

Avaliação de Modelos de Redes Neurais Artificiais na Classificação de Similaridade Estrutural Molecular

Isabeli Rosana Reik¹, Giancarlo Dondoni Salton¹

¹Universidade Federal da Fronteira Sul
Campus Chapecó

isabeli.reik@estudante.uffrs.edu.br, gian@uffrs.edu.br

Resumo. A similaridade molecular é um conceito central na quimioinformática, com aplicações relevantes na avaliação de novos medicamentos, em especial os medicamentos órfãos. Os métodos tradicionais para essa tarefa baseiam-se em fingerprints combinados a modelos de regressão logística, que, embora apresentem bom desempenho, possuem limitações à perda de informação estrutural e à linearidade dos classificadores. Este trabalho investiga o desempenho de modelos de deep learning aplicados diretamente à classificação de similaridade molecular a partir de representações 2D e 3D, por meio de três arquiteturas convolucionais inspiradas na LeNet-5, adaptadas para o processamento simultâneo de pares de imagens 2D e representações 3D. Os resultados indicam que os modelos propostos não superaram o baseline de regressão logística da literatura, atingindo ROCAUC entre 0,5627 e 0,6356 frente a valores entre 0,845 e 0,988 dos modelos de referência. O volume reduzido de dados de treinamento (200 pares moleculares) é apontado como principal fator limitante. O estudo contribui para a área ao reportar de forma transparente os desafios da abordagem e indicar caminhos para pesquisas futuras, como a ampliação dos conjuntos de dados, o balanceamento de classes e a exploração de transferência de aprendizado.

Abstract. Molecular similarity is a central concept in cheminformatics, with relevant applications in the evaluation of new drugs, especially orphan drugs. Traditional methods for this task rely on fingerprints combined with logistic regression models, which, despite their good performance, have inherent limitations related to structural information loss and classifier linearity. This work investigates the performance of deep learning models applied directly to molecular similarity classification from 2D and 3D representations, through three convolutional architectures inspired by LeNet-5 and adapted for the simultaneous processing of pairs of 2D images and 3D representations. The results indicate that the proposed models did not outperform the logistic regression baseline reported in the literature, achieving ROCAUC values between 0.5627 and 0.6356, compared to values between 0.845 and 0.988 for the reference models. The reduced training data volume (200 molecular pairs) is identified as the main limiting factor. The study contributes to the field by transparently reporting the challenges of the approach and identifying directions for future research, such as dataset expansion, class balancing, and the exploration of transfer learning architectures.

1. Introdução

A quimioinformática tem avançado significativamente nas últimas décadas, consolidando-se como uma área fundamental para a descoberta e o desenvolvimento de novos medicamentos. Entre suas principais aplicações, destaca-se a similaridade molecular, baseada no princípio de que “moléculas semelhantes tendem a ter propriedades semelhantes”[Franco et al. 2014]. Esse conceito orienta buscas em bancos de dados químicos, modelagem de propriedades e etapas de *virtual screening*, além de apoiar avaliações regulatórias[Pereira Franco 2015].

A similaridade molecular é particularmente relevante na análise de novos candidatos a fármacos, incluindo os medicamentos órfãos¹. Nesses casos, especialistas avaliam manualmente se um composto difere dos já registrados, um processo suscetível a subjetividade, variações individuais e inconsistências ao longo do tempo[Gandini et al. 2022]. Essa natureza subjetiva também foi observada em estudos que mostram divergências entre avaliadores e a influência de fatores cognitivos e contextuais[Kutchukian et al. 2012].

Modelos de regressão logística com base em *fingerprints* têm sido amplamente utilizados na tarefa[Franco et al. 2014, Pereira Franco 2015, Franco et al. 2017, Gandini et al. 2022] apresentam desempenhos satisfatórios. No entanto, enfrentam limitações inerentes: *fingerprints* capturam apenas parte das características moleculares[Franco et al. 2017], e regressões logísticas são pouco flexíveis frente a relações não lineares e dados complexos, como imagens[Kelleher et al. 2020]. Estudos também reportam erros recorrentes em moléculas maiores e perdas de informação em descritores simplificados[Franco et al. 2014]. Outras abordagens, como medidas baseadas em subgrafo comum máximo (MCS)[Raymond and Willett 2002, Ehrlich and Rarey 2011], oferecem maior detalhamento estrutural, mas têm alto custo computacional e dependem de definições específicas de subestruturas.

Atualmente, não existem muitas pesquisas na utilização de *Deep Learning* para a avaliação de similaridade entre moléculas, com exceção do estudo de [Nasser et al. 2022], que avaliou 3 arquiteturas de *autoencoders* para realizar a redução de dimensionalidade de algumas *fingerprints*, posteriormente utilizadas na avaliação de similaridade molecular com regressões logísticas. Este panorama nos mostra lacunas que ainda não foram exploradas, e novas perspectivas podem ser obtidas na utilização de modelos de *Deep Learning* investigando, se as arquiteturas convolucionais aplicadas diretamente às representações estruturais 2D e 3D são capazes de capturar padrões mais complexos do que os métodos tradicionais baseados em *fingerprints*, tanto sobre os diagramas de estrutura 2D quanto nas representações estruturais 3D das moléculas e, por consequência, nos cálculos de similaridade. Desta forma, espera-se trazer mais informações para auxiliar no desenvolvimento de novos modelos e contribuir em estudos futuros para a área em questão.

Diante disso, este trabalho busca responder à pergunta: “Qual é o desempenho da utilização de técnicas de *Deep Learning* aplicadas na classificação de similaridade entre moléculas?” Nesse contexto, o presente estudo busca preencher essa lacuna, investigando se arquiteturas convolucionais aplicadas diretamente às representações estruturais 2D e 3D são capazes de capturar padrões mais complexos do que os métodos tradicionais baseados em *fingerprints*. Acredita-se que os resultados obtidos poderão contribuir para o avanço da

¹Produtos destinados ao tratamento de doenças raras.

área, indicando novos caminhos para a construção de modelos mais robustos e informativos no processo de avaliação de similaridade molecular.

Para tal, serão avaliadas arquiteturas de *machine learning* aplicadas diretamente às representações estruturais 2D e 3D de moléculas, investigando sua capacidade de capturar padrões complexos e de superar as limitações dos métodos convencionais. Espera-se, assim, contribuir para o avanço de ferramentas mais robustas e adequadas à natureza multidimensional das estruturas moleculares.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os conceitos de similaridade molecular e as principais formas de representação computacional de moléculas; a Seção 3 descreve os algoritmos de *machine learning* utilizados como base e as arquiteturas de *deep learning* consideradas; a Seção 4 discute os trabalhos relacionados que fundamentam a proposta; a Seção 5 detalha a metodologia adotada, os conjuntos de dados e a configuração experimental; a Seção 6 apresenta e analisa os resultados obtidos; e, por fim, a Seção 7 sintetiza as conclusões e indica direções para trabalhos futuros.

2. Similaridade Molecular

Pode-se definir o conceito moderno de molécula como uma estrutura composta por dois ou mais átomos, idênticos ou distintos, unidos pelo compartilhamento de elétrons, formando assim ligações covalentes [Lewis 1916].

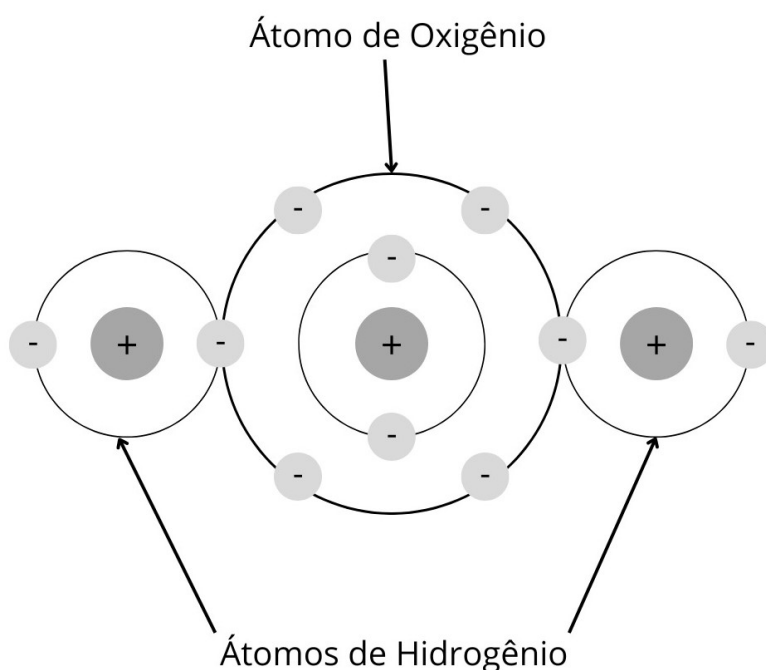


Figure 1. Molécula de água. Fonte: elaborado pela autora.

[Gandini et al. 2022] descreve a similaridade entre moléculas, em especial medicamentos, como uma combinação de três critérios: estrutura molecular, indicação terapêutica e mecanismo de ação, bastando uma divergência significativa em um deles para que sejam considerados não similares. O conceito é naturalmente subjetivo, variando conforme a idade, experiências e estado de espírito [Gandini et al. 2022]; [Pereira Franco 2015] re-

força isso, observando que até um mesmo especialista pode ter percepções diferentes em momentos diversos.

