UNIVERSIDADE FEDERAL DE SÃO CARLOS
DEPARTAMENTO DE COMPUTAÇÃO
ENGENHARIA DE COMPUTAÇÃO

Felipe do Nascimento Torrieri

Análise comparativa de desempenho da geração de conjuntos de moléculas utilizando redes generativas adversárias

São Carlos - SP

2025

Felipe do Nascimento Torrieri

Análise comparativa de desempenho da geração de conjuntos de moléculas utilizando redes generativas adversárias

Trabalho de Conclusão de Curso apresentado ao curso de Engenharia de Computação da Universidade Federal de São Carlos, como requisito para a obtenção do título de Bacharel em Engenharia de Computação.

Orientação Prof. Dr. Alan Demétrius Baria Valejo

São Carlos - SP

2025

Dedico este trabalho a minha família e a todos os meus amigos.

Agradecimentos

Agradeço à minha mãe, Erika, ao meu pai, Enio, e ao meu irmão, Fábio, pelo apoio que tive deles em todas as etapas da minha vida até agora e por estarem comigo em todos os bons momentos e também os momentos mais difíceis da minha vida. E também a Luna, minha cachorrinha que chegou faz pouco tempo e virou uma parte essencial da família.

Agradeço a todas as pessoas que conheci graças à graduação e que se tornaram pessoas muito importantes na minha vida e que fizeram com que a faculdade fosse a melhor época da minha vida. Em especial, gostaria de citar aqui o Caique, o Vinícius, a Jhulie, a Amanda, o Allan, a Julia, o João V., a Emilia, o João H., a Sara, o João P., o Tulio, o Pedro e todos os outros amigos que estiveram comigo em momentos especiais durante esse período.

Ao meu orientador, Prof. Dr. Alan Valejo, por sua paciência e por sempre ser solícito ao me auxiliar em questões na orientação deste trabalho. E aos outros professores que, ao longo de toda graduação, foram essenciais na minha formação didática.

Por último, mas não menos importante, agradeço a todas as pessoas não citadas aqui mas que também geraram algum impacto positivo na minha vida.

Resumo

A geração de novas moléculas e compostos moleculares é uma tarefa que possui variados números de aplicações em diversas áreas da ciência, sendo relevante atualmente para o desenvolvimento de novas drogas, remédios e compostos químicos que podem ter benefícios para a vida humana ou para diversas utilidades químicas. Para isso, é fundamental que os métodos computacionais sejam evoluídos de forma a otimizar essa geração, garantindo a aprendizagem dos padrões de compostos químicos e moleculares para gerar moléculas válidas de um ponto de vista físico-químico. Dentre os métodos de geração de moléculas, destacam-se as redes generativas adversárias (GANs), que são algoritmos recentes capazes de aprender padrões de complexidade em grandes volumes de dados e proporcionam a geração de novos exemplos a partir do conjunto de dados original. Este trabalho busca avaliar o desempenho das GANs no contexto de geração de moléculas relacionando as métricas de avaliação necessárias com outros métodos de IA generativa e também com diferentes conjuntos de dados relacionados a fármacos e compostos moleculares, contribuindo, assim, na identificação dos pontos fortes e fracos do uso das GANs comparado ao uso de outros métodos de geração de moléculas, como os VAEs e as GNNs, em diferentes contextos e conjuntos de dados.

Palavras-chave: Geração de Moléculas; Redes Neurais; GANs; IA; Aprendizado de Máquina.

Abstract

The generation of new molecules and molecular compounds is a task with numerous applications in various scientific fields, currently playing a crucial role in the development of new drugs, medicines, and chemical compounds that may benefit human life or serve various chemical purposes. To achieve this, it is essential to advance computational methods to optimize this generation process, ensuring that machine learning algorithms effectively learn the patterns of chemical and molecular compounds to generate valid molecules from a physicochemical perspective. Among the molecular generation methods, generative adversarial networks (GANs) stand out as recent algorithms capable of learning complex patterns from large datasets and generating new examples based on the original data. This study aims to evaluate the performance of GANs in the context of molecular generation by comparing the necessary evaluation metrics with other generative AI methods and different datasets related to pharmaceuticals and molecular compounds. This analysis contributes to identifying the strengths and weaknesses of GANs compared to other molecular generation methods, such as VAEs and GNNs, across various contexts and datasets.

Keywords: Molecule Generation; Neural Networks; GANs; AI; Machine Learning.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A geração de novas moléculas e compostos moleculares é uma tarefa que possui variados números de aplicações em diversas áreas da ciência, tais quais a química, a medicina e a biotecnologia. Com o objetivo de criar conjuntos moleculares com características específicas para determinados fins, tem surgido a necessidade do desenvolvimento de métodos computacionais com desempenhos melhores com relação aos já existentes. Entre estes métodos, pode-se destacar o uso de redes adversárias generativas (GANs - Generative Adversarial Networks), uma técnica recentemente criada na área de aprendizado de máquina, que é capaz de aprender padrões de complexidade em volumes grandes de dados e, a partir, deles, proporcionar a criação e geração de novos exemplos daquele conjunto de dados (ZHANG et al., 2022).

Gerar moléculas significa criar novas estruturas moleculares a partir de dados ou modelos computacionais. Isso é feito com algoritmos de inteligência artificial e aprendizado de máquina, principalmente para encontrar compostos com propriedades desejáveis, como eficácia farmacológica, baixo custo de produção, ou características específicas para aplicações em áreas como a medicina, a química e a biotecnologia (XU et al., 2019).

As GANs foram propostas inicialmente por Ian Goodfellow no ano de 2014 e consistiam em duas redes neurais: uma geradora e uma discriminadora. A rede geradora possui a função de tentar criar dados sintéticos que sejam semelhantes aos dados reais, e a rede discriminadora possui a função de tentar diferenciar os dados gerados dos dados já existentes nas bases de dados reais. Esse processo faz com que a rede geradora produza dados que vão se aproximando cada vez mais dos esperados conforme a rede discriminadora se torna mais precisa. Por este processo, as GANs e sua capacidade de geração de novos conjuntos de dados tornaram-se relevantes na descoberta de novos compostos químicos (GOODFELLOW et al., 2014).

Na química computacional, as GANs têm sido aplicadas para gerar conjuntos de moléculas com propriedades desejáveis, como por exemplo determinada característica de conformação estrutural ou baixa toxicidade. Levando em consideração o trabalho de Kadurin et al. (2017), as GANs foram usadas para gerar novos compostos que poderiam ser candidatos a se tornarem fármacos com características desejadas otimizadas. Outros métodos de aprendizado, como redes neurais recorrentes, também vêm sendo explorados na geração de moléculas, motivando este estudo para avaliação do desempenho das GANs em relação a outras abordagens (KADURIN et al., 2017).

Este trabalho tem como objetivo realizar um estudo comparativo entre diferentes métodos de geração de moléculas, com ênfase nas GANs. Para isso, serão considerados

aspectos como a capacidade do modelo gerar moléculas válidas no âmbito químico, a capacidade do modelo de gerar moléculas diferentes e a capacidade do modelo de gerar moléculas novas (diferentes das que já estão na base de dados), baseadas nas métricas estatísticas gerais de validação definidas pela literatura (RAZAVI-FAR et al., 2022).

Além desse trabalho, outros estudos como o de Blaschke et al. (2018) destacam a importância de realizar comparações de desempenho, uma vez que diferentes modelos e métodos podem apresentar pontos fortes e fracos em diferentes métricas, se comparados uns com os outros. Portanto, esse trabalho irá buscar entender as capacidades de geração molecular de diferentes modelos e, além disso, fornecer insights sobre suas limitações e caminhos possíveis para que possam haver futuras melhorias (BLASCHKE et al., 2018).

1.1 Objetivo

O objetivo deste trabalho é estudar como a geração de moléculas é realizada por meio de uma GAN, analisar e comparar fatores de performance das suas execuções e fazer uma análise do desempenho da mesma, relacionando com a execução de outros algoritmos de IA generativa no mesmo propósito.

1.2 Objetivo específico

Dados os métodos distintos e diferentes conjuntos de dados que serão usados nesta análise comparativa, um objetivo específico importante é a definição de quais algoritmos podem ser mais recomendados ou possuírem melhor desempenho no contexto de geração de moléculas com base nas métricas de avaliação e definir se as GANs são as mais recomendadas para determinado contexto ou se os outros algoritmos usados possuem desempenho melhor nas tarefas detalhadas neste trabalho.

2 Fundamentação Teórica

Este capítulo contém os conceitos que são abordados neste trabalho. Além disso, serve como um guia para entender os assuntos tratados.

2.1 Aprendizado de máquina

O aprendizado de máquina é um campo da inteligência artificial em que há o desenvolvimento e aplicação de algoritmos que permitem que computadores aprendam padrões de comportamento a partir de dados já existentes. Como esses sistemas podem identificar padrões complexos e ajudar no processo de tomada de decisões em diversos segmentos, esta área se tornou cada vez mais relevante nos dias atuais, podendo ser relevante na área da saúde, educação, marketing e modelos financeiros (JORDAN; MITCHELL, 2015).

O aprendizado de máquina pode ser dividido em algumas categorias, 3 delas são: o aprendizado supervisionado, em que o algoritmo prevê resultados a partir de dados que possuem rótulos; o aprendizado não supervisionado, em que o algoritmo busca padrões em dados que não são rotulados; e o aprendizado por esforço, em que o modelo aprende por meio de interações ambientais (MORALES; ESCALANTE, 2022). De especial interesse para este trabalho é o aprendizado não supervisionado, detalhado na próxima subseção.

2.1.1 Aprendizado não supervisionado

O aprendizado não supervisionado é uma das categorias do aprendizado de máquina em que dados sem rótulos são analisados. Com isso, os algoritmos tentam buscar padrões de comportamento ou estrutura nos dados, explorando-os para identificar grupos ou encontrar representações relevantes (NAEEM et al., 2023).

Um método bem conhecido de aprendizado não supervisionado é o agrupamento, em que os dados são divididos em subgrupos com base em características similares que os elementos desse grupo possuam entre si em relação a outros grupos, e entre os algoritmos que podem ser usados nesse método, destaca-se o K-Means (BOCK, 2007).

A geração de moléculas utilizando algoritmos de IA generativa está inserida no contexto de aprendizado de máquina não supervisionado, uma vez que os dados não possuem rótulos e os métodos apresentados tentam buscar padrões de comportamento e de estruturas moleculares nas entradas, aprendendo sobre sua complexidade para que consigam gerar moléculas válidas do ponto de vista químico.

2.2 Redes Neurais Artificiais

As redes neurais artificiais são modelos computacionais que se espelham na estrutura de funcionamento do cérebro humano, e são organizadas por partes que recebem os dados, partes que processam os dados, e partes que geram previsões com base nesses dados. São usadas atualmente em técnicas de *deep learning*, como no reconhecimento de imagem, processamento de linguagem natural e tradução automática (WU; FENG, 2018).

O aprendizado de uma rede neural artificial é feito pelo ajuste do peso das conexões dessa rede para que seja minimizada a diferença entre a saída real e a saída desejada do modelo usado. Esse processo é chamado de retropropagação e é uma técnica criada por Rumelhart, Hinton e Williams (1986), que consiste em calcular os gradientes de erro e, assim, ajustar os pesos das conexões pelo algoritmo (RUMELHART; HINTON; WILLIAMS, 1986). Entre os diversos tipos de redes propostos, as GANs foram escolhidas para serem investigadas neste trabalho e são detalhadas a seguir.

