realizado para disponibilizar essa ferramenta.

REFERÊNCIAS

ABDEL-JABER, H. et al. *A Review of Deep Learning Algorithms and Their Applications in Healthcare*. 2022.

ALFEI, S.; SCHITO, A. M. *Positively charged polymers as promising devices against multidrug resistant gram-negative bacteria: A Review*. 2020.

ASIF, N. A. et al. *Graph Neural Network: A Comprehensive Review on Non-Euclidean Space*. 2021.

BRASIL. *PORTARIA GM/MS Nº 3.398, DE 7 DE DEZEMBRO DE 2021*. 2021.

CDC. Antibiotic resistance threats in the united states, 2019, atlanta, ga: U.s. department of health and human services. *Center for Disease control and Prevention*, 2019.

CHEN, H. et al. *The rise of deep learning in drug discovery*. 2018.

CHICCO, D.; TöTSCH, N.; JURMAN, G. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, v. 14, 2021. ISSN 17560381.

CHICCO, D.; WARRENS, M. J.; JURMAN, G. The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *PeerJ Computer Science*, v. 7, 2021. ISSN 23765992.

DAS, P. et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature Biomedical Engineering*, v. 5, 2021. ISSN 2157846X.

DOYTCHINOVA, I. *Drug Design—Past, Present, Future*. 2022.

ECE, A. *Computer-aided drug design*. 2023.

FERNANDES, P. O. et al. Molecular insights on abl kinase activation using tree-based machine learning models and molecular docking. *Molecular Diversity*, v. 25, 2021. ISSN 1573501X.

FEY, M.; LENSSEN, J. E. Fast graph representation learning with pytorch geometric. *arXiv*, 2019.

FOURCHES, D.; MURATOV, E.; TROPSHA, A. *Trust, but verify: On the importance of chemical structure curation in cheminformatics and QSAR modeling research*. 2010.

FRADKOV, A. L. Early history of machine learning. In: . [S.l.: s.n.], 2020. v. 53. ISSN 24058963.

GADALETA, D. et al. Applicability domain for qsar models. *International Journal of Quantitative Structure-Property Relationships*, v. 1, 2016. ISSN 2379-7487.

GILMER, J. et al. Neural message passing for quantum chemistry. In: . [S.l.: s.n.], 2017. v. 3.

GODBOLE, V. et al. *Deep Learning Tuning Playbook*. 2023.

GRAMATICA, P.; SANGION, A. *A Historical Excursus on the Statistical Validation Parameters for QSAR Models: A Clarification Concerning Metrics and Terminology*. 2016.

GUO, M. et al. *Pharmaceutical cocrystals: A review of preparations, physicochemical properties and applications*. 2021.

HIGHAM, C. F.; HIGHAM, D. J. Deep learning: An introduction for applied mathematicians. *SIAM Review*, v. 61, 2019. ISSN 00361445.

HOANG, V. T. et al. *Graph Representation Learning and Its Applications: A Survey*. 2023.

HODSON, T. O. *Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not*. 2022.

JADON, S. A survey of loss functions for semantic segmentation. In: . [S.l.: s.n.], 2020.

JUNIOR, L. R. A. et al. Virtual screening of antibacterial compounds by similarity search of enoyl-acp reductase (fabi) inhibitors. *Future Medicinal Chemistry*, v. 12, 2019. ISSN 17568927.

KADURIN, A. et al. Drugan: An advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular Pharmaceutics*, v. 14, 2017. ISSN 15438392.

LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. *Nature*, v. 521, p. 436–444, 2015. ISSN 0028-0836.

LEESON, P. D.; YOUNG, R. J. Molecular property design: Does everyone get it? *ACS Medicinal Chemistry Letters*, v. 6, 2015. ISSN 19485875.

LI, J. et al. Drug discovery approaches using quantum machine learning. In: . [S.l.: s.n.], 2021. v. 2021-December. ISSN 0738100X.

LIAW, R. et al. Tune: A research platform for distributed model selection and training. *arXiv*, 2018.

LO, Y. C. et al. *Machine learning in chemoinformatics and drug discovery*. 2018.