Uma das aplicações mais importantes da similaridade molecular é na aprovação de medicamentos órfãos, voltados para doenças raras, no qual possuem um número de pacientes consideravelmente menor em relação à doenças mais comuns; geralmente com incentivo de exclusividade de mercado [Gandini et al. 2022]. Segundo [Franco et al. 2014], as solicitações de registro são repassadas a órgãos competentes, onde especialistas avaliam e votam pela aprovação ou não do medicamento, incluindo sua classificação como órfão.

Com os avanços da quimioinformática², a quantidade de compostos catalogados é crescente: a plataforma CAS Registry³ reúne aproximadamente 196 milhões de substâncias orgânicas e inorgânicas e 70 milhões de proteínas e ácidos nucleicos, com cerca de 15.000 novos compostos por dia, tornando a representação dessas informações em um padrão compreensível para os computadores um ponto importante da área.

2.1. Representações Computacionais e Descritores Moleculares

[Pereira Franco 2015] cita notações lineares, grafos matemáticos e tabelas de conexão como técnicas comuns para traduzir estruturas químicas a um formato manipulável por computadores. A notação linear mais conhecida, SMILES (*Simplified Molecular Input Line Entry System*), foi criada por David Weininger em 1980 e representa moléculas em sequências compactas de letras e números [Pereira Franco 2015]. Algumas das especificações mais usadas, conforme [Pereira Franco 2015], são:

- Cada átomo é representado pelo seu símbolo químico com a letra maiúscula, exceto o hidrogênio, que está oculto na grande maioria dos casos.
- As ligações duplas são representadas pelo símbolo =.
- As ligações triplas são representadas pelo símbolo #.
- As ramificações das moléculas são representadas por parênteses.
- Os anéis presentes nas moléculas são representados por números.
- Os átomos em anéis aromáticos são representados pelo seu símbolo químico com a letra minúscula.

Descritores moleculares são formas matemáticas de representar características moleculares a partir de um método padronizado, nenhum descritor isolado captura toda a complexidade de uma estrutura molecular; e por conta disso, existe o desenvolvimento de inúmeros descritores com diferentes níveis de complexidade [Pereira Franco 2015], classificados⁴ pela sua dimensionalidade.

Os Descritores Unidimensionais (1D) representam a molécula por um único número — como refração molar, peso molar ou logP — e, segundo [Pereira Franco 2015], são pouco robustos para comparações aprofundadas, servindo mais para prever propriedades.

²O termo quimioinformática (do inglês *chemoinformatics*) é definido por [Brown et al. 1998] como a área de estudo na resolução de problemas da química utilizando-se de técnicas da informática, em especial, na descoberta de novos medicamentos. Segundo [Pereira Franco 2015], esse termo surgiu como uma nova definição para um campo já conhecido da química combinatória.

³Disponível em: <https://www.cas.org/cas-data/cas-registry>. Acesso em: jun. 2025.

⁴Apesar de [Pereira Franco 2015] usar esta determinada classificação, existem autores que usam outras nomenclaturas. Para este trabalho, será considerada a nomenclatura e classificação elaborada por [Pereira Franco 2015] como base.

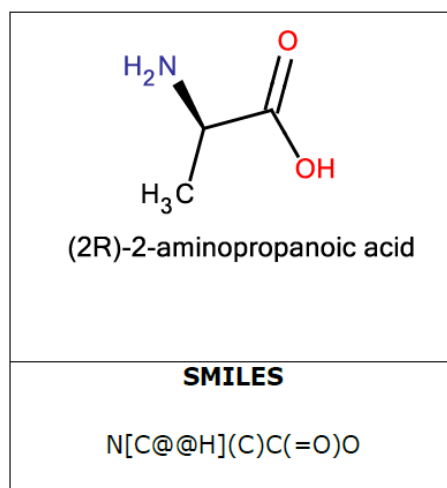


Figure 2. Representação SMILES da molécula Alanina (Ácido 2-aminopropiônico).
Fonte: [Pereira Franco 2015].

Já os Descritores Bidimensionais (2D) representam a molécula por grafos moleculares, capturando mais informação e sendo usados na busca de subestruturas, similaridade e QSAR (*Quantitative Structure-Activity Relationship*) [Pereira Franco 2015]. O autor cita como exemplos:

- *2D fingerprints*: um vetor de binários em que cada posição indica a presença (1) ou a ausência (0) de uma subestrutura específica, dividindo a molécula em conjuntos predefinidos de subestruturas.

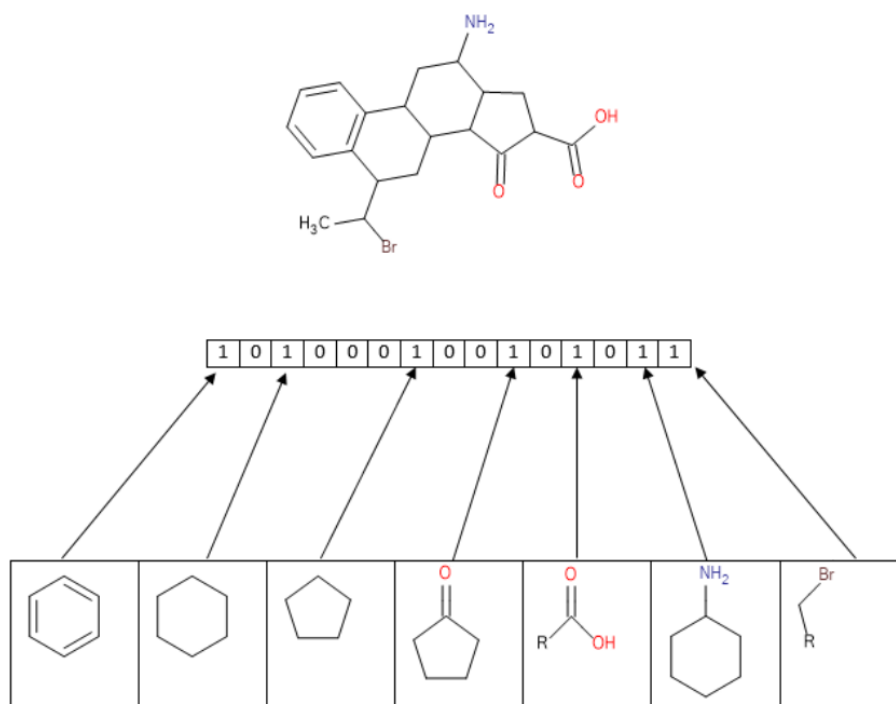


Figure 3. Exemplo de um descritor 2D fingerprint. Fonte: [Pereira Franco 2015].

- *Hashed fingerprints*: padrões coletados das subestruturas moleculares e transformados em código *hash* a partir de um algoritmo *hash*.

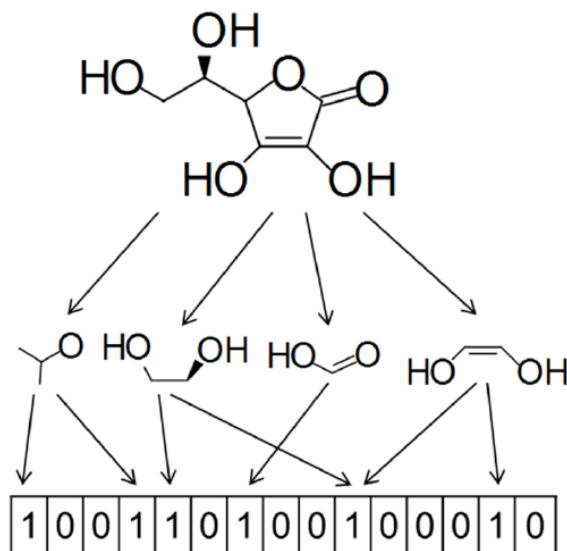


Figure 4. Exemplo de um descritor hashed fingerprints. Fonte: [Śmieja and Warszycki 2016].

- *Circular fingerprints*: informações geradas a partir de círculos centrados em cada átomo, com diâmetros crescentes, capturando átomos e ligações até o círculo máximo.

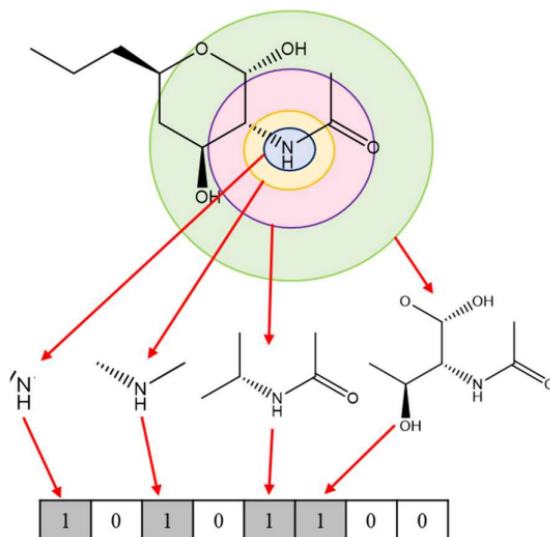


Figure 5. Exemplo de um descritor circular fingerprint. Fonte: [Zarnecka 2018].