2.2.1 Redes Adversárias Generativas (GANs)

As Redes Adversárias Generativas, ou GANs, foram propostas por Ian Goodfellow em 2014 e são consideradas uma das maiores inovações atuais na área de *deep learning*. O novo modelo de aprendizado não supervisionado foi baseado em uma estrutura que possui dois componentes: um gerador e um discriminador, em que esses dois componentes se comunicam um com o outro de forma a se obter uma saída ou resultado esperado para o algoritmo (GOODFELLOW et al., 2014).

Essa arquitetura se destacou por sua capacidade de gerar dados cada vez mais próximos do esperado, e tornou-se uma ferramenta muito relevante em áreas como geração de imagens, bioinformática e na descoberta de fármacos.

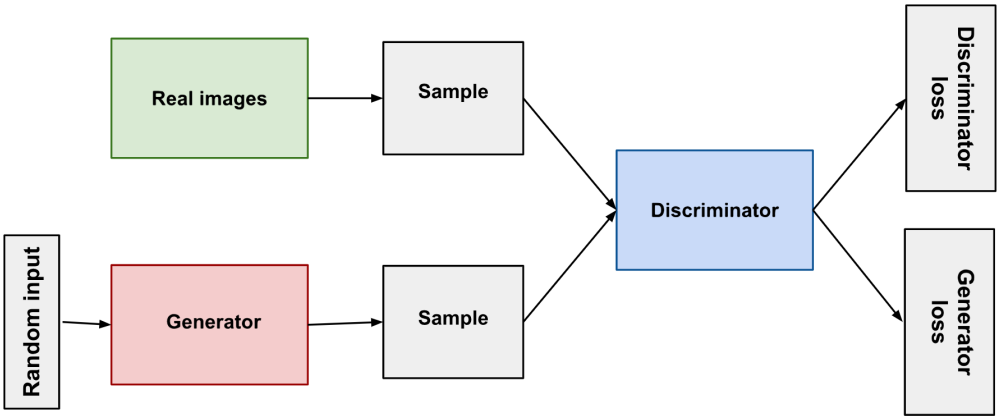
2.2.2 Arquitetura das GANs

A arquitetura das GANs é composta por dois componentes principais:

- Gerador: O componente gerador da rede tem a função de criar dados que sejam semelhantes aos dados reais. Pode começar com uma entrada aleatória e, conforme atravessa as camadas de *deep learning*, pode gerar amostras cada vez mais fiéis à semelhança da base de dados.
- Discriminador (Discriminator): O componente discriminador da rede tem a função de diferenciar os dados gerados pelo gerador e os dados reais do conjunto de dados de treinamento, agindo como um classificador que aprende a identificar se um conjunto de dados é “real” ou “falso”.

O aprendizado de máquina adversário entre esses dois componentes é capaz de criar um ciclo de melhoria contínua, pois enquanto o gerador tenta criar dados cada vez mais semelhantes à base de dados para que o discriminador não seja capaz de distingui-lo, o discriminador aumenta cada vez mais sua capacidade de diferenciar o que é falso e o que é real(AGGARWAL; MITTAL; BATTINENI, 2021).

Figura 1 – Funcionamento de uma GAN



Fonte: (Google Developers, 2022)

A Figura 1 ilustra o processo de aprendizado de uma GAN, em que o discriminador é alimentado com dados reais e com dados criados pelo gerador, influenciando a competição entre os dois componentes.

O conceito de aprendizado de máquina adversário é inspirado na teoria dos jogos (MOGHADAM et al., 2022), em que os dois componentes se enfrentam em um jogo em que, caso um tenha sucesso, o outro falha. Enquanto o gerador tenta minimizar a função de custo, o discriminador tenta maximizá-la. Isso é representado pela função matemática 2.1:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))], \qquad (2.1)$$

onde G(z) é o gerador que transforma uma entrada em uma amostra gerada e D(x) é o discriminador que tenta prever se a amostra vem da distribuição real ou do gerador.

A função matemática 2.1 evidencia que, enquanto o treinamento progride, o discriminador fica mais habilidoso na identificação de entradas geradas pelo gerador, e o gerador melhora suas amostras para que o discriminador não consiga identificá-las, logo o objetivo é alcançar um equilíbrio em que o discriminador não tenha mais capacidade de distinguir entre amostras geradas e amostras reais.

2.3 Algoritmos usados nos experimentos

2.3.1 VAEs - Autoencoders Variacionais

Os Autoencoders Variacionais (VAEs) são modelos de *deep learning* generativos que possuem capacidade de aprendizado de representações de dados e são capazes de gerar novos dados semelhantes aos dados de treinamento (KINGMA; WELLING, 2019).

Os VAEs são compostos por duas redes neurais: um encoder e um decoder.

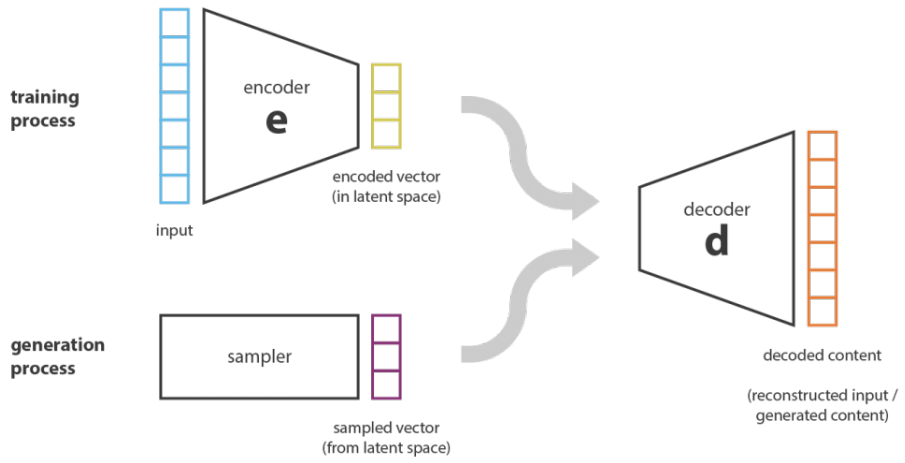
- Encoder: é responsável por mapear os dados de entrada para um espaço latente menor, que contém uma representação compacta dos dados originais.
- Decoder: é responsável por reconstruir os dados de entrada a partir da representação do espaço latente, impondo uma distribuição de probabilidade sobre esse espaço, o que permite que o modelo gere novos dados de forma realista.

O funcionamento desse algoritmo é dado por suas fases: codificação, amostragem e decodificação (CINELLI et al., 2021).

Na codificação, o encoder recebe dados de entrada e mapeia esses dados para um espaço latente. Ele produz os parâmetros de uma distribuição de probabilidade, que é geralmente gaussiana, em vez de distribuir um único ponto.

Na amostragem, uma amostra é retirada da distribuição de probabilidade no espaço latente e, na decodificação, o decoder usa a amostra para reconstruir o dado original.

Figura 2 – Funcionamento de um VAE



Fonte: (SOUSA, 2022)

Em sua função de perda, a perda é medida pela diferença entre a entrada original e a reconstrução produzida pelo decoder, garantindo que a distribuição do espaço latente seja próxima de uma distribuição normal.

A Figura 2 ilustra o processo de aprendizagem de um VAE, em que os dados são mapeados pelo encoder e repassados para o decoder para que ele consiga reconstruir as entradas baseado no que recebeu, gerando novas saídas.

Por sua versatilidade, os VAEs podem ser utilizados na geração de dados (gerando novas imagens, músicas, textos e outros tipos de dados), redução de dimensionalidade (projeção de dados de alta dimensionalidade), denoising (remoção de ruídos nos dados) e até mesmo em aplicações financeiras e de biosinais (SINGH; OGUNFUNMI, 2022).

2.3.2 GNNs - Redes Neurais de Grafos

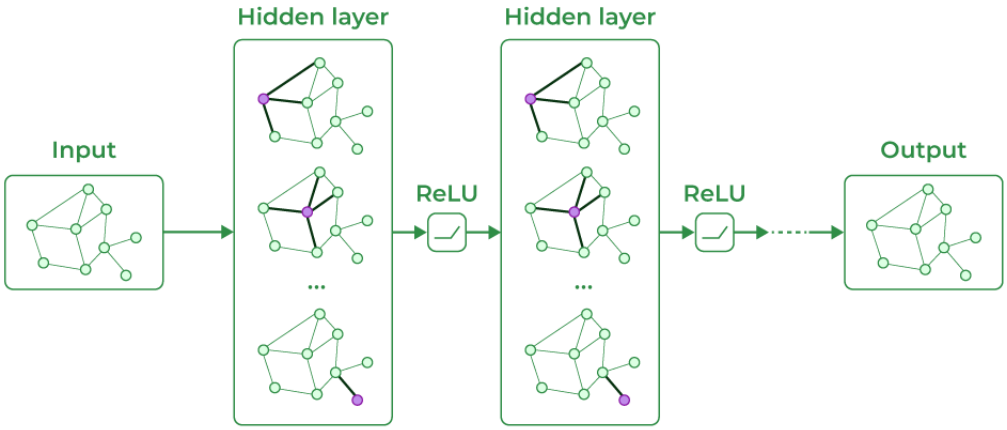
As Redes Neurais de Grafos (GNNs) são um modelo de aprendizado de máquina projetado para operar em dados estruturados em forma de grafos. Elas são capazes de capturar relações complexas e interdependências entre nós e arestas de um grafo (WU et al., 2021).

Essas redes neurais utilizam de sua estrutura para aprender representações vetoriais dos nós e arestas dos gráficos, aprendendo assim sobre as características locais do mesmo.

Seu funcionamento é dado pelos seguintes passos (KARAGIANNAKOS, 2020):

- Inicialização: Na inicialização, cada nó do grafo é inicializado com um vetor de características sobre os dados.
- Propagação: As informações são propagadas entre os nós vizinhos em múltiplas camadas e, com isso, as representações dos nós vão sendo atualizadas.
- Agregação: As informações advindas dos nós vizinhos são agregadas para atualizar a representação de cada nó.
- Atualização: A representação dos nós é atualizada com base nas informações que foram agregadas anteriormente.

Figura 3 – Funcionamento de uma GNN



Fonte: (GEEKSFORGEEKS, 2024)

A Figura 3 ilustra o processo de aprendizagem de uma GNN, em que as entradas representadas por grafos recebem informações de seus nós vizinhos, gerando agregações sucessivas até o ponto em que todos os nós possuam informações sobre os outros nós, fazendo com que o modelo seja capaz de criar mais saídas baseado nos padrões entre arestas e nós aprendidos por ele.

Pelo seu funcionamento e aprendizado, as GNNs podem ser aplicadas em diversas áreas, como: descoberta de drogas (previsão de propriedades moleculares e geração de novos compostos), recomendação de itens (detecção de preferência de usuário), detecção de fraudes (identificação de transações fraudulentas) e processamento de linguagem natural (modelagem da estrutura semântica e sintática dos textos), entre outros (WAIKHOM; PATGIRI, 2021).

3 Revisão da Literatura

Para o avanço deste trabalho, uma etapa fundamental foi a revisão da literatura sobre o tema abordado. Essa seção contém os detalhes e resultados encontrados com a etapa de levantamento bibliográfico e leitura das referências.