LUKASHINA, N. et al. Lipophilicity prediction with multitask learning and molecular substructures representation. *arXiv*, 2020.

MACALINO, S. J. Y. et al. *Role of computer-aided drug design in modern drug discovery*. 2015.

MAIER, A. et al. *A gentle introduction to deep learning in medical image processing*. 2019.

MENDEZ, D. et al. Chembl: Towards direct deposition of bioassay data. *Nucleic Acids Research*, v. 47, 2019. ISSN 13624962.

NADEEM, S. F. et al. *Antimicrobial resistance: more than 70 years of war between humans and bacteria*. 2020.

PANTALEãO, S. Q. et al. *Recent Advances in the Prediction of Pharmacokinetics Properties in Drug Design Studies: A Review*. 2022.

PASZKE, A. et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv*, 2019.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, v. 65, 1958. ISSN 0033295X.

ROY, K.; KAR, S.; AMBURE, P. On a simple approach for determining applicability domain of qsar models. *Chemometrics and Intelligent Laboratory Systems*, v. 145, 2015. ISSN 18733239.

SAHIGARA, F. et al. Comparison of different approaches to define the applicability domain of qsar models. *Molecules*, v. 17, 2012. ISSN 14203049.

SCHNEIDER, P. et al. *Rethinking drug design in the artificial intelligence era*. 2020.

SERAFIM, M. S. M. et al. *The application of machine learning techniques to innovative antibacterial discovery and development*. 2020.

SHI, J.; ZHAO, G.; WEI, Y. Computational qsar model combined molecular descriptors and fingerprints to predict hdac1 inhibitors. *Medecine/Sciences*, v. 34, 2018. ISSN 19585381.

SIMõES, R. S. et al. *Transfer and multi-task learning in QSAR modeling: Advances and challenges*. 2018.

SORKUN, M. C.; KHETAN, A.; ER, S. Aqsoldb, a curated reference set of aqueous solubility and 2d descriptors for a diverse set of compounds. *Scientific Data*, v. 6, 2019. ISSN 20524463.

STOKES, J. M. et al. A deep learning approach to antibiotic discovery. *Cell*, v. 180, 2020. ISSN 10974172.

SWAIN, M.; MEYERS, J. *MolVS*. 2018.

TETKO, I. V. et al. Development of dimethyl sulfoxide solubility models using 163 000 molecules: Using a domain applicability metric to select more reliable predictions. *Journal of Chemical Information and Modeling*, v. 53, 2013. ISSN 1549960X.

UDDIN, S. et al. Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, v. 19, 2019. ISSN 14726947.

VALERO-CARRERAS, D.; ALCARAZ, J.; LANDETE, M. Comparing two svm models through different metrics based on the confusion matrix. *Computers and Operations Research*, v. 152, 2023. ISSN 03050548.

VAMATHEVAN, J. et al. *Applications of machine learning in drug discovery and development*. 2019.

VEMULA, D. et al. *CADD, AI and ML in drug discovery: A comprehensive review*. [S.l.]: Elsevier B.V., 2023.

VERíSSIMO, G. C. et al. The brazilian compound library (bracoli) database: a repository of chemical and biological information for drug design. *Molecular Diversity*, v. 26, 2022. ISSN 1573501X.

VERíSSIMO, G. C. et al. Massa algorithm: automated rational sampling of training and test subsets for qsar modelling. *ChemRxiv*, Cambridge Open Engage, 2022.

WALDEN, D. M. et al. *Molecular Simulation and Statistical Learning Methods toward Predicting Drug–Polymer Amorphous Solid Dispersion Miscibility, Stability, and Formulation Design*. 2021.

WANG, L. et al. Review of classification methods on unbalanced data sets. *IEEE Access*, v. 9, 2021. ISSN 21693536.

YANG, X. et al. *Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery*. 2019.

ZHANG, Z.; CUI, P.; ZHU, W. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, v. 34, 2022. ISSN 15582191.

ZHOU, J. et al. *Graph neural networks: A review of methods and applications*. 2020.