Já os Descritores Tridimensionais (3D) representam a molécula a partir de seu arranjo espacial; [Niazi and Mariam 2023] citam como exemplos os descritores de superfície: volume, autocorrelação, quantum, 3D-MoRSE, GETAWAY e WHIM, exemplificando [Pereira Franco 2015] o seguinte tipo:

- 3D *screens*: descritores que armazenam, em sequências de bits, ângulos e distâncias entre átomos e partes da molécula, com a faixa de distância de cada par dividida em intervalos.



Figure 6. Ácido quisqualico na sua representação 3D. Fonte: [Ebejer et al. 2012].

2.2. Limitações dos Métodos Tradicionais

A conversão de estruturas moleculares em descritores numéricos permite medidas quantitativas de similaridade, das quais o Coeficiente de Tanimoto — métrica sobre dados binários entre zero e um — é uma das mais difundidas [Pereira Franco 2015]. Apesar de objetivo, esse coeficiente trata subestruturas moleculares com pesos equivalentes, ignorando que certas características farmacofóricas são mais determinantes para a atividade biológica; isso contrasta com a avaliação humana, complexa e estocástica, que pondera implicitamente o contexto terapêutico e a experiência prévia [Gandini et al. 2022], podendo falhar em moléculas de topologias distintas e mesma bioatividade.

Nesse cenário, *Machine Learning* surge como uma abordagem capaz de modelar essa complexidade, estruturando a avaliação de similaridade como classificação supervisionada, em que se induz $f : X \rightarrow Y$ do espaço de características moleculares X para a classe de similaridade Y . As primeiras abordagens usaram modelos lineares, como o caso de [Franco et al. 2014], com *fingerprints* como entrada de uma Regressão Logística, mas [Kelleher et al. 2020] aponta que tais modelos apresentam um bom desempenho apenas quando os dados são linearmente separáveis, dependendo ainda da qualidade dos descritores pré-calculados (*feature engineering*).

O Aprendizado Profundo (*Deep Learning*) surge para superar essas limitações, usando Redes Neurais Artificiais com múltiplas camadas para modelar abstrações de alto nível, com vantagem no aprendizado de representação (*representation learning*): em vez de exigir características pré-extraídas, opera sobre dados brutos — como a imagem do diagrama 2D ou a matriz de coordenadas 3D —, captando características que escapam às *fingerprints* tradicionais e superando as barreiras de subjetividade dos métodos manuais.

3. Algoritmos de *Machine Learning*

No presente trabalho, foram utilizados algoritmos de aprendizado supervisionado. Segundo [Murphy 2012], estes métodos possuem como característica principal, dado um

determinado conjunto de dados rotulados com pares de entradas e saídas $D = \{(x_i, y_i)\}_{i=1}^N$, a realização de um mapeamento das entradas x_i para suas saídas y_i .

A seguir, são descritos os algoritmos citados e utilizados no estudo.

3.1. Regressão Logística

Há casos em que uma equação da regressão linear não consegue abranger todos os pontos em uma reta; para isso, usamos uma adequação dessa regressão, chamada de regressão logística. [James et al. 2023] mostra que essa adequação ocorre pelo fato de que a regressão logística modela a probabilidade de que y pertença a uma determinada classificação ou categoria que está sendo avaliada. Para as regressões logísticas, a variável dependente é categórica. O autor ainda explica que a equação básica de uma regressão logística pode ser definida da seguinte forma:

$$P(Y = 1|X) = \frac{e^{(\beta_0 + \beta_1 x_1^i)}}{e^{(\beta_0 + \beta_1 x_1)} + 1} \quad (1)$$

onde: $P(Y = 1|X)$ denota a probabilidade condicional de que y seja igual a 1 (determinado evento ocorrer), dado o valor da variável x_1^i ; e denota o número de Euler⁵; β_0 denota o intercepto; β_1 denota o coeficiente angular; e x_1 indica o conjunto de todas as observações da variável x_1 .

Segundo [Kelleher et al. 2020], assim como na regressão linear, pode se medir o ajuste do modelo da regressão logística utilizando o SSE de forma adaptada, juntamente com o gradiente descendente. As duas fórmulas para β_0 e β_1 , adaptadas, seguem abaixo:

$$\frac{\partial}{\partial \beta_0} = \left(\sum_{i=1}^m (\beta_0 + \sum_{j=1}^n \beta_j x_j^i - y^i) \right) \sigma' \quad (2)$$

$$\frac{\partial}{\partial \beta_1} = \left(\sum_{i=1}^m (\beta_0 + \sum_{k=1}^n \beta_k x_k^i - y^i) x_j^i \right) \sigma' \quad (3)$$

onde: σ' denota a derivada da equação logística; o restante dos componentes nas equações segue a mesma definição das equações de regressão linear.

No presente trabalho, a regressão logística foi utilizada como modelo de *baseline*, com base nos modelos construídos por [Gandini et al. 2022], permitindo a comparação direta com os modelos de *Deep Learning* propostos.

3.2. Redes Neurais Convolucionais

A rede neural convolucional (do inglês *Convolutional Neural Network* – CNN) possui esta nomenclatura por utilizar a convolução em pelo menos uma de suas camadas e é comumente aplicada em problemas que envolvem dados com uma estrutura em grade, como é o caso das imagens. A convolução se trata de uma operação matemática de operações

⁵Valor aproximado de 2.71828.

lineares, na qual são utilizados filtros (do inglês *kernels*) para extrair características locais dos dados de entrada, eliminando os ruídos [Goodfellow et al. 2016].

[Goodfellow et al. 2016] define a fórmula de uma convolução como:

$$S_{(i,j)} = (I * K) = \sum_{i=1}^m \sum_{j=1}^n I(i+m, j+n) \times K(m, n) \quad (4)$$

onde: (i, j) denota as coordenadas espaciais da saída; m e n denotam as dimensões do filtro; I denota a entrada; K denota o filtro; $*$ denota a operação de convolução; e S denota a saída.

[Hijazi et al. 2015] define que os principais componentes de uma CNN são as camadas de convolução e as camadas de *pooling*. A camada de convolução tem como objetivo principal a extração de diferentes características da entrada: nas camadas mais iniciais, geralmente são extraídas características mais simples, como linhas e bordas; e nas camadas mais avançadas, características mais detalhadas. O filtro percorre a imagem de entrada desde o início, deslocando-se ponto a ponto até que toda a imagem seja varrida.

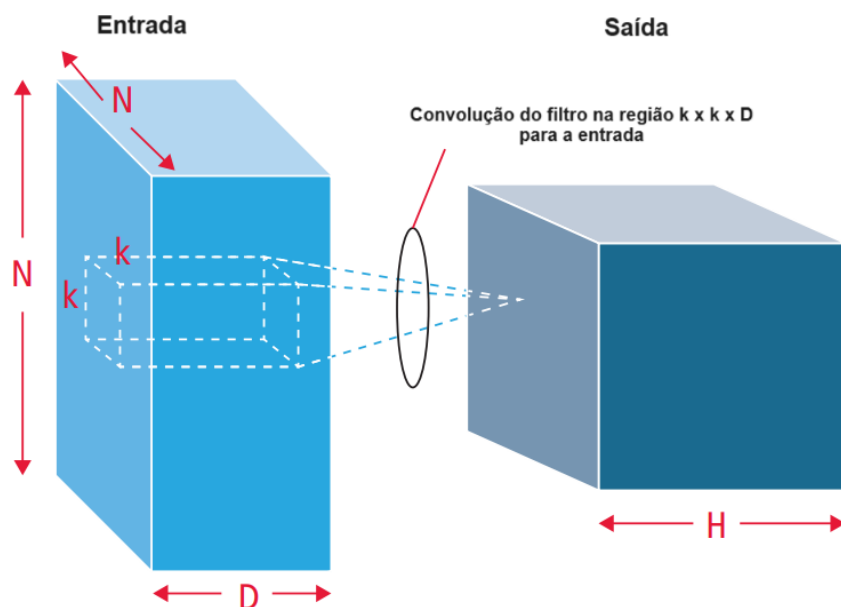


Figure 7. Ilustração de uma camada de convolução. Fonte: [Hijazi et al. 2015]. Traduzido e adaptado pela autora.

Já nas camadas de *pooling*, é realizado o processo de redução da resolução das características por uma outra saída estatisticamente mais próxima. A realização desse processo pode ocorrer de duas maneiras: pelo *max pooling* (seleciona-se o valor máximo da região) e pelo *average pooling* (seleciona-se o valor médio da região).

No presente trabalho, foram construídas CNNs adaptadas da arquitetura LeNet-5, capazes de processar simultaneamente as representações 2D e 3D das moléculas, conforme detalhado na Seção 5.1.