3.1 Mapeamento Bibliográfico

As GANs podem ser usadas na geração de novos compostos químicos pela sua habilidade de geração de novos dados, os principais trabalhos que exemplificam isso se encontram a seguir:

- Goodfellow et al. (2014), "Generative Adversarial Networks": Teoria inicial sobre GANs, que inspirou sua aplicação em vários domínios, incluindo a geração de moléculas.
- Cao e Kipf (2022), "MolGAN: An implicit generative model for small molecular graphs": Artigo que adapta GANs para geração de moléculas representadas como grafos, evitando alguns problemas que podem surgir na utilização de SMILES. Neste trabalho, é utilizado o conjunto de dados QM9 (RAMAKRISHNAN et al., 2014), que é um conjunto de dados composto por aproximadamente 134 mil compostos orgânicos contendo os átomos de carbono (C), oxigênio (O), nitrogênio (N) e flúor (F), além de detalhes sobre suas propriedades quânticas.

As métricas usadas nesse trabalho, que são validade, unicidade e novidade são amplamente discutidas para avaliar a qualidade da geração de moléculas e têm foco em assegurar a relevância teórica e prática desses métodos. Os trabalhos que discutem sobre essas métricas são citados a seguir:

- Polykovskiy et al. (2018), "Molecular Sets (MOSES): A benchmarking platform for molecular generation models": Discussão sobre métricas como validade, unicidade e novidade.
- Razavi-Far et al. (2022), "Generative adversarial learning: Architectures and applications. Intelligent Systems Reference Library, Volume 217. Springer": Livro que explora os fundamentos teóricos e as aplicações práticas do aprendizado das GANs, abordando suas arquiteturas e diversas implementações.

3.2 Principal Referência

As GANs são um marco do aprendizado de máquina, e foram introduzidas por Goodfellow et al. (2014) no artigo "Generative Adversarial Nets". Nesse modelo, dois componentes principais são usados, um gerador e um discriminador, competindo entre si em um processo de aprendizado que leva à criação de dados que imitam padrões complexos baseado em conjuntos de dados reais (GOODFELLOW et al., 2014).

Alguns artigos recentes ampliaram o escopo das GANs, sendo possível a representação de moléculas baseadas em grafos para lidar com complexidade estrutural. Com isso, pode-se citar alguns trabalhos:

- Arús-Pous et al. (2020). "Randomized SMILES strings improve the quality of molecular generative models. Journal of Cheminformatics": Neste artigo mais atual, é explorado o uso de representações SMILES otimizadas para melhorar o desempenho de determinados modelos generativos, incluindo GANs. Neste estudo, há um foco na geração de moléculas analisando as representações SMILES e, utilizando-se do conjunto de dados CheMBL para o treinamento do modelo, a conclusão obtida foi que os modelos usados foram capazes de gerar uma grande quantidade de saídas no domínio de moléculas e que com apenas 0.001% do conjunto de dados usado no treinamento, foi gerado mais da metade das moléculas do conjunto real de dados.

- Maziarka et al. (2020). "Mol-CycleGAN: A generative model for molecular optimization. Journal of Cheminformatics": Neste trabalho, as GANs são combinadas com algoritmos de aprendizado por reforço para otimizações das propriedades moleculares geradas. Utilizando-se dos conjuntos de dados ZINC e CheMBL, este trabalho conclui que o novo modelo criado pelos autores é capaz de gerar moléculas com propriedades desejáveis semelhantes às moléculas de teste e que os resultados obtidos têm performances superiores ao resultado obtido pelas GANs em sua arquitetura original.

- Gangwal et al. (2024). "Recent Advances of Generative Models for Drug Design. Frontiers in Pharmacology, 15, 1331062": Este trabalho é importante pois aborda o uso de GANs na geração de moléculas, analisando avanços em inteligência artificial generativa aplicada na descoberta de novos fármacos, discutindo como GANs, juntamente com outras abordagens, como RNNs e VAEs têm sido usadas para essas gerações e explora também novos frameworks e desafios futuros na área. Utilizando-se de diversos conjuntos de dados, como o ZINC, este trabalho conclui que todos algoritmos podem ser eficientes à sua maneira, mas que ainda há um grande potencial para a criação de novos métodos de forma que a inteligência artificial generativa seja aproveitada ao seu máximo potencial e que ainda existem limitações, como da-

dos de treinamento contendo ruído e o grande tempo e custo para se obter acesso às bases de dados.

Nota-se que, desde sua idealização teórica em 2014, as GANs passaram por diversas evoluções em relação às suas aplicações, sendo citada em diversos artigos e estudos como um método viável e útil na geração de moléculas (ZHANG et al., 2022; RENDO, 2022; BLANCHARD; STANLEY; BHOWMIK, 2021).

Os trabalhos citados nesta seção mostram que a geração de moléculas é um tema atual e recorrente nas áreas de química avançada e descoberta de novos fármacos e relacionam-se com esse trabalho pois utilizam algoritmos e conjuntos de dados semelhantes aos que foram usados nele para a mesma finalidade.

Com isso, há uma base sólida para a discussão do desempenho da aplicação de GANs na geração de moléculas, abrangendo comparações com outros métodos, adaptações e avanços recentes e motivando a criação deste trabalho de análise comparativa.

4 Metodologia

O objetivo desse capítulo é informar sobre a base de dados, algoritmos, métodos e outros aspectos da metodologia utilizados durante o trabalho.

4.1 Ambiente de execução

Para a realização dos experimentos presentes nesse trabalho, a plataforma Google Colab foi utilizada, que é um ambiente de desenvolvimento baseado na nuvem que oferece suporte para Jupyter Notebooks e permite a execução de códigos Python. Após os experimentos nessa plataforma, as implementações foram exportadas para o GitHub, que se encontra disponível no link <https://github.com/midas-ufscar/2024-2-TCC-2-Felipe-do-Nascimento-Torrieri>.

Os experimentos foram conduzidos utilizando o ambiente do Colab com CPU para que todos os métodos fossem analisados seguindo os mesmos parâmetros de configuração. A configuração específica geralmente incluiu:

- Sistema operacional: Linux baseado no Ubuntu (fornecido pelo Google Colab)
- Processador: Intel Xeon (núcleos e threads variáveis)
- Memória RAM: 8GB
- Versão do Python: 3.11.11

As principais bibliotecas utilizadas foram:

- RDKit-pypi version: 2022.09.5
- pandas version: 2.2.2
- torch version: 2.5.1+cu121
- torchvision version: 0.20.1+cu121
- torch-geometric version: 2.6.1
- matplotlib version: 3.10.0

As principais utilidades dessas bibliotecas foram: a implementação de algoritmos de IA generativa (torch, torch-geometric), a manipulação de moléculas em formato SMILES (rdkit), manipulação de dados (pandas), visualizações de métricas (matplotlib).

Com essa infraestrutura, o ambiente estava adequado para a execução dos modelos propostos e permitiu que o custo computacional e a disponibilidade de recursos estivessem balanceados.

4.2 Base de dados

Para a comparação da análise de desempenho dos algoritmos, é crucial que a escolha das bases de dados utilizadas nos testes seja feita considerando a aplicação viável nos métodos que serão utilizados na comparação.

Com isso, foram escolhidas as bases ZINC (IRWIN; SHOICHET, 2005; TINGLE et al., 2023), ChEMBL (GAULTON et al., 2012; ZDRAZIL et al., 2024) e PubChem (KIM et al., 2016; KIM; BOLTON, 2024), que são conhecidas na área do aprendizado de máquina por conta da sua diversidade e grande detalhamento de informações químicas dos compostos.

- ZINC: é uma base pública que contém acima de 37 bilhões de moléculas comercialmente disponíveis em formatos como MOL2, SDF, SMILES e Flexibase, organizadas com base em suas propriedades químicas. Por conta disso, ela permite o treinamento de modelos generativos com a finalidade de criação de novas moléculas com propriedades otimizadas (IRWIN; SHOICHET, 2005; TINGLE et al., 2023).
- CheMBL: é uma base de dados bioativa que contém mais de 1.6 milhões de compostos químicos nos formatos SMILES e SDF e que contém informações com muitos detalhes sobre essas moléculas e suas atividades biológicas e por isso é muito usada em modelagem molecular (GAULTON et al., 2012; ZDRAZIL et al., 2024).
- PubChem: é uma das maiores bases de dados de domínio público, contendo mais de 293 milhões de moléculas nos formatos SDF, SMILES, InChI e XML e que contém informações abrangentes sobre esses compostos químicos, como suas propriedades, atividades biológicas e também dados sobre toxicidade (KIM et al., 2016; KIM; BOLTON, 2024).

4.3 Algoritmos Utilizados

Para realizar a análise comparativa de forma satisfatória, é necessário comparar o desempenho dos algoritmos das GANs com outros algoritmos conhecidos na geração

de moléculas presentes na bibliografia. Com isso, além dos algoritmos das GANs, serão utilizados:

4.3.1 Autoencoders Variacionais (VAEs)

Os Autoencoders Variacionais (VAEs) são modelos baseados em redes neurais que utilizam da probabilidade e estatística para identificarem padrões de representação contínua das moléculas, o que torna possível a geração de novas estruturas.

O estudo de Gómez-Bombarelli et al. (2018) deu introdução ao uso dos VAEs na geração de moléculas com base em propriedades desejáveis, utilizando-se de representações moleculares do formato SMILES e da manipulação do espaço contínuo em que as propriedades moleculares são otimizadas (GóMEZ-BOMBARELLI et al., 2018).

4.3.2 Redes Neurais de Grafos (GNNs)

As Redes Neurais de Grafos (GNNs) são modelos de aprendizado de máquina projetados para operar em dados estruturados na forma de grafos. Por sua grande capacidade de captação de relações e interdependências complexas entre nós e arestas de um grafo, elas são ótimas ferramentas no contexto de inteligência artificial generativa (WU et al., 2021).

Moléculas podem ser representadas por grafos, em que átomos correspondem a nós e ligações químicas podem ser representadas por arestas, permitindo às GNNs reconhecerem padrões de complexidade química e interações entre átomos, facilitando a modelagem de propriedades moleculares (BONGINI; BIANCHINI; SCARSELLI, 2021).

4.4 Medidas de Avaliação

Para a análise comparativa dos métodos computacionais de geração de moléculas, necessitam-se estabelecer critérios para garantir a geração mais efetiva possível e analisar os resultados com base nessas métricas. Para isso, serão utilizadas as seguintes métricas que estão descritas no Capítulo 11 do livro Generative Adversarial Learning: Architectures and Applications (RAZAVI-FAR et al., 2022):

4.4.1 Validade

A validade se refere à proporção de amostras geradas que satisfazem as regras e restrições do domínio molecular, garantindo que as moléculas sejam válidas para o fim que foram geradas, podendo-se observar fatores e regras químicas, como por exemplo a valência dos átomos.

$$\text{Validade} = \frac{\text{Número de amostras válidas geradas}}{\text{Número total de amostras geradas}} \tag{4.1}$$

A validade é calculada pela Equação 4.1, em que se obtém a porcentagem de amostras válidas geradas na amostra em relação ao número total de amostras que foram geradas.

Os requisitos para definir a validade de uma molécula são:

- Sintaxe SMILES correta: A string deve seguir a sintaxe SMILES, então a presença de caracteres inválidos ou símbolos desconhecidos tornam o SMILES inválido.
- Balanceamento de átomos e valências: todo átomo deve seguir sua valência, como o carbono (C) que pode ter só até 4 ligações químicas. Logo, estruturas que violam a valência permitida para qualquer átomo tornam o SMILES inválido.
- Ramificações e fechamento de anéis corretos: Os anéis, representados por números (C1CCCCC1 para ciclo-hexano, por exemplo) devem ser corretamente balanceados e ramificações devem estar unidas por parênteses para evitar estruturas incorretas.
- Uso correto de ligações: Ligações podem ser definidas como simples (-), duplas (=), triplas (#) ou aromáticas (:). Caso a conectividade entre dois átomos esteja incorreta, o SMILES será considerado inválido.
- Carga e balanceamento eletrônico: A representação de íons e neutralidade deve ser corretamente representada, como em [Na+] e [Cl-]. Logo, estruturas que não cumprem com as regras de neutralidade são consideradas inválidas.

A validade é importante, pois se um modelo possui baixa validade, isso significa que o modelo não está aprendendo de forma adequada as restrições que os conjuntos de dados reais possuem e, assim, sua aplicabilidade prática acaba sendo limitada.

4.4.2 Unicidade

A unicidade avalia a diversidade de exemplos gerados, verificando a proporção de exemplos únicos que estão contidos no conjunto gerado.

$$\text{Unicidade} = \frac{\text{Número de amostras únicas geradas}}{\text{Número total de amostras geradas}} \tag{4.2}$$

A unicidade é calculada pela Equação 4.2, em que se obtém a porcentagem de amostras únicas geradas na amostra em relação ao número total de amostras que foram geradas. A unicidade é calculada após a execução de todas as épocas, assim, ela tem a

capacidade de definir se o modelo está gerando sempre os mesmos resultados de saída ou se o modelo está gerando amostras diferentes dentre todas as suas execuções.

A unicidade é importante, pois seu cálculo garante que o modelo não seja redundante de forma que gere apenas resultados iguais e garantindo que a quantidade de dados gerados seja explorada de forma eficaz.

4.4.3 Novidade

A novidade avalia e mede a capacidade do modelo de criar amostras novas, ou seja, amostras que sejam diferentes daquelas amostras que já estão presentes no conjunto de treinamento original.

$$\text{Novidade} = \frac{\text{Número de amostras novas geradas}}{\text{Número total de amostras geradas}} \tag{4.3}$$

A novidade é calculada pela Equação 4.3, em que se obtém a porcentagem de amostras novas geradas na amostra em relação ao número total de amostras que foram geradas. Como essa métrica define a capacidade da criação de novas moléculas, ela é comparada em relação ao espaço de possibilidades fora dos dados observados, o que permite definir se o modelo está tendo a capacidade de produzir respostas novas e variadas.

A novidade é importante, pois o foco principal da tarefa de geração de moléculas é a descoberta de dados novos que sejam diferentes dos dados já presentes na base de dados original.

4.4.4 Média de perda por época

Na análise comparativa de algoritmos de aprendizado de máquina, a métrica de perda por época é fundamental, pois fornece uma visão clara da convergência, estabilidade e eficiência do treinamento conforme o tempo passa (WANG et al., 2022).

Para calcular a perda por época, utiliza-se o Erro Quadrático Médio (MSE), que é uma métrica amplamente utilizada para avaliar a precisão de um modelo de aprendizado de máquina (ALEXANDER; ALEXANDER, 1986). A fórmula do MSE é dada por:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{4.4}$$

onde n é o número de exemplos no conjunto de dados, y_i são os valores reais (valores de saída desejados) e $\hat{y}_i$ são os valores previstos ou gerados pelo modelo.

A Equação 4.4 mede a média das diferenças quadráticas entre os valores previstos e os valores reais. Quanto menor o valor da MSE, melhor o modelo, pois significa que as previsões ou que as amostras geradas estão mais próximas dos valores reais.

Nesta análise, é possível verificar se o algoritmo está aprendendo de forma consistente ou se há oscilações abruptas de erro nas épocas, identificar qual algoritmo converge mais rapidamente pra estabilidade e entender se o modelo está aprendendo bem o padrão dos dados ou não.

4.4.5 Tempo de execução

O tempo de execução é uma métrica de eficiência computacional que determina o tempo necessário para a execução de um algoritmo com base na quantidade de entradas geradas. Nessa métrica, é mais interessante um tempo menor de execução.

Avaliar o tempo de execução é fundamental para medir a eficiência de um algoritmo, pois um tempo de execução menor indica um uso mais otimizado dos recursos computacionais. Isso possibilita a obtenção de resultados de forma mais rápida e com menor custo computacional, sem comprometer a qualidade da geração de moléculas (PAPADIMITRIOU, 2003).

4.5 Configuração Experimental

4.5.1 Divisão dos Dados

Para os conjuntos de dados CheMBL e PubChem, foram feitos recortes de conjunto de dados com 1000 entradas aleatórias em cada, sendo assim divididos em 800 entradas para treino, 100 entradas para validação e 100 entradas para teste.

Para o conjunto de dados ZINC, foi utilizada uma biblioteca que importa diretamente os dados do seu banco, então a divisão foi feita em 220.011 entradas para treino, 24.445 entradas para validação e 5.000 entradas para teste.

4.5.2 Treinamento e Avaliação

Na seção anterior, é possível perceber que o conjunto de dados ZINC é um conjunto de dados consideravelmente maior que os conjuntos de dados CheMBL e PubChem. Por esta razão e pela consideração do custo computacional, o número de épocas escolhidas para os conjuntos de dados CheMBL e PubChem foi de 100 e o número de épocas escolhidas para o conjunto de dados ZINC foi de 50.

As métricas de avaliação foram calculadas após o experimento de cada um dos métodos, de forma que as moléculas geradas eram armazenadas em uma variável e os dados dessa variável eram depois validados (validação), comparados entre si (unicidade) e comparados com o conjunto de dados real (novidade) para que cada uma dessas métricas de avaliação fosse calculada com exatidão.

5 Análise e discussão dos resultados

Nesta seção, é apresentada a análise dos resultados obtidos com os experimentos realizados para cada um dos conjuntos de dados e para cada um dos métodos de geração de moléculas usados. Serão usadas as métricas de qualidade das moléculas de validade, unicidade e novidade e que conclusões podem ser obtidas a partir desses resultados.

5.1 Gráficos de perda por execução

5.1.1 Conjunto de dados CheMBL - GANs

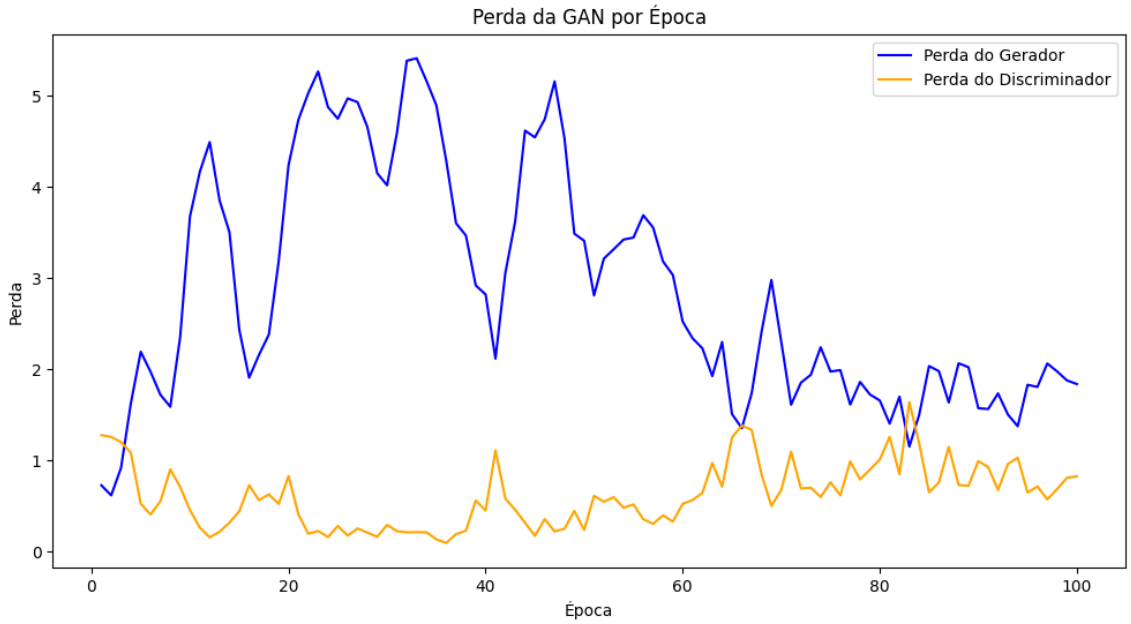
Nas GANs, existem dois tipos de perda: a perda do gerador e a perda do discriminador.

A perda do gerador está relacionada à qualidade dos dados gerados, medindo a distância entre os dados gerados pela rede geradora e os dados reais do conjunto de treinamento. Logo, minimizar essa perda significa que o gerador está produzindo dados cada vez mais realistas e difíceis de serem distinguidos dos dados reais (PAN et al., 2020).

Enquanto isso, a perda do discriminador está relacionada à sua capacidade de distinguir os dados reais dos dados gerados. Com isso, essas duas perdas são inversamente proporcionais, pois aumentar a perda de um implica em diminuir a perda do outro.

Para a comparação do desempenho, é mais importante considerar a análise da perda do gerador, pois ele é o componente responsável por gerar novos dados de saída e, assim, seu valor de perda representa a distância entre os dados gerados por ele e os dados presentes no conjunto de dados real. Logo, neste trabalho, em todas as análises referentes ao desempenho da perda por época das GANs, será considerada a perda do gerador.

Figura 4 – Perda por Época - CheMBL - GANs



Fonte: Elaborado pelo autor no Google Colab.

A Figura 4 ilustra o gráfico da perda por época das GANs na base CheMBL. O eixo x indica a época e o eixo y indica a perda. Pela ilustração, pode-se observar que os valores de perda obtidos pelas GANs foram:

- Perda do gerador: com o valor mínimo de 0.62 e valor máximo de 5.46, possui uma média de 2.97.
- Perda do discriminador: com o valor mínimo de 0.12 e valor máximo de 1.64, possui uma média de 0.55.

Com isso, os dois componentes das GANs apresentam uma margem de perda estabelecida que vai se estabilizando conforme as épocas vão acontecendo.

O ponto importante de observação desses gráficos é que um valor numérico de perda menor representa um desempenho melhor, uma vez que está sendo avaliada a qualidade dos dados gerados e, assim, a distância entre o conjunto de dados gerados e o conjunto de dados real deve ser a menor possível para a avaliação desta métrica.

5.1.2 Conjunto de dados CheMBL - VAEs

Figura 5 – Perda por Época - CheMBL - VAEs

Fonte: Elaborado pelo autor no Google Colab.

A Figura 5 ilustra o gráfico da perda por época dos VAEs na base CheMBL. O eixo x indica a época e o eixo y indica a perda. Pela ilustração, pode-se observar que a perda obtida pelos VAEs possui um valor mínimo de 1794.04, valor máximo de 17523.47, e média aproximada de 3007.21.