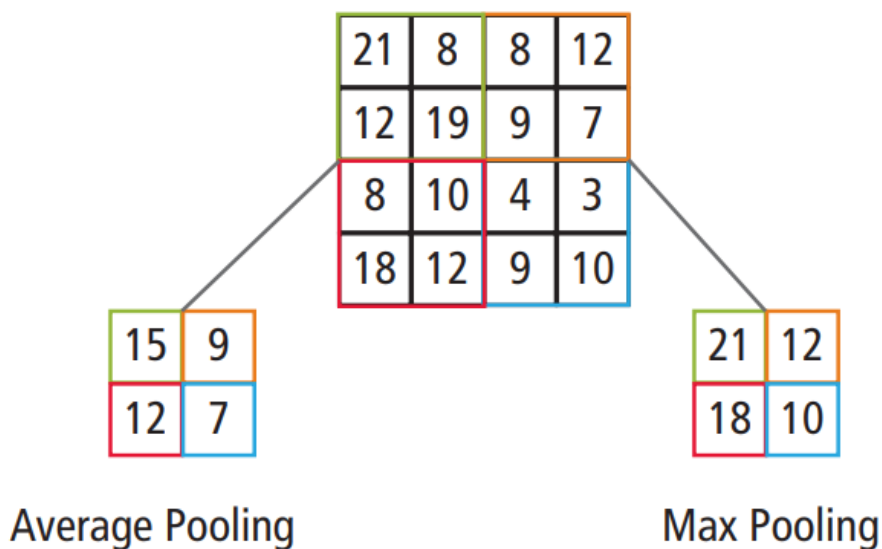


Figure 8. Representação do max pooling e average pooling. Fonte: [Hijazi et al. 2015].

4. Trabalhos Relacionados

Em [Franco et al. 2014], os autores investigaram a utilização de 6 tipos de fingerprints 2D de código fechado, a partir do diagrama de estrutura 2D, para avaliar a similaridade estrutural entre pares de moléculas no contexto da legislação de medicamentos órfãos da União Europeia, comparando os resultados de modelos de *machine learning* ao julgamento de especialistas. O estudo coletou o julgamento de 143 especialistas para respostas de sim ou não na avaliação de similaridade de 100 pares de moléculas do banco DrugBank 3.0; para os mesmos pares, aplicaram cada uma das 6 *fingerprints* selecionadas (BCI, Daylight, ECFC4, ECFP4, MDL e Unity), calcularam a similaridade pelo Coeficiente de Tanimoto e construíram um modelo de regressão logística para cada uma, avaliando seu desempenho preditivo em relação às respostas dos especialistas.

Os modelos apresentaram um bom poder preditivo e forte compatibilidade com os especialistas, utilizando boas métricas de avaliação (tROC, Probability, Sensitivity, Specificity, Precision, Acurácia, F, Youden e Matthews); a *fingerprint* BCI apresentou o melhor desempenho geral, sendo considerada pelos autores a melhor escolha para estudos futuros. Entretanto, o estudo se limitou a 6 *fingerprints*, cada uma exigindo um modelo próprio — já que cada uma foca em determinadas características, com algumas se saindo melhor que outras —, evidenciando a falta de uniformização e a adaptabilidade de um único modelo para conseguir extrair todas as características de cada molécula. Por utilizar apenas o diagrama 2D, muitas informações moleculares foram desconsideradas, e outras 23 características testadas obtiveram resultados inferiores às *fingerprints*, expondo a limitação da regressão logística frente à complexidade dos dados. O conjunto de teste careceu de diversidade (5 classes farmacológicas frente às 42 do treino), os dados são confidenciais, a coleta de opiniões não foi aleatória, e não foram consideradas representações 3D.

Já em [Franco et al. 2017], os autores aprofundaram o estudo com 12 *fingerprints* de código aberto (CDK Extended, CDK Standard, Estate, PubChem, MACCS, Morgan, Feat Morgan, Atom Pair, Torsion, RDKit, Avalon e Layers) sobre o mesmo conjunto de

100 pares e as mesmas respostas dos 143 especialistas, buscando superar a restrição de acesso das *fingerprints* do trabalho anterior. Os modelos também apresentaram bom poder preditivo: quase todas as *fingerprints*, exceto Estate, Atom Pair e RDKit, preveram com êxito 95 ou mais pares de teste, e a Morgan errou apenas um — o mesmo par que errou em todos os demais modelos, referente às estruturas mais complexas do conjunto, com os maiores pesos moleculares (superiores a 500).

Assim como no estudo anterior, os autores utilizaram boas métricas de avaliação, mas o estudo manteve as mesmas limitações: um modelo específico por *fingerprint*, perda de informação por usar somente o diagrama 2D, dados de teste confidenciais e indisponíveis, coleta não aleatória de respostas e ausência de representações 3D.

No trabalho de [Gandini et al. 2022], os autores construíram modelos de *machine learning* para prever a percepção humana de similaridade molecular, utilizando tanto o diagrama 2D quanto a representação 3D das moléculas, e disponibilizaram uma aplicação online que recebe a representação SMILES ou o diagrama 2D de um par e retorna o percentual de similaridade. Para 100 pares do banco ChEMBL 27, avaliados por 418 usuários (5 pares aleatórios cada), aplicaram a *fingerprint* CDK Extended ao diagrama 2D e o software ROCS da OpenEye à representação 3D, calculando a similaridade pelos Coeficientes de Tanimoto e Tanimoto Combo, respectivamente; construíram, então, dois modelos de regressão logística: um para os diagramas 2D (tXT) e outro para as representações 3D (tCS).

Os modelos tiveram bom desempenho na fase de treino, com o conjunto de [Franco et al. 2017], mas mostraram-se inferiores na fase de teste, com o conjunto próprio dos autores. Observou-se que os modelos baseados apenas no diagrama 2D já se correlacionavam bem com as respostas dos usuários, mesmo com a melhora trazida pela adição das representações 3D; os autores levantaram a hipótese de que, nos casos de dúvida, os usuários tendem a priorizar mais as informações em diagramas 2D para a sua resposta final.

Entre os pontos positivos do trabalho estão a implantação pública e online do modelo e a coleta aleatória de respostas por usuários. Por outro lado, os autores delimitaram o *threshold* de classificação de forma manual, quando o mais adequado seria uma métrica como o ROC score (AUC); e avaliaram os modelos somente pela acurácia, uma medida pouco apropriada para dados não lineares e complexos. A adição da representação 3D trouxe ganhos, mas ainda exigiu modelos separados, dificultando um modelo único e mais robusto — papel que um modelo de *machine learning* baseado em redes neurais artificiais cumpriria de forma mais abrangente.

Analisando-se os trabalhos relacionados, podemos observar que o progresso de modelos para a atividade de avaliação de similaridade molecular foi desenvolvida predominantemente por regressões logísticas a partir da utilização de *fingerprints*. Os bons resultados obtidos pelos modelos desenvolvidos ainda dependem da necessidade de transformar as informações das moléculas para que as regressões possam ser aplicadas, além do fato de que cada *fingerprint* utilizada necessita de uma regressão logística específica. A partir desse contexto, o presente trabalho implementou e analisou uma abordagem alternativa, baseada em Redes Neurais Convolucionais, para explorar o comportamento de modelos quando construídos para se utilizar informações diretamente sobre os diagramas

2D e as suas representações estruturais 3D das moléculas. Para isso, foi utilizado o mesmo conjunto de dados do trabalho de [Gandini et al. 2022], garantindo assim, uma comparação direta do trabalho aqui desenvolvido com o do autor. O estudo contribuiu com a construção e análise de novos modelos de *machine learning* para a área de similaridade molecular.

5. Metodologia e Experimentos

O desenvolvimento deste estudo e a análise do desempenho dos modelos de *machine learning* para a avaliação de similaridade molecular seguem um fluxo de trabalho inspirado na metodologia CRISP-DM [Chapman 2000]. Esta abordagem, adotada em ciência de dados devido à sua flexibilidade, fundamentou a estruturação da pesquisa, conforme ilustrado na Figura 9.

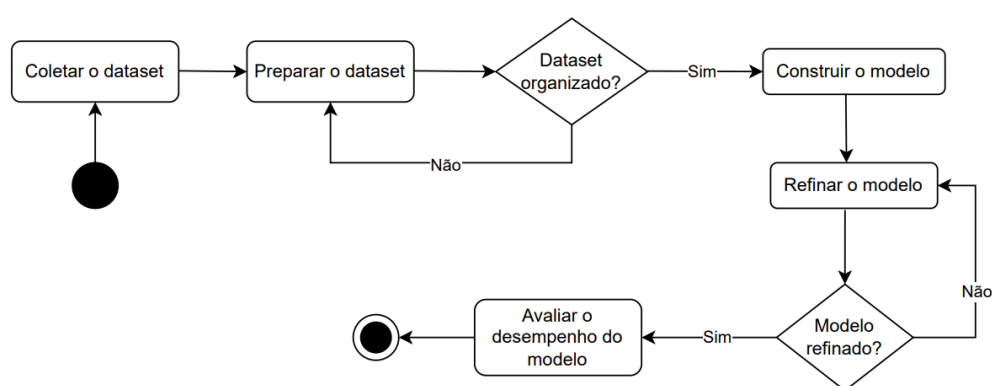


Figure 9. Fluxograma da metodologia aplicada. Fonte: elaborado pela autora.