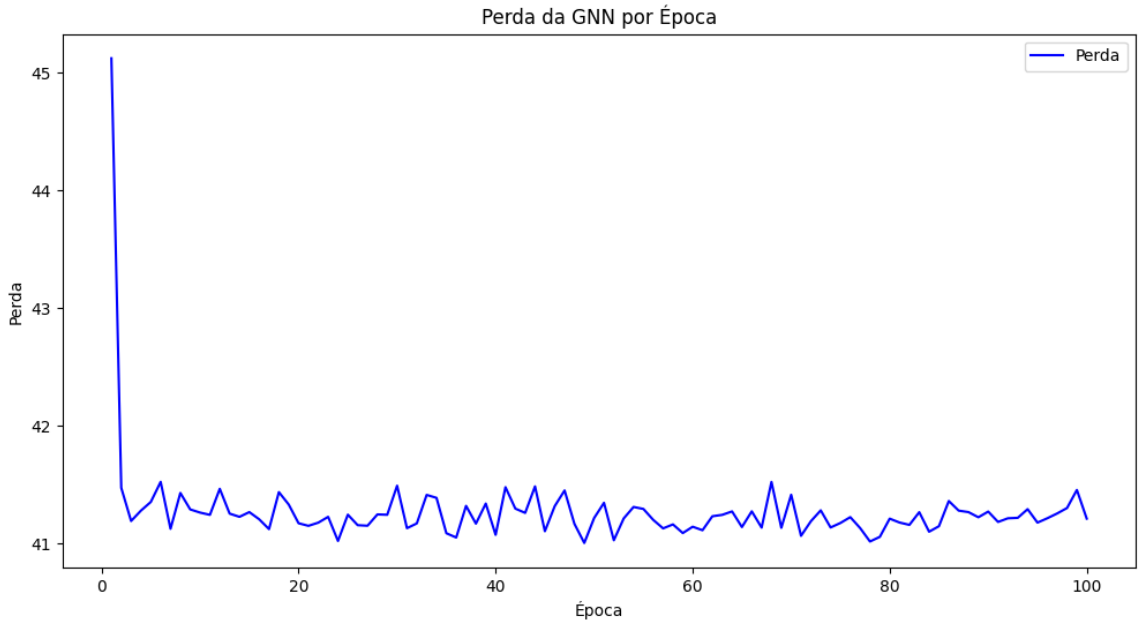
É muito evidente pelo gráfico que a perda dos VAEs começa com valores muito altos e ele consegue diminuir e se estabilizar rapidamente com o tempo, porém, ainda assim, ele apresenta valores de perda muito altos e o seu início aumenta em grande quantidade a média de perda por época do conjunto todo.

5.1.3 Conjunto de dados CheMBL - GNNs

A Figura 6 ilustra o gráfico da perda por época das GNNs na base CheMBL. O eixo x indica a época e o eixo y indica a perda. Pela ilustração, pode-se observar que a perda obtida pelas GNNs possui um valor mínimo de 41.05, valor máximo de 45.12, e média aproximada de 41.21.

Pelo gráfico, é possível perceber que as primeiras épocas da GNN possuem perda maior que as épocas restantes, que vão se estabilizando e atingindo valores próximos cada vez mais rápido, é possível perceber que, de todas execuções para este conjunto de dados, a execução do GNN é a que possui menor diferença entre os valores mínimos e máximos de perda em toda sua execução.

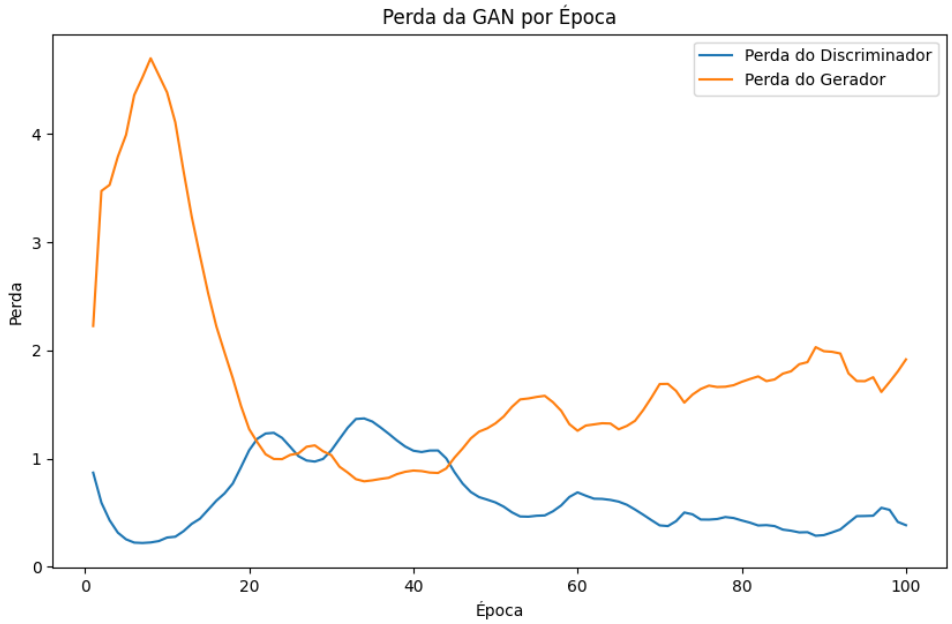
Figura 6 – Perda por Época - CheMBL - GNNs



Fonte: Elaborado pelo autor no Google Colab.

5.1.4 Conjunto de dados PubChem - GANs

Figura 7 – Perda por Época - PubChem - GANs



Fonte: Elaborado pelo autor no Google Colab.

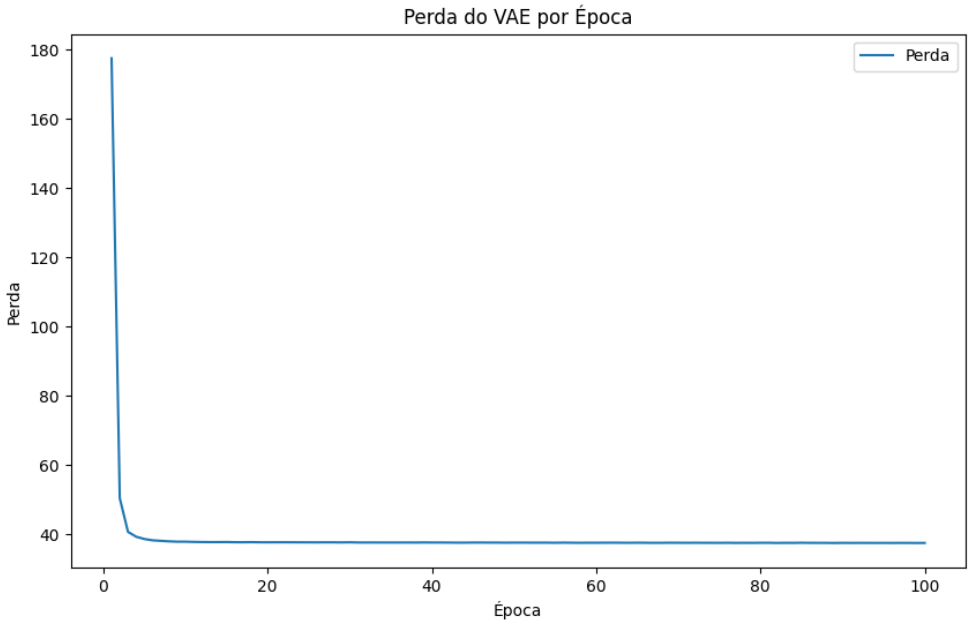
A Figura 7 ilustra o gráfico da perda por época das GANs na base PubChem. O eixo x indica a época e o eixo y indica a perda. Pela ilustração, pode-se observar que os valores de perda obtidos pelas GANs foram:

- Perda do gerador: com o valor mínimo de 0.70 e valor máximo de 4.95, possui uma média de 1.33.
- Perda do discriminador: com o valor mínimo de 0.20 e valor máximo de 1.44, possui uma média de 0.62.

Com isso, os dois componentes das GANs apresentam uma margem de perda estabelecida que vai se estabilizando conforme as épocas vão acontecendo. É possível perceber que o gerador apresenta perdas maiores no começo do experimento e, com o tempo, essas perdas diminuem e se estabilizam.

5.1.5 Conjunto de dados PubChem - VAEs

Figura 8 – Perda por Época - PubChem - VAEs



Fonte: Elaborado pelo autor no Google Colab.

A Figura 8 ilustra o gráfico da perda por época dos VAEs na base PubChem. O eixo x indica a época e o eixo y indica a perda. Pela ilustração, pode-se observar que a perda obtida pelos VAEs possui um valor mínimo de 36.79, valor máximo de 178.30, e média aproximada de 37.59.

Pelo gráfico, é possível notar que a perda dos VAEs começa com um valor muito alto, mas rapidamente ele reduziu e estabilizou para uma faixa de perdas com valores muito semelhantes, indicando que o modelo convergiu pra uma região estável.

5.1.6 Conjunto de dados PubChem - GNNs

Figura 9 – Perda por Época - PubChem - GNNs

Fonte: Elaborado pelo autor no Google Colab.

A Figura 9 ilustra o gráfico da perda por época das GNNs na base PubChem. O eixo x indica a época e o eixo y indica a perda. Pela ilustração, pode-se observar que a perda obtida pelas GNNs possui um valor mínimo de 39.91, valor máximo de 44.00, e média aproximada de 40.19.

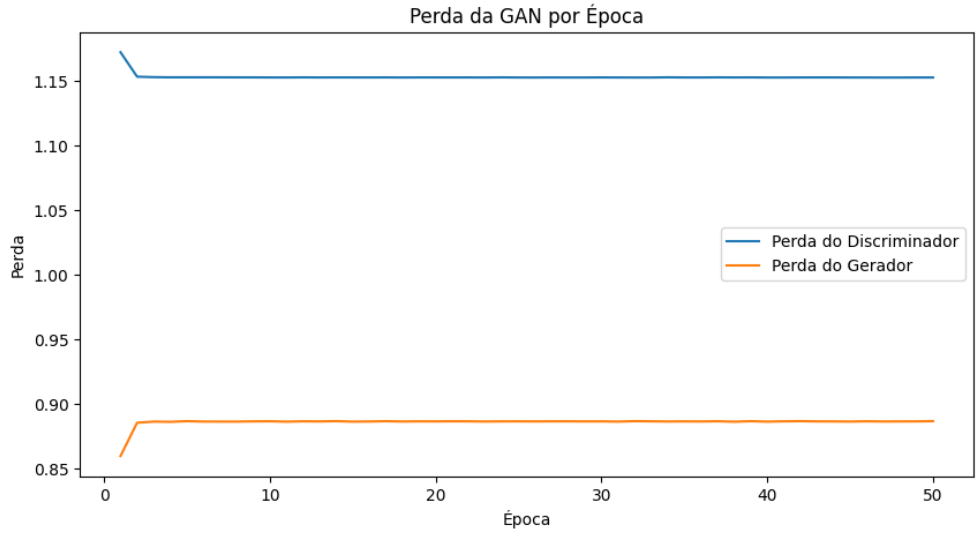
Novamente, as GNNs possuem uma grande estabilidade no que se diz a perda referente às suas épocas, e nesta execução, há uma estabilidade nos valores de perda que possui alguns picos maiores de perda, mas que num geral, não interferem na média geral de perda que se mantém próximo ao valor mínimo de perda da amostra.

5.1.7 Conjunto de dados ZINC - GANs

A Figura 10 ilustra o gráfico da perda por época das GANs na base ZINC. O eixo x indica a época e o eixo y indica a perda. Pela ilustração, pode-se observar que os valores de perda obtidos pelas GANs foram:

- Perda do gerador: com o valor mínimo de 0.87 e valor máximo de 0.89, possui uma média de 0.88.
- Perda do discriminador: com o valor mínimo de 1.15 e valor máximo de 1.16, possui uma média de 1.15.

Figura 10 – Perda por Época - ZINC - GANs



Fonte: Elaborado pelo autor no Google Colab.

Por ser um conjunto de dados com uma quantidade de dados bem superior aos demais, o conjunto de dados ZINC garante uma grande estabilidade nas perdas por época, com poucas variações e registra também a única vez neste experimento em que a perda do gerador é inferior à perda do discriminador durante todas as épocas.

5.1.8 Conjunto de dados ZINC - VAEs

Figura 11 – Perda por Época - ZINC - VAEs

Fonte: Elaborado pelo autor no Google Colab.

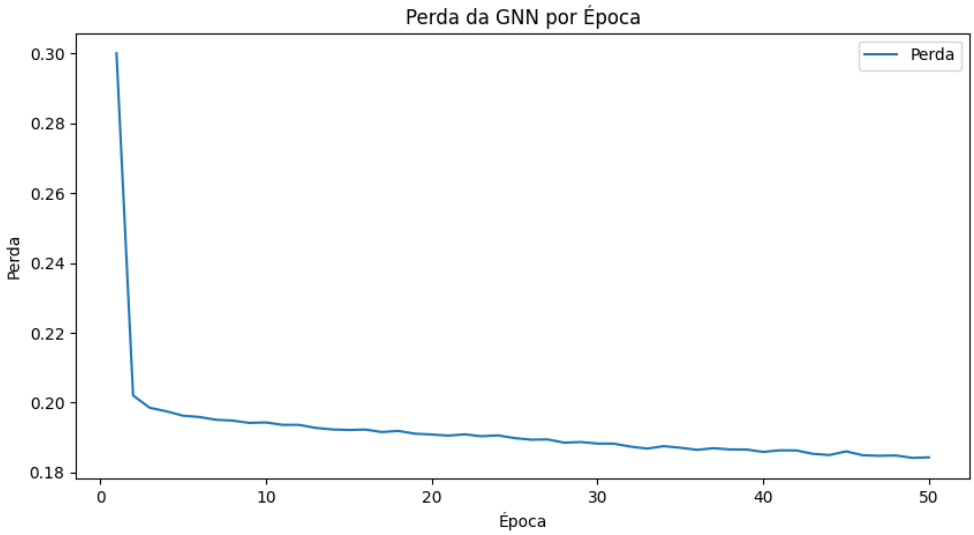
A Figura 11 ilustra o gráfico da perda por época dos VAEs na base ZINC. O eixo x indica a época e o eixo y indica a perda. Pela ilustração, pode-se observar que a perda

obtida pelos VAEs possui um valor mínimo de 178.55, valor máximo de 225.34, e média aproximada de 179.46.

Novamente, o gráfico aponta uma perda alta no começo do experimento que rapidamente é reduzido e estabilizado para uma faixa de perdas com valores muito semelhantes, assim, o modelo converge para uma região estável.

5.1.9 Conjunto de dados ZINC - GNNs

Figura 12 – Perda por Época - ZINC - GNNs



Fonte: Elaborado pelo autor no Google Colab.

A Figura 12 ilustra o gráfico da perda por época das GNNs na base ZINC. O eixo x indica a época e o eixo y indica a perda. Pela ilustração, pode-se observar que a perda obtida pelas GNNs possui um valor mínimo de 0.18, valor máximo de 0.30, e média aproximada de 0.19.

Com uma diferença de 0.12 entre seus valores mínimo e máximo, a GNN possui uma ótima performance na média de perdas por época para este conjunto de dados, indicando também perdas muito menores se comparada com os outros conjuntos de dados desse experimento e apresenta uma média de perda muito próxima do valor mínimo de perda da amostra.

De maneira geral, observando os gráficos desta seção, é notável que a média de perda por época dos VAEs e das GNNs possui comportamento semelhante para todos os conjuntos de dados, com valores maiores de início e valores de perda menores e estabilizados para o restante das épocas.

Para as GANs, é possível notar que os gráficos obtiveram comportamentos diferentes nas médias de perda por época, principalmente considerando a distância entre as

perdas do gerador e discriminador e a variação dos valores de perda. A principal diferença é vista no conjunto de dados ZINC, pois obtém valores de perda muito semelhantes para todas as épocas, enquanto que nos outros conjuntos, há uma variação perceptível nos valores de perda durante as épocas.

5.2 Comparativo geral por métricas

5.2.1 Validade

A validade se refere à proporção de amostras geradas que satisfazem as regras e restrições do domínio molecular, garantindo que as moléculas sejam válidas para o fim que foram geradas, podendo-se observar fatores e regras químicas, como por exemplo a valência dos átomos.

Tabela 1 – Validade para diferentes métodos e conjuntos de dados.

Conjunto de dados	**GANs**	**VAEs**	**GNNs**
CheMBL	100%	100%	100%
PubChem	100%	100%	100%
ZINC	100%	100%	100%

Fonte: Elaborado pelo autor.

A Tabela 1 exibe um comparativo do desempenho de cada um dos algoritmos para cada conjunto de dados baseado na métrica de validade.

Os resultados para esse teste foram muito positivos, pois indica que todos os modelos conseguiram gerar 100% de moléculas válidas independente do tamanho do conjunto de dados e do formato de seus dados, indicando que todos modelos são recomendados para a métrica de validade pois geram moléculas válidas do ponto de vista químico.

5.2.2 Unicidade

A unicidade avalia a diversidade de exemplos gerados, verificando a proporção de exemplos únicos que estão contidos no conjunto gerado.

A Tabela 2 exibe um comparativo do desempenho de cada um dos algoritmos para cada conjunto de dados baseado na métrica de unicidade.

Para a métrica de unicidade, podem-se inferir algumas conclusões com base nos resultados apresentados:

Em geral, todos os modelos obtiveram 0% de moléculas únicas para o conjunto de dados ZINC, e isso pode indicar que, como o conjunto de dados ZINC era um conjunto de

Tabela 2 – Unicidade para diferentes métodos e conjuntos de dados.

Conjunto de dados	GANs	VAEs	GNNs
CheMBL	4%	4%	14%
PubChem	4%	4%	14%
ZINC	0%	0%	0%

Fonte: Elaborado pelo autor.

dados com número de amostras muito superior aos outros conjuntos de dados, os modelos deste experimento não performam muito bem na geração de moléculas distintas umas das outras se o conjunto de dados do experimento for muito grande.

Agora, analisando os números dos conjuntos de dados CheMBL e PubChem, que são conjuntos de dados que possuem um número inferior de amostras em seu conjunto, foi possível obter valores melhores que aqueles observados no conjunto de dados do ZINC, indicando que, em conjuntos de dados menores, há um espaço amostral menor de dados, logo o espaço complementar para gerar dados diferentes é maior e, com isso, os algoritmos possuem maior abrangência para explorar e criar novos dados, o que aumenta a eficiência desses algoritmos para esta métrica.

Ainda analisando cada método de geração de moléculas, pode-se notar que, para os conjuntos de dados CheMBL e PubChem, as GANs e os VAEs possuem desempenho de 4% enquanto que as GNNs possuem desempenho de 14%, isso indica que, para estes conjuntos de dados, as GNNs possuem um melhor desempenho na geração de moléculas únicas e deve ser a mais indicada para essa métrica de avaliação.

5.2.3 Novidade

A novidade avalia e mede a capacidade do modelo de criar amostras novas, ou seja, amostras que sejam diferentes daquelas amostras que já estão presentes no conjunto de treinamento original.

Tabela 3 – Novidade para diferentes métodos e conjuntos de dados.

Conjunto de dados	GANs	VAEs	GNNs
CheMBL	100%	100%	100%
PubChem	100%	100%	100%
ZINC	0%	0%	0%

Fonte: Elaborado pelo autor.

A Tabela 3 exibe um comparativo do desempenho de cada um dos algoritmos para cada conjunto de dados baseado na métrica de novidade.

Para a métrica de novidade, podem-se inferir algumas conclusões com base nos resultados apresentados:

Em geral, todos os modelos obtiveram 0% de moléculas novas para o conjunto de dados ZINC, e isso pode indicar que, como o conjunto de dados ZINC era um conjunto de dados com número de amostras muito superior aos outros conjuntos de dados, os modelos deste experimento não performam muito bem na geração de moléculas novas e diferentes das moléculas presentes no conjunto de dados original se o conjunto de dados do experimento for muito grande.

Agora, analisando os números dos conjuntos de dados CheMBL e PubChem, que são conjuntos de dados que possuem um número inferior de amostras em seu conjunto, foi possível obter valores muito melhores que aqueles observados no conjunto de dados do ZINC, indicando que conjuntos de dados menores podem ter performance melhor na criação de moléculas novas (diferentes das moléculas presentes no conjunto de dados original) entre as moléculas geradas.

Ainda analisando cada método de geração de moléculas, pode-se notar que, para os conjuntos de dados CheMBL e PubChem, todos os métodos de geração de moléculas possuem desempenho de 100%. Isso indica que, para estes conjuntos de dados, todos os algoritmos deste experimento possuem um ótimo desempenho na geração de moléculas novas e podem ser indicados para essa métrica de avaliação.

5.2.4 Média de perda por época

A média de perda por época avalia numericamente qual a distância e a diferença entre os dados gerados pelos modelos e os dados reais presentes no conjunto de dados. É uma métrica importante para analisar se os dados gerados destoam muito dos dados reais presentes nos conjuntos de dados.

Tabela 4 – Média de perda por época para diferentes métodos e conjuntos de dados.

Conjunto de dados	**GANs**	**VAEs**	**GNNs**
CheMBL	2.97	3007.21	41.21
PubChem	1.33	37.59	40.19
ZINC	0.88	179.46	0.19

Fonte: Elaborado pelo autor.

A Tabela 4 exibe um comparativo do desempenho de cada um dos algoritmos para cada conjunto de dados baseado na métrica de média de perda por época.

Com uma análise da tabela de resultados de média por perda de época, é possível inferir algumas conclusões:

Em relação às GANs, é possível notar valores de perda muito baixos para todos os conjuntos de dados, indicando que a distância entre os conjuntos de dados gerados com os conjuntos de dados reais é muito pequena, o que é um ponto positivo para os experimentos.

Outra análise possível de se inferir é que a perda do gerador é ainda menor para o conjunto de dados ZINC (que é um conjunto de dados maior) do que para os outros, indicando uma melhor performance do experimento para conjuntos de dados maiores do que para conjuntos de dados menores, visto também que a perda do gerador foi menor que a perda do discriminador apenas no conjunto de dados ZINC.

Em relação aos VAEs, a média de perdas pode parecer um pouco confusa, dado que os dois conjuntos de dados menores (CheMBL e PubChem) possuem média de perda muito discrepantes, com um tendo valor de 3007.21 e o outro tendo valor de 37.59. Alguns motivos para essa diferença tão acentuada podem ocorrer pois as moléculas do CheMBL podem apresentar estruturas mais complexas que dificultam a aprendizagem do modelo ou que simplesmente os VAEs não sejam adaptados para ter uma boa performance com os dados presentes no conjunto de dados do CheMBL e seja necessário ajustes técnicos para melhorar esse desempenho.

Em relação às GNNs, é possível notar que os valores de perda são baixos para conjuntos de dados menores (CheMBL e PubChem), mas que são muito menores para o conjunto de dados ZINC, então, a performance das GNNs para conjuntos de dados maiores é ótima e apresenta médias de perdas por época próximas a zero.

Em relação ao conjunto de dados CheMBL, é possível notar que a média de perda por época dos VAEs resultou em um número muito grande, logo seus dados gerados são muito destoantes do conjunto de dados original e com isso sua performance não é muito positiva. Dentre os restantes, as GNNs performam bem, mas as GANs possuem uma performance muito superior, dado que sua perda é a menor dentre os métodos, e deve ser o método mais indicado para este conjunto de dados com base na perda.

Em relação ao conjunto de dados PubChem, é possível notar que a média de perda por época dos VAEs e das GNNs resultam em perdas não muito grandes, porém, novamente, as GANs possuem uma performance muito superior, dado que sua perda é a menor dentre os métodos, e deve ser o método mais indicado para este conjunto de dados com base na perda.