A etapa inicial consistiu na revisão da literatura e no entendimento dos métodos atuais de predição, identificando a necessidade de explorar modelos não lineares frente à subjetividade da avaliação humana. A partir desse contexto, definiu-se a estratégia para a coleta e preparação dos dados que sustentam os experimentos.

Para a realização deste estudo, foram utilizados dois conjuntos de dados distintos, disponibilizados na plataforma Kaggle⁶ e consolidados no estudo de [Gandini et al. 2022]. A seleção destes conjuntos justifica-se pela presença de avaliações de especialistas humanos (*ground truth*), elemento crucial para treinar modelos que mimetizem a percepção de similaridade. O detalhamento do rigoroso processo de coleta realizado pelos autores originais é descrito a seguir.

O primeiro conjunto de dados foi originalmente coletado e processado por [Pereira Franco 2015] para o treinamento de seus modelos. A partir do banco de dados Drug Bank 3.0, o autor selecionou inicialmente moléculas que possuíam pelo menos um átomo de carbono em sua estrutura SMILES. Na sequência, aplicou-se o filtro de Lipinski para restringir a seleção a compostos com características farmacológicas adequadas, selecionando moléculas com no máximo 10 receptores de ligações de hidrogênio, no máximo 5 doadores, massa molecular inferior a 500 Daltons e logP máximo de 5. Após a filtragem, realizou-se o cálculo do Coeficiente de Tanimoto para os 1.140.624

⁶Disponível em: <https://www.kaggle.com/datasets/wary-slabs/molecular-similarity-pairs-2d3d-and-human-ratings>. Acesso em: jun. 2025.

pares moleculares resultantes. Para evitar viés de seleção, foram escolhidas 200 moléculas ativas e, dentre estas, selecionados aleatoriamente 40 pares distribuídos em 5 intervalos do Coeficiente de Tanimoto (0,0–0,19; 0,2–0,39; 0,4–0,59; 0,6–0,79; 0,8–1,0). O conjunto final foi composto por 100 pares selecionados aleatoriamente. A validação humana foi obtida através de formulários enviados por e-mail a 143 especialistas, que classificaram os pares como similares ou não similares.

Com os rótulos deste conjunto, foi construído um histograma, conforme apresentado na figura 10, no qual é possível observar algumas informações: a existência de uma concentração extremamente alta no valor próximo de 0 (classe não similar); uma outra concentração relativamente alta nos valores próximos de 1 (classe similar); e poucas amostras nos valores intermediários.

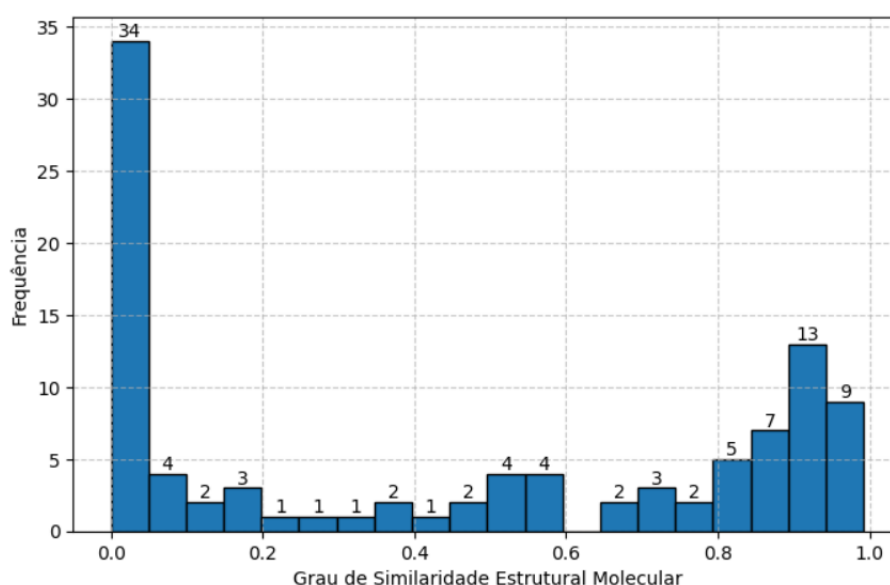


Figure 10. Distribuição dos valores do grau de similaridade estrutural molecular do conjunto de treino. Fonte: elaborado pela autora.

O segundo conjunto de dados foi elaborado por [Gandini et al. 2022] visando a fase de teste e a mitigação de limitações do conjunto anterior. A coleta iniciou-se no banco de dados ChEMBL 27, focando em três alvos biológicos específicos: HERG, 5HT2B e CYP2D6, escolhidos por possuírem sítios de ligação bem definidos. Utilizou-se o InChikey para a remoção de compostos duplicados. Diferentemente do primeiro conjunto, aplicaram-se protocolos para gerar conformadores tanto em 2D quanto em 3D, descartando compostos que falharam nesta etapa. A seleção dos pares seguiu uma estratégia de estratificação baseada na divergência entre representações: foram selecionados aleatoriamente 3.000 pares para cada alvo, divididos em quatro subgrupos de acordo com a concordância ou discordância de similaridade entre os diagramas 2D e as estruturas 3D. Ao final, selecionaram-se 25 pares de cada subconjunto, totalizando 100 pares. A avaliação destes foi realizada via plataforma web, onde 418 usuários julgaram 5 pares aleatórios cada.

Para este conjunto de dados também foi construído um histograma com os seus rótulos, presente na figura 11; onde podemos observar que: existe uma quantidade considerável de amostras nas faixas de valores intermediárias; e a sua distribuição é mais espalhada.

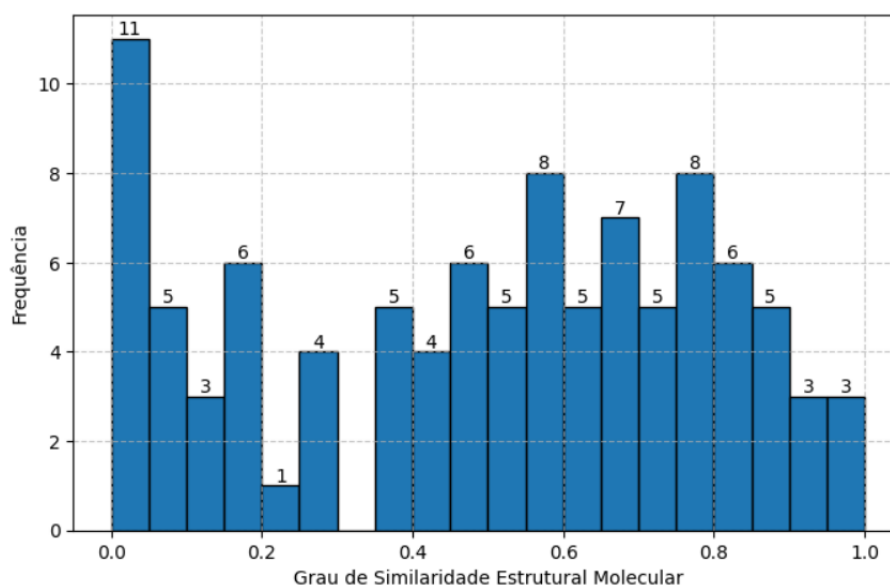


Figure 11. Distribuição dos valores do grau de similaridade estrutural molecular do conjunto de teste. Fonte: elaborado pela autora.

Ambos os conjuntos possuem os diagramas de estrutura 2D, as representações estruturais 3D, os arquivos contendo as avaliações de similaridade e as suas representações SMILES.

Após a coleta, procedeu-se com a preparação dos dados. As informações foram analisadas para verificar a necessidade de transformações e agregações. Para manter o mesmo estado comparativo com o trabalho de [Gandini et al. 2022], optou-se em manter a mesma configuração pelo autor na divisão dos dados de treino e teste: 50% dos dados, conjunto de dados de [Pereira Franco 2015], destinados para a etapa de treino; e 50% dos dados, conjunto de dados de [Gandini et al. 2022], destinados para a etapa de teste. Reconhece-se, no entanto, que conjuntos de 200 pares moleculares representam um volume reduzido para o treinamento de redes convolucionais; como estratégia de mitigação, foram aplicadas técnicas de *data augmentation* durante o treinamento dos modelos, conforme descrito na Seção 5.1.

Na fase de modelagem, foram implementados modelos de *Deep Learning* para serem aplicados aos dados de teste. Para fins de comparação direta e estabelecimento de um *baseline*, os resultados obtidos foram comparados aos modelos de regressão logística reportados por [Gandini et al. 2022], construídos sobre os mesmos conjuntos de dados utilizados neste estudo; essa comparação é realizada com base nos resultados publicados pelos autores e não por meio de uma reimplementação independente do modelo.

A avaliação do desempenho dos modelos construídos foi realizada utilizando métricas padrão de classificação, conforme abordadas por [Kelleher et al. 2020]. Foram analisadas a matriz de confusão, a precisão (*precision*), a revocação (*recall*), o *F1-score* e a área sob a curva ROC. Adicionalmente, para manter a comparabilidade com os trabalhos relacionados, reportou-se a taxa de acertos (acurácia), ressaltando-se suas limitações em cenários de dados complexos. Por se tratar de um estudo focado na análise de desempenho, a fase de implantação prevista no CRISP-DM não foi realizada.