Em relação ao conjunto de dados ZINC, é possível notar que a média de perda por época dos VAEs deu um número muito grande, com isso sua performance não é muito positiva. Dentre os restantes, as GANs e as GNNs têm uma performance ótima, com valores de perda próximos a zero, porém é válido destacar que as GNNs possuem uma performance melhor que as GANs, dado que sua perda é a menor dentre os métodos

utilizados.

5.2.5 Tempo de execução

O tempo de execução avalia, em segundos, quanto tempo foi necessário para cada método gerar as moléculas desde a primeira até a última época em uma execução, e é importante para avaliar custo computacional e desempenho, pois um modelo mais rápido é mais interessante para um grande número de execuções dos algoritmos.

Tabela 5 – Tempos de execução para diferentes métodos e conjuntos de dados.

Conjunto de dados	**GANs**	**VAEs**	**GNNs**
CheMBL	42.84	16.97	39.30
PubChem	22.79	20.61	26.81
ZINC	440.66	156.07	179.87

Fonte: Elaborado pelo autor.

A Tabela 5 exibe um comparativo do desempenho de cada um dos algoritmos para cada conjunto de dados baseado na métrica de tempo de execução. Com uma análise dessa tabela, é possível inferir algumas conclusões:

Em relação às GANs, é possível notar que o tempo de execução para os conjuntos de dados menores é menor, com destaque para o conjunto de dados PubChem, que possui o menor tempo dentre os experimentos. Como esperado, o tempo de execução para o conjunto de dados ZINC é o maior dentre as execuções.

Em relação aos VAEs, é possível notar que o tempo de execução para os conjuntos de dados menores é menor, com destaque para o conjunto de dados CheMBL, que possui o menor tempo dentre os experimentos. Como esperado, o tempo de execução para o conjunto de dados ZINC é o maior dentre as execuções.

Em relação aos GNNs, é possível notar que o tempo de execução para os conjuntos de dados menores é menor, com destaque para o conjunto de dados PubChem, que possui o menor tempo dentre os experimentos. Como esperado, o tempo de execução para o conjunto de dados ZINC é o maior dentre as execuções.

Em relação ao conjunto de dados CheMBL, pode-se destacar que o VAE possui o menor tempo de execução dentre os experimentos. Dentre os métodos restantes, as GANs e as GNNs possuem tempos de execução muito semelhantes, com um desempenho das GNNs um pouco melhor que o das GANs.

Em relação ao conjunto de dados PubChem, todos os métodos possuem tempos de execução muito semelhantes, com segundos de diferença. Logo, os VAEs possuem um

desempenho um pouco melhor, seguidos pelas GANs e as GNNs com o maior tempo de execução dentre esses experimentos.

Em relação ao conjunto de dados ZINC, é possível notar que os VAEs e as GNNs possuem os melhores tempos de execução, abaixo de 200 segundos, com os VAEs tendo um desempenho melhor nesse quesito do que os GNNs. Em contrapartida, as GANs possuem o maior tempo de execução, excedendo em até duas vezes mais o tempo dos outros métodos, o que pode ser um indicativo de que as GANs podem ter um desempenho bem mais lento quando estão usando conjuntos de dados maiores para a geração de moléculas.

5.3 Avaliativo geral por algoritmos

Utilizando os dados da seção anterior, foram construídas tabelas para exibir os resultados obtidos pelos algoritmos em cada uma das execuções, divididos por algoritmo e por conjunto de dados, para avaliar o desempenho daquela execução no geral, destacando seus pontos fortes, medianos e fracos.

5.3.1 GANs

5.3.1.1 Conjunto de dados CheMBL

Tabela 6 – Métricas de avaliação para GANs - CheMBL

Conjunto	Validade	Unicidade	Novidade	Média de perda	Execução (s)
CheMBL	100%	4%	100%	2.97	42.84

Fonte: Elaborado pelo autor.

A Tabela 6 exibe todos os resultados apresentados pelas GANs nas métricas de avaliação quando executadas utilizando o conjunto de dados CheMBL.

Analisando esses resultados, é possível perceber que, para este conjunto de dados, a GAN apresenta ótimos resultados na validação de moléculas geradas (validade), na criação de moléculas novas (novidade) e também apresenta uma média de perda baixa, que é um ponto positivo na geração de moléculas. Porém, apresenta uma baixa porcentagem de unicidade e sua execução é um pouco mais lenta em comparação com outros métodos.

5.3.1.2 Conjunto de dados PubChem

A Tabela 7 exibe todos os resultados apresentados pelas GANs nas métricas de avaliação quando executadas utilizando o conjunto de dados PubChem.

Analisando esses resultados, é possível perceber que, para este conjunto de dados, a GAN apresenta ótimos resultados na validação de moléculas geradas (validade), na criação

Tabela 7 – Métricas de avaliação para GANs - PubChem

Conjunto	Validade	Unicidade	Novidade	Média de perda	Execução (s)
PubChem	100%	4%	100%	1.33	22.79

Fonte: Elaborado pelo autor.

de moléculas novas (novidade) e também apresenta uma média de perda baixa, que é um ponto positivo na geração de moléculas. Porém, apresenta uma baixa porcentagem de unicidade e seu tempo de execução está na média em comparação com outros métodos.

5.3.1.3 Conjunto de dados ZINC

Tabela 8 – Métricas de avaliação para GANs - ZINC

Conjunto	Validade	Unicidade	Novidade	Média de perda	Execução (s)
ZINC	100%	0%	0%	0.88	440.66

Fonte: Elaborado pelo autor.

A Tabela 8 exibe todos os resultados apresentados pelas GANs nas métricas de avaliação quando executadas utilizando o conjunto de dados ZINC.

Analisando esses resultados, é possível perceber que, para este conjunto de dados, a GAN apresenta ótimos resultados na validação de moléculas geradas (validade) e também apresenta uma média de perda bem baixa, que é um ponto positivo na geração de moléculas. Porém, não consegue gerar moléculas únicas ou novas (unicidade e novidade) e sua execução é bem mais lenta em comparação com outros métodos.

5.3.2 VAEs

5.3.2.1 Conjunto de dados CheMBL

Tabela 9 – Métricas de avaliação para VAEs - CheMBL

Conjunto	Validade	Unicidade	Novidade	Média de perda	Execução (s)
CheMBL	100%	4%	100%	3007.21	16.97

Fonte: Elaborado pelo autor.

A Tabela 9 exibe todos os resultados apresentados pelos VAEs nas métricas de avaliação quando executados utilizando o conjunto de dados CheMBL.

Analisando esses resultados, é possível perceber que, para este conjunto de dados, o VAE apresenta ótimos resultados na validação de moléculas geradas (validade), na criação de moléculas novas (novidade) e também apresenta um tempo de execução bem mais rápido do que em comparação com outros métodos. Porém, apresenta uma baixa porcentagem de unicidade e possui uma média de perda muito grande, indicando que os dados gerados destoam muito dos dados presentes no conjunto de dados original.

5.3.2.2 Conjunto de dados PubChem

Tabela 10 – Métricas de avaliação para VAEs - PubChem

Conjunto	Validade	Unicidade	Novidade	Média de perda	Execução (s)
PubChem	100%	4%	100%	37.59	20.61

Fonte: Elaborado pelo autor.

A Tabela 10 exibe todos os resultados apresentados pelos VAEs nas métricas de avaliação quando executados utilizando o conjunto de dados PubChem.

Analisando esses resultados, é possível perceber que, para este conjunto de dados, o VAE apresenta ótimos resultados na validação de moléculas geradas (validade), na criação de moléculas novas (novidade) e também apresenta um tempo de execução rápido e melhor que os demais. Porém, apresenta uma baixa porcentagem de unicidade e possui uma média de perda razoável em comparação com outros métodos.

5.3.2.3 Conjunto de dados ZINC

Tabela 11 – Métricas de avaliação para VAEs - ZINC

Conjunto	Validade	Unicidade	Novidade	Média de perda	Execução (s)
ZINC	100%	0%	0%	179.46	156.07

Fonte: Elaborado pelo autor.

A Tabela 11 exibe todos os resultados apresentados pelos VAEs nas métricas de avaliação quando executados utilizando o conjunto de dados ZINC.

Analisando esses resultados, é possível perceber que, para este conjunto de dados, o VAE apresenta ótimos resultados na validação de moléculas geradas (validade) e também apresenta o tempo de execução mais baixo dentre todos os algoritmos. Porém, não consegue gerar moléculas únicas ou novas (unicidade e novidade) e sua média de perda é muito grande em comparação com os outros métodos.

5.3.3 GNNs

5.3.3.1 Conjunto de dados CheMBL

Tabela 12 – Métricas de avaliação para GNNs - CheMBL

Conjunto	Validade	Unicidade	Novidade	Média de perda	Execução (s)
CheMBL	100%	14%	100%	41.21	39.30

Fonte: Elaborado pelo autor.

A Tabela 12 exibe todos os resultados apresentados pelas GNNs nas métricas de avaliação quando executadas utilizando o conjunto de dados CheMBL.

Analisando esses resultados, é possível perceber que, para este conjunto de dados, a GNN apresenta ótimos resultados na validação de moléculas geradas (validade), na criação de moléculas novas (novidade) e também consegue gerar moléculas únicas em um desempenho melhor que os demais algoritmos. Porém, apresenta uma média de perda razoável e um tempo de execução mediano em comparação com os outros métodos.

5.3.3.2 Conjunto de dados PubChem

Tabela 13 – Métricas de avaliação para GNNs - PubChem

Conjunto	Validade	Unicidade	Novidade	Média de perda	Execução (s)
PubChem	100%	14%	100%	40.19	26.81

Fonte: Elaborado pelo autor.

A Tabela 13 exibe todos os resultados apresentados pelas GNNs nas métricas de avaliação quando executadas utilizando o conjunto de dados PubChem.

Analisando esses resultados, é possível perceber que, para este conjunto de dados, a GNN apresenta ótimos resultados na validação de moléculas geradas (validade), na criação de moléculas novas (novidade) e também consegue gerar moléculas únicas em um desempenho melhor que os demais algoritmos. Porém, apresenta uma média de perda razoável e um tempo de execução mediano em comparação com os outros métodos.

5.3.3.3 Conjunto de dados ZINC

Tabela 14 – Métricas de avaliação para GNNs - ZINC

Conjunto	Validade	Unicidade	Novidade	Média de perda	Execução (s)
ZINC	100%	0%	0%	0.19	179.87

Fonte: Elaborado pelo autor.

A Tabela 14 exibe todos os resultados apresentados pelas GNNs nas métricas de avaliação quando executadas utilizando o conjunto de dados ZINC.