5.1. Experimentos

Para analisar o desempenho de técnicas de *deep learning* na classificação de similaridade molecular, foram construídos 3 modelos de *machine learning* na arquitetura LeNet-5, de forma adaptada, utilizando as bibliotecas apresentadas na Tabela 1. Optou-se por esta arquitetura por se tratar de uma rede neural relativamente simples e de fácil implementação [LeCun et al. 1998].

Table 1. Versões das bibliotecas utilizadas

Biblioteca	Versão
tensorflow	2.19.0
scikit-learn	1.6.1
pandas	2.2.2
bayesian-optimization	3.1.0
biopython	1.86

A biblioteca *biopython* foi utilizada para o processamento e leitura dos arquivos de estruturas moleculares nos formatos padrão da área (*e.g.*, SDF e MOL2), etapa necessária para a geração das representações de entrada dos modelos.

Cada um dos modelos construídos foram denominado de Modelo A, B e C, respectivamente. A arquitetura LeNet-5 foi adaptada para o processamento simultâneo de moléculas em imagens 2D e 3D, a partir da organização de duas redes paralelas em um único modelo; mantendo-se o seu princípio da extração hierárquica de características através de camadas convolucionais seguidas de *pooling*. Também foi adaptada a entrada do modelo, visando atender as características das imagens, além do fato de serem analisadas em pares. Na saída do modelo, também houve uma adaptação, onde a mesma foi transformada em um único valor contínuo que representa o grau de similaridade do par de moléculas analisado; e, por fim, outras melhorias adicionais para testes de refinamento, como compartilhamento de pesos entre entradas pareadas, fusão multimodal por concatenação de características, convolução 3D, regularização L2, *dropout*, *data augmentation*, *early stopping* e otimização bayesiana de hiperparâmetros.

Os três modelos possuem como entrada um par de moléculas. A rede 2D é formada por duas camadas convolucionais (Conv2D) com função de ativação ReLU, cada uma seguida por uma camada de *Average Pooling* 2D, após a qual uma camada de *Flatten* precede uma camada densa responsável pela geração dos vetores de características. Já na rede 3D, são também duas camadas convolucionais (Conv3D) com função de ativação ReLU, cada uma seguida por uma camada de *Max Pooling* 3D, com estrutura análoga de *Flatten* e camada densa ao final. As duas saídas são concatenadas e então processadas por três camadas densas, das quais se gera o índice de similaridade molecular do par. Uma ilustração detalhada do modelo pode ser observada na Figura 12.

A configuração detalhada das camadas de cada modelo, com as respectivas taxas de aprendizado, épocas de treinamento e valores de *batch size*, é apresentada na Tabela 2.

Os Modelos A e C compartilham a mesma arquitetura de camadas, taxa de aprendizado, épocas e *batch size*, diferindo exclusivamente na estratégia de definição do threshold de decisão na fase de avaliação: o Modelo A e B utilizou um valor de *threshold* fixo

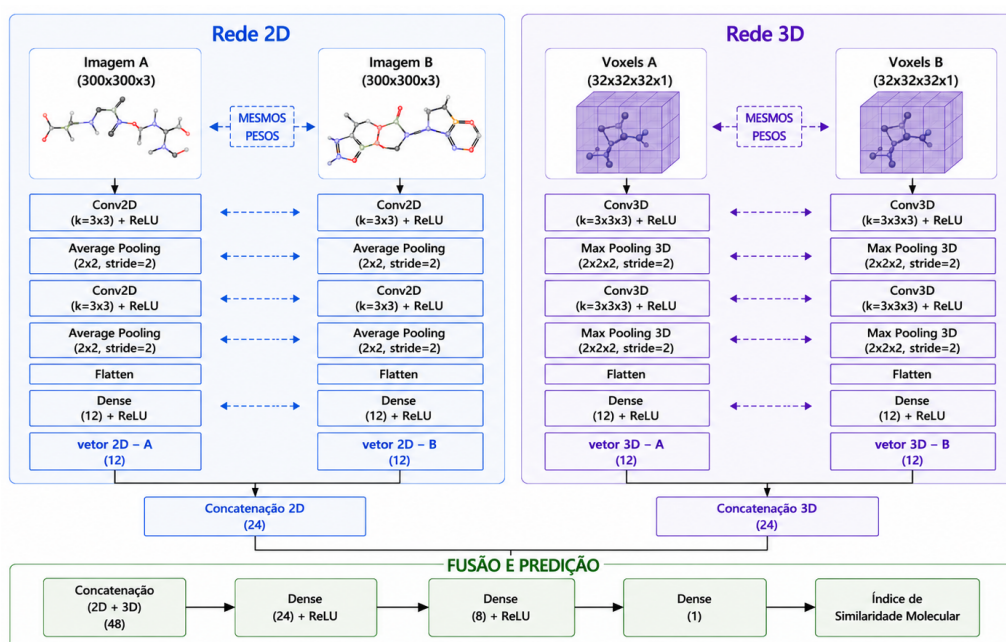


Figure 12. Arquitetura geral dos modelos de machine learning construídos. Fonte: elaborado pela autora.

(mesma estratégia utilizada por [Gandini et al. 2022], a fim de se manter um mesmo estado comparativo com o autor), enquanto o Modelo C utilizou um *threshold* determinado automaticamente a partir dos dados de validação. O Modelo B, por sua vez, difere dos demais tanto na arquitetura quanto nos hiperparâmetros, incorporando otimização bayesiana para a busca dos melhores valores de filtros e camadas densas, além de regularização L2 e *dropout* explícitos.

6. Resultados

Para a avaliação do desempenho dos modelos, foram utilizadas as métricas de curva ROC, *Precision*, *Recall*, *F1-score* e acurácia. Em conjunto, foram realizados testes e obtidos resultados com duas estratégias de atribuição ao valor de *threshold*: um valor de *threshold* fixo e outro com valor automático.

Os resultados obtidos apresentam diferenças significativas entre os modelos e a estratégia de atribuição do valor de *threshold*. É importante ressaltar que, em função de restrições de tempo computacional durante a fase de experimentação, os Modelos A e B foram avaliados apenas com a estratégia de *threshold* fixo, enquanto o Modelo C foi avaliado apenas com a estratégia de *threshold* automático. Essa assimetria limita a comparação direta entre os três modelos sob uma mesma condição experimental, sendo retomada como limitação na Seção 7.

Na tabela 3, apresentam-se os resultados com *threshold* fixo; o modelo A apresentou um comportamento relativamente equilibrado entre as classes, com boa tendência de estabilidade na classificação. Para a classe 1, o valor de *F1-score* de 0,6981 revela uma capacidade moderada de identificação correta para a mesma, ainda que distante dos valores reportados na literatura para modelos baseados em *fingerprints* e regressão logística (Seção 6.1). Já o modelo B mostrou um desempenho desbalanceado entre as classes, não

Table 2. Arquiteturas dos modelos

Parâmetro	Modelo A	Modelo B	Modelo C
Camada convolucional 2D	4 filtros, kernel 3x3, ReLU Avg. Pooling 2x2, stride 2 18 filtros, kernel 3x3, ReLU Avg. Pooling 2x2, stride 2	6–32 filtros (step 6) ou 6 fixos, kernel 3x3, ReLU, L2 (0.0005) Avg. Pooling 2x2, stride 2 12–36 filtros (step 6) ou 18 fixos, kernel 3x3, ReLU, L2 (0.0005) Avg. Pooling 2x2, stride 2	4 filtros, kernel 3x3, ReLU Avg. Pooling 2x2, stride 2 18 filtros, kernel 3x3, ReLU Avg. Pooling 2x2, stride 2
Camada convolucional 3D	14 filtros, kernel 3x3x3, ReLU Max Pooling 2x2x2, stride 2 18 filtros, kernel 3x3x3, ReLU Max Pooling 2x2x2, stride 2	2–16 filtros (step 2) ou 2 fixos, kernel 3x3x3, ReLU, L2 (0.0005) Max Pooling 2x2x2, stride 2 8–32 filtros (step 2) ou 18 fixos, kernel 3x3x3, ReLU, L2 (0.0005) Max Pooling 2x2x2, stride 2	14 filtros, kernel 3x3x3, ReLU Max Pooling 2x2x2, stride 2 18 filtros, kernel 3x3x3, ReLU Max Pooling 2x2x2, stride 2
Camada Densa	1 camada de 24 1 camada de 8	1 camada de 4–64 (step 8) ou 4 fixos dropout de 0.2–0.5 (step 0.1) ou 0.4 fixo 1 camada de 8–128 (step 16) ou 8 fixos dropout de 0.2–0.5 (step 0.1) ou 0.4 fixo	1 camada de 24 1 camada de 8
Taxa de aprendizado	0.0005	0.0001	0.0005
Épocas	100	50	100
Batch size	2	32	2
Threshold	Fixo	Fixo	Automático

conseguindo identificar corretamente nenhum valor da classe 0. Essa performance, aliada a uma acurácia aparentemente elevada (0,7100), sugere um possível viés do modelo para a classe 1, comportamento discutido em detalhe na Seção 6.2.

Table 3. Resultados com threshold fixo

Modelo	Curva ROC	Precision	Recall	F1-score	Acurácia
Modelo A	0.6356	0 0.6327	0 0.6889	0 0.6596	0.6800
		1 0.7255	1 0.6727	1 0.6981	
Modelo B	0.5838	0 0.0000	0 0.0000	0 0.0000	0.7100
		1 0.7100	1 1.0000	1 0.8304	

Já na tabela 4, visualizamos os resultados com o *threshold* automático; e o modelo C obteve uma melhor performance de modo geral em termos de acurácia (0,8200). Entretanto, essa métrica isolada esconde uma fragilidade importante: o *recall* de apenas 0,0526 para a classe 0 indica que o modelo identificou corretamente menos de 6% dos pares efetivamente dissimilares do conjunto de teste, classificando a grande maioria deles incorretamente como similares. Esse padrão, combinado ao *recall* perfeito (1,0000) para a classe 1, sugere que o ganho de acurácia do Modelo C decorre majoritariamente do acerto da classe majoritária, e não de uma capacidade discriminativa superior, demonstrando ainda uma dificuldade relevante na identificação da classe 0.

Table 4. Resultados com treshold automático

Modelo	Curva ROC	Precision		Recall		F1-score		Acurácia
Modelo C	0.5627	0	1.0000	0	0.0526	0	0.1000	0.8200
		1	0.8182	1	1.0000	1	0.9000	

6.1. Comparação com o *Baseline* da Literatura

Conforme estabelecido na metodologia (Seção 5.1), o trabalho de [Gandini et al. 2022] foi utilizado como referência de *baseline*, por apresentar modelos de regressão logística construídos sobre os mesmos dois conjuntos de dados utilizados neste estudo (o conjunto coletado pelos próprios autores e o conjunto de [Pereira Franco 2015]). As tabelas 5 e 6 reproduzem os resultados reportados pelos autores para os modelos *single-feature* (t_{XT} , baseado no diagrama estrutural 2D; t_{CS} , baseado na representação 3D) e *double-feature* (combinação de t_{XT} e t_{CS}).

Table 5. Modelos treinados no conjunto coletado por [Gandini et al. 2022] e validados no conjunto de [Pereira Franco 2015]

Tipo de modelo	Variáveis	Fit		Validação	
		$N_{correct}$	ROCAUC	$N_{correct}$	ROCAUC
<i>single-feature</i>	t_{XT}	81	0.920	92	0.988
	t_{CS}	70	0.845	92	0.970
<i>double-feature</i>	t_{XT}, t_{CS}	84	0.924	95	0.988

Table 6. Modelos treinados no conjunto de [Pereira Franco 2015] e validados no conjunto coletado por [Gandini et al. 2022]

Tipo de modelo	Variáveis	Fit		Validação	
		$N_{correct}$	ROCAUC	$N_{correct}$	ROCAUC
<i>single-feature</i>	t_{XT}	93	0.988	81	0.920
	t_{CS}	91	0.970	69	0.845
<i>double-feature</i>	t_{XT}, t_{CS}	95	0.988	81	0.916

A comparação entre os resultados de [Gandini et al. 2022] e os modelos construídos neste trabalho revela uma diferença expressiva de desempenho. Os modelos de regressão logística dos autores atingiram ROCAUC entre 0,845 e 0,988, enquanto os Modelos A, B e C deste estudo obtiveram ROCAUC entre 0,5627 e 0,6356 — valores próximos ao limite de uma classificação aleatória (ROCAUC = 0,5). Esse contraste indica que, ao menos na configuração experimental adotada, a arquitetura de *Deep Learning* proposta não foi capaz de superar o *baseline* de regressão logística sobre *fingerprints*, contrariando a expectativa inicial de que modelos convolucionais aplicados diretamente às representações 2D e 3D capturassem padrões estruturais mais ricos.

Alguns fatores ajudam a explicar essa diferença. Primeiro, os modelos de [Gandini et al. 2022] operam sobre descritores numéricos (*fingerprint CDK Extended*

e o Coeficiente de Tanimoto Combo calculada via ROCS), que já encapsulam conhecimento de domínio cuidadosamente desenvolvido pela comunidade de quimioinformática ao longo de décadas; já as redes convolucionais construídas neste trabalho precisam aprender, a partir de diagrama de estrutura 2D e representações 3D brutos, tanto a extração de características relevantes quanto a função de similaridade propriamente dita — uma tarefa consideravelmente mais difícil quando o volume de dados de treinamento é pequeno (200 pares de moléculas). Segundo, modelos de regressão logística possuem poucos parâmetros e, por isso, tendem a generalizar melhor em regimes de dados escassos, ao passo que arquiteturas convolucionais, mesmo em arquiteturas mais simples como a LeNet-5 adaptada e utilizada aqui, possuem maior capacidade de *overfitting* quando treinadas com poucas centenas de exemplos.

6.2. Análise do Desbalanceamento de Classes

Um padrão recorrente nos Modelos B e C é a baixa capacidade de identificação da classe 0 (pares dissimilares), com métricas de *precisão*, *recall* e *F1-score* próximas de zero para essa classe, enquanto a classe 1 (pares similares) é identificada com *recall* próximo ou igual a 1,0000 em ambos os modelos. Esse comportamento é uma assinatura típica de desbalanceamento de classes ou de calibração inadequada do limiar de decisão e merece ser discutido com mais profundidade.

Não foi possível, no momento da redação desta seção, confirmar com exatidão a proporção real de pares similares e dissimilares no conjunto de dados utilizado. O *dataset* de [Gandini et al. 2022] foi construído com estratificação em quatro subgrupos de 25 pares cada (similar em 2D e 3D; similar em 2D e dissimilar em 3D; dissimilar em 2D e similar em 3D; dissimilar em ambos), o que sugere — mas não comprova — um desenho aproximadamente balanceado entre as classes. Entretanto, ao se observar o histograma da figura 11, fica nítido uma distribuição mais uniforme para os valores intermediários e valores desbalanceados nos extremos. Já para o *dataset* de [Pereira Franco 2015], é provável que o desbalanceamento entre as classes seja maior em comparação com [Gandini et al. 2022], analisando o histograma da figura 10 observamos um pico grande para os valores próximo a 0, um segundo pico para os valores próximos a 1 e poucas amostras nos valores intermediários. Desta forma, podemos afirmar que os modelos foram treinados com exemplos extremos e avaliados com exemplos intermediários, possivelmente tendo uma maior dificuldade de generalizar exemplos ambíguos presentes nas faixas de valores intermediários.

Além disso, tanto [Pereira Franco 2015] quanto [Gandini et al. 2022] não reportam explicitamente o número final de pares rotulados como similares versus dissimilares após a avaliação humana (nem a distribuição por partição de treino/teste, quando aplicável), o que é um problema metodológico relevante do estudo original: sem essa informação, torna-se difícil (i) diagnosticar desbalanceamento de classes com precisão, (ii) interpretar corretamente métricas sensíveis à prevalência (como acurácia e *F1-score*), (iii) comparar de modo auditável resultados entre trabalhos e (iv) reproduzir análises de erro e calibração. Na prática, essa lacuna de reporte torna ambígua a leitura do desempenho observado aqui, pois parte do comportamento dos modelos pode ser explicado tanto por vieses de prevalência quanto por limitações do classificador em si.

Caso o conjunto de teste utilizado neste trabalho seja, de fato, desbalanceado, isso explicaria diretamente o comportamento observado: um modelo que aprende a prever

majoritariamente a classe mais frequente, a partir de valores extremos, obtém acurácia artificialmente alta (como os 0,7100 do Modelo B e os 0,8200 do Modelo C), mas falha em capturar as classes nas faixas de valores intermediárias — exatamente o padrão presente nas tabelas 3 e 4. Alternativamente, mesmo em um cenário balanceado, a definição do *threshold* de decisão (fixo, no caso dos Modelos A e B; automático, no caso do Modelo C) pode ter sido inadequada para o comportamento da distribuição de saída do modelo, deslocando a fronteira de decisão a favor da classe 1.

Invariavelmente, esse resultado reforça uma crítica já levantada na Seção 4 deste trabalho aos estudos de [Franco et al. 2014] e [Franco et al. 2017]: a definição manual ou pouco criteriosa do limiar de classificação compromete a interpretação do desempenho real do modelo, sendo recomendável, em trabalhos futuros, a aplicação sistemática de métricas como o ponto de corte ótimo da curva ROC (estatística J de Youden) ou técnicas de balanceamento de classes durante o treinamento, como ponderação de classes (*class weighting*) ou sobreamostragem (*oversampling*) da classe minoritária.

6.3. Síntese Comparativa

De forma geral, o Modelo A apresentou o comportamento mais equilibrado entre as três arquiteturas testadas, ainda que com desempenho discriminativo modesto (curva ROC de 0,6356). Os Modelos B e C, por sua vez, embora apresentem acurácia nominal mais alta, sofrem de um desbalanceamento severo na identificação da classe dissimilar, o que compromete sua utilidade prática para a tarefa de avaliação de similaridade molecular — especialmente em um contexto de aplicação real, como a avaliação de medicamentos órfãos, no qual identificar corretamente pares dissimilares é tão ou mais importante do que identificar pares similares. Em comparação com o *baseline* de [Gandini et al. 2022], nenhum dos três modelos propostos atingiu desempenho competitivo, o que indica que a estratégia de aplicar *Deep Learning* diretamente sobre representações 2D e 3D brutas, ao menos na configuração e no volume de dados utilizados neste estudo, não foi suficiente para superar os métodos tradicionais baseados em *fingerprints*.

7. Conclusões

Este trabalho teve como objetivo responder à pergunta de pesquisa: qual é o desempenho da utilização de técnicas de *deep learning* aplicadas à classificação de similaridade entre moléculas? Para isso, foram construídas três modelos de redes neurais convolucionais, inspiradas na arquitetura LeNet-5 e adaptadas para o processamento simultâneo de representações estruturais 2D e 3D em pares de moléculas, treinadas e avaliadas sobre os conjuntos de dados consolidados por [Pereira Franco 2015] e [Gandini et al. 2022].

Os resultados obtidos indicam que, na configuração experimental adotada, os modelos de *Deep Learning* propostos não superaram o desempenho do *baseline* de regressão logística sobre *fingerprints* reportado por [Gandini et al. 2022]: enquanto os modelos da literatura atingiram ROCAUC entre 0,845 e 0,988, os Modelos A, B e C deste estudo obtiveram ROCAUC entre 0,5627 e 0,6356, valores próximos ao de uma classificação aleatória. Adicionalmente, os Modelos B e C apresentaram forte desbalanceamento na identificação da classe de pares dissimilares, sugerindo problemas relacionados à distribuição das classes no conjunto de dados e à calibração do *threshold* de decisão, e não necessariamente uma limitação intrínseca da abordagem de *Deep Learning* em si.

Esses resultados não invalidam a hipótese de que arquiteturas profundas possam, em princípio, capturar padrões estruturais mais ricos do que os *fingerprints* tradicionais; eles indicam, sim, que a viabilização dessa abordagem depende de condições experimentais que não foram plenamente alcançadas neste estudo, em especial o volume de dados de treinamento disponível. Os conjuntos de dados utilizados, com 100 pares moleculares cada, são adequados para o treinamento de modelos lineares com poucos parâmetros, mas insuficientes para o treinamento estável de redes convolucionais, mesmo em arquiteturas relativamente simples como as aqui utilizadas. Esse é, provavelmente, o principal fator explicativo da diferença de desempenho observada, e deve ser considerado a limitação central deste trabalho.

Outras limitações importantes merecem destaque. A avaliação dos Modelos A e B restringiu-se à estratégia de *threshold* fixo, e a do Modelo C à estratégia de *threshold* automático, o que impede uma comparação direta entre as três arquiteturas sob a mesma condição experimental; trabalhos futuros devem padronizar essa avaliação para todos os modelos construídos. Além disso, e de forma mais crítica, não foi possível confirmar com exatidão a proporção de pares similares e dissimilares nos conjuntos de teste utilizados, pois a literatura de base [Pereira Franco 2015, Gandini et al. 2022] não reporta explicitamente a contagem final de rótulos após a avaliação humana (nem, quando aplicável, sua distribuição por partição). Esse é um problema do estudo original que afeta diretamente a validade e a reprodutibilidade de comparações posteriores: sem conhecer a prevalência real, métricas como acurácia e *F1-score* podem ser infladas por vieses de classe, e análises de calibração/erro ficam comprometidas. Assim, para pesquisas futuras aconselhamos (i) reconstruir e publicar essa contagem a partir dos dados brutos e (ii) descrever de forma auditável o procedimento de rotulagem e as distribuições por divisão, para permitir diagnósticos e comparações mais sólidos sobre desbalanceamento de classes.

Como trabalhos futuros, sugere-se: (i) a ampliação do conjunto de dados de treinamento, seja por meio da coleta de novos pares com avaliação humana, seja por meio de técnicas mais agressivas de *data augmentation* específicas para representações moleculares 2D e 3D; (ii) a aplicação de técnicas de balanceamento de classes, como ponderação de classes durante o treinamento ou sobreamostragem da classe minoritária; (iii) a definição do *threshold* de decisão a partir de métricas como o ponto de corte ótimo da curva ROC, de forma consistente entre todos os modelos avaliados; (iv) a exploração de arquiteturas de transferência de aprendizado (*transfer learning*), utilizando redes pré-treinadas em tarefas relacionadas de visão computacional ou quimioinformática, de modo a mitigar a escassez de dados rotulados; (v) testes utilizando-se de Redes Neurais Orientadas a Grafos (Graph Neural Networks – GNNs); (vi) realização de um refinamento mais aprofundado sobre os hiperparâmetros; (vii) explorar o comportamento de redes neurais artificiais com *fingerprints*; e (viii) a realização de uma análise qualitativa dos pares de moléculas classificados incorretamente, nos moldes da análise realizada por [Franco et al. 2017] para seu próprio conjunto de dados, buscando identificar padrões relacionados à complexidade estrutural ou ao peso molecular que possam explicar os erros de classificação observados.

O presente trabalho contribui para a área de avaliação de similaridade molecular ao reportar os resultados e suas limitações, obtidos pela aplicação de *Deep Learning* diretamente sobre representações estruturais 2D e 3D de moléculas; além de posicionar

esses resultados em comparação direta com o estado da arte baseado em *fingerprints* e regressão logística. Embora os modelos aqui propostos não tenham superado o *baseline* da literatura, o estudo evidencia caminhos concretos — sobretudo relacionados ao volume de dados e à calibração de *threshold* — que podem orientar pesquisas futuras na construção de modelos mais robustos para essa tarefa.

References

- Brown, F. K. et al. (1998). Chemoinformatics: what is it and how does it impact drug discovery. *Annual reports in medicinal chemistry*, 33:375–384.
- Chapman, P. (2000). Crisp-dm 1.0: Step-by-step data mining guide.
- Ebejer, J.-P., Morris, G. M., and Deane, C. M. (2012). Freely available conformer generation methods: how good are they? *Journal of chemical information and modeling*, 52(5):1146–1158.
- Ehrlich, H.-C. and Rarey, M. (2011). Maximum common subgraph isomorphism algorithms and their applications in molecular science: a review. *WIREs Computational Molecular Science*, 1(1):68–79.
- Franco, P., Porta, N., Holliday, J. D., and Willett, P. (2014). The use of 2d fingerprint methods to support the assessment of structural similarity in orphan drug legislation. *Journal of Cheminformatics*, 6:5 – 5.
- Franco, P., Porta, N., Holliday, J. D., and Willett, P. (2017). Molecular similarity considerations in the licensing of orphan drugs. *Drug discovery today*, 22 2:377–381.
- Gandini, E., Marcou, G., Bonachéra, F., Varnek, A., Pieraccini, S., and Sironi, M. (2022). Molecular similarity perception based on machine-learning models. *International Journal of Molecular Sciences*, 23.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.
- Hijazi, S., Kumar, R., Rowen, C., et al. (2015). Using convolutional neural networks for image recognition. *Cadence Design Systems Inc.: San Jose, CA, USA*, 9(1).
- James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An introduction to statistical learning: With applications in python*. Springer Nature.
- Kelleher, J. D., Namee, B. M., and D’Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples and Case Studies*. MIT Press, 2 edition.
- Kutchukian, P. S., Vasilyeva, N. Y., Xu, J., Lindvall, M. K., Dillon, M. P., Glick, M., Coley, J. D., and Brooijmans, N. (2012). Inside the mind of a medicinal chemist: The role of human bias in compound prioritization during drug discovery. *PLOS ONE*, 7(11):1–18.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.
- Lewis, G. N. (1916). The atom and the molecule. *Journal of the American Chemical Society*, 38(4):762–785.
- Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

- Nasser, M., Salim, N., Saeed, F., Basurra, S., Rabi, I., Hamza, H., and Alsoufi, M. A. (2022). Feature reduction for molecular similarity searching based on autoencoder deep learning. *Biomolecules*, 12(4).
- Niazi, S. K. and Mariam, Z. (2023). Recent advances in machine-learning-based chemoinformatics: A comprehensive review. *International Journal of Molecular Sciences*, 24(14).
- Pereira Franco, J. P. (2015). *Similarity in the context of the Orphan Drug Legislation*. PhD thesis, University of Sheffield.
- Raymond, J. W. and Willett, P. (2002). Maximum common subgraph isomorphism algorithms for the matching of chemical structures. *Journal of Computer-Aided Molecular Design*, 16:521–533.
- Śmieja, M. and Warszycki, D. (2016). Average information content maximization—a new approach for fingerprint hybridization and reduction. *PloS one*, 11(1):e0146666.
- Zarnecka, J. M. (2018). *Assessing and developing methods to explore the role of molecular shape in computer-aided drug design*. Liverpool John Moores University (United Kingdom).