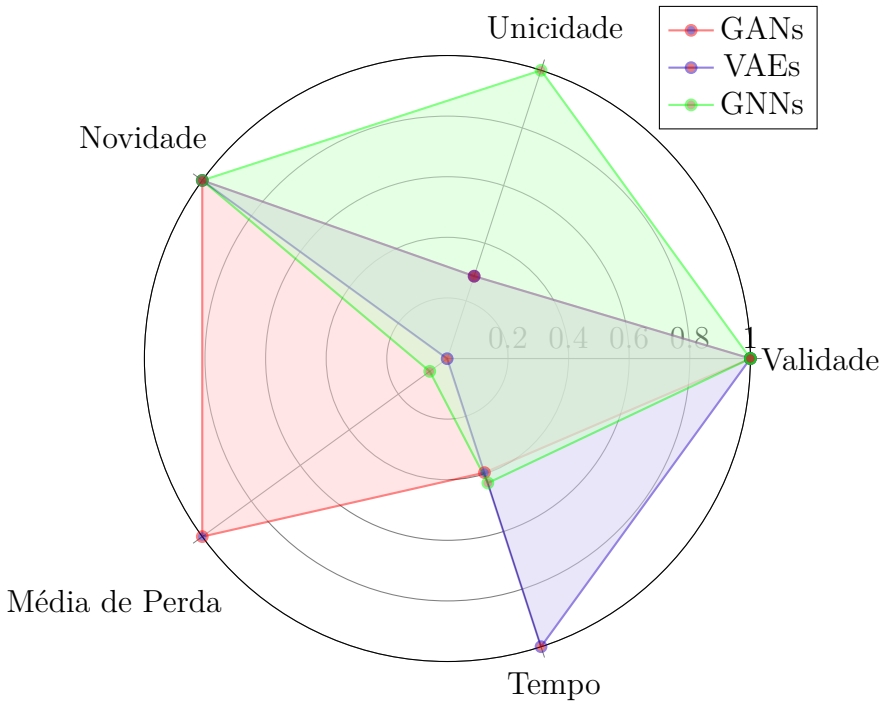
Analisando esses resultados, é possível perceber que, para este conjunto de dados, a GNN apresenta ótimos resultados na validação de moléculas geradas (validade), apresenta a menor média de perda de todos os métodos e também um tempo de execução médio em comparação aos outros algoritmos. Porém, não consegue gerar moléculas únicas ou novas (unicidade e novidade).

5.4 Desempenho geral por conjuntos de dados

Utilizando os dados da seção anterior, foram construídos gráficos de radar para ilustrar de forma visual o desempenho de cada um dos algoritmos para cada uma das métricas de avaliação utilizadas, divididos pelo conjunto de dados.

5.4.1 Desempenho para o conjunto de dados CheMBL

Figura 13 – Gráfico de radar de desempenho para o conjunto CheMBL.



Fonte: Elaborado pelo autor.

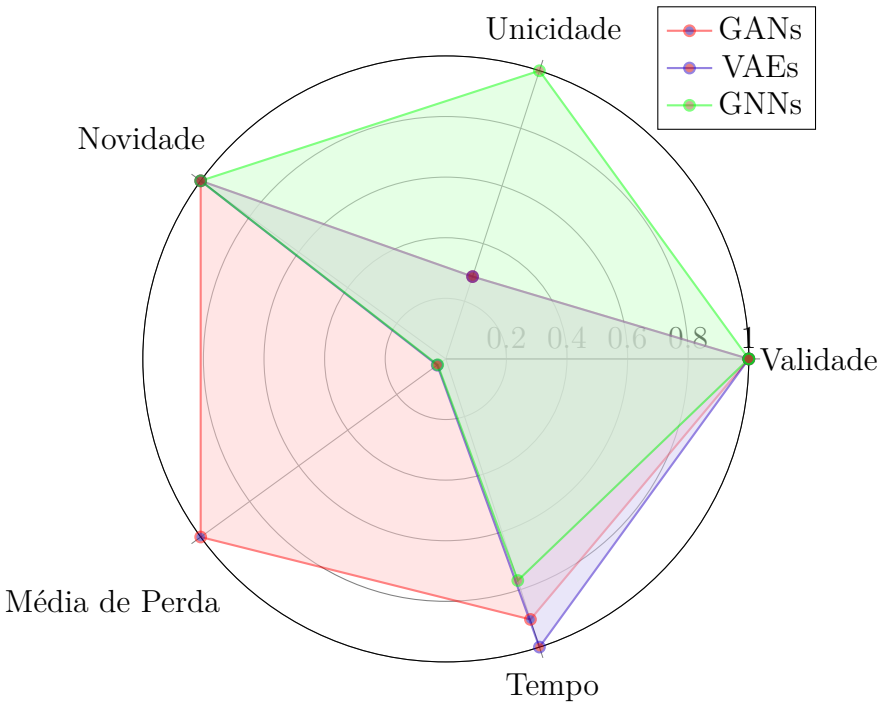
A Figura 13 apresenta uma visualização multidimensional do desempenho dos algoritmos GANs, VAEs e GNNs em relação às cinco métricas utilizadas neste trabalho para o conjunto de dados CheMBL. Cada eixo representa uma métrica e os valores nor-

malizados são plotados ao longo desses eixos de forma que os valores mais próximos da borda representam o melhor desempenho em determinada métrica.

A visualização desse gráfico permite afirmar que as GANs se destacam pelo seu desempenho na média de perda por época, as GNNs se destacam por sua unicidade e os VAEs se destacam pelo seu tempo de execução. Para a validade e para a novidade, todos os algoritmos atingem a performance de desempenho máxima.

5.4.2 Desempenho para o conjunto de dados PubChem

Figura 14 – Gráfico de radar de desempenho para o conjunto PubChem.



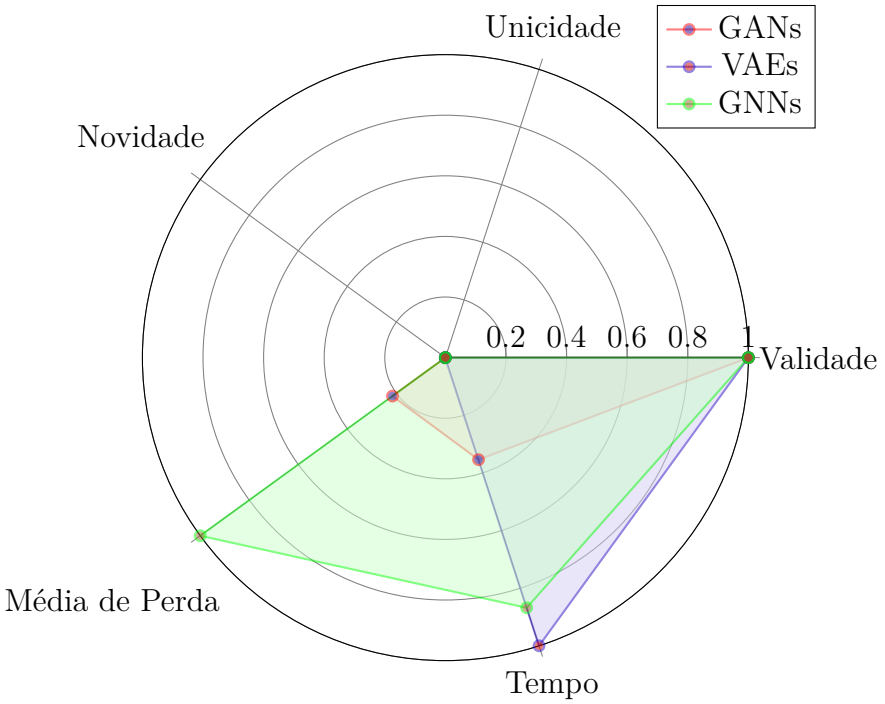
Fonte: Elaborado pelo autor.

A Figura 14 apresenta uma visualização multidimensional do desempenho dos algoritmos GANs, VAEs e GNNs em relação às cinco métricas utilizadas neste trabalho para o conjunto de dados PubChem. Cada eixo representa uma métrica e os valores normalizados são plotados ao longo desses eixos de forma que os valores mais próximos da borda representam o melhor desempenho em determinada métrica.

A visualização desse gráfico permite afirmar que as GANs se destacam pelo seu desempenho na média de perda por época e as GNNs se destacam por sua unicidade. Para a validade, a novidade e o tempo de execução, todos os algoritmos atingem valores próximos à performance de desempenho máxima.

5.4.3 Desempenho para o conjunto de dados ZINC

Figura 15 – Gráfico de radar de desempenho para o conjunto ZINC.



Fonte: Elaborado pelo autor.

A Figura 15 apresenta uma visualização multidimensional do desempenho dos algoritmos GANs, VAEs e GNNs em relação às cinco métricas utilizadas neste trabalho para o conjunto de dados ZINC. Cada eixo representa uma métrica e os valores normalizados são plotados ao longo desses eixos de forma que os valores mais próximos da borda representam o melhor desempenho em determinada métrica.

A visualização desse gráfico permite afirmar que as GNNs e as GANs se destacam pelo seu desempenho na média de perda por época e que os VAEs e as GNNs apresentam os melhores tempos de execução. Para a validade, todos os algoritmos atingem a performance de desempenho máxima. Porém, nenhum algoritmo conseguiu obter performance não nula na unicidade e na novidade, que são métricas muito importantes para esse estudo.

Para todos os conjuntos de dados, no geral, a visualização dos resultados evidencia que as GANs destacam-se de forma significativa nas médias de perda por época, apresentando um desempenho superior que os demais algoritmos não conseguem alcançar ou se aproximar. Enquanto isso, os VAEs e as GNNs demonstram desempenhos notáveis nas métricas de tempo de execução e unicidade, respectivamente. No entanto, nessas métricas, os outros algoritmos conseguem atingir valores próximos aos valores obtidos por eles.

Logo, considerando todas as análises e comparações entre algoritmos, métricas de avaliação e conjuntos de dados usados, é possível definir que, no geral, as GANs atingem desempenhos melhores que os outros algoritmos na geração de moléculas neste trabalho. Em segundo lugar, as GNNs apresentam melhores desempenhos para esta tarefa e os VAEs possuem o pior desempenho nos experimentos realizados seguindo as métricas utilizadas neste trabalho.

6 Conclusão

A geração de conjuntos de moléculas é um tema recorrente e que vem ganhando importância atualmente juntamente com os avanços da biotecnologia e da química medicinal. Com isso, torna-se importante o estudo dos melhores algoritmos e aplicações para garantir que o desempenho e o custo computacional sejam otimizados da melhor forma possível.

O foco de estudo desse trabalho foram as GANs, porém é interessante notar que, dependendo do contexto e das aplicações, outras opções de algoritmos podem ser mais interessantes de serem usadas devido à melhor aplicação e resultados provenientes do uso deles. Com base na análise e discussão de resultados, algumas conclusões podem ser constatadas.

Em geral, todos os algoritmos estudados nesse trabalho conseguiram gerar moléculas válidas de um ponto de vista físico-químico, provando que todos os algoritmos foram capazes de entender os padrões de complexidade de compostos moleculares e replicá-los na criação de novos.

Quando se é analisada a capacidade dos algoritmos de gerarem moléculas únicas e não acabarem gerando apenas moléculas iguais, observa-se que as GNNs possuem uma performance melhor do que a performance das GANs e dos VAEs, no geral.

Para conjuntos de dados específicos, também foi observado que todos os algoritmos estudados neste trabalho conseguiram gerar moléculas novas, ou seja, moléculas diferentes das moléculas que estão presentes no conjunto de dados de treinamento, o que é uma métrica boa para medir quanto esses algoritmos podem aprender sobre padrões químicos de compostos moleculares e replicar esses padrões na geração de novas moléculas válidas.

Em geral, foi possível perceber também que a média de perda por época das GANs obteve valores baixos, ou seja, a diferença entre as moléculas geradas pelas GANs e as moléculas dos dados reais era baixa, e isso significa que o gerador dessas redes neurais conseguiu obter sucesso na criação de dados que eram semelhantes aos conjuntos de dados reais e próximos aos padrões de complexidade presentes neles.

Quanto ao tempo de execução dos algoritmos, foi possível notar que as GANs não obtiveram o melhor tempo de execução dentre o restante dos algoritmos do estudo, porém seu desempenho não teve uma diferença tão grande do desempenho dos VAEs e das GNNs. Notou-se também que, para conjuntos de dados maiores, as GANs possuíam um tempo de execução que excedia duas vezes mais o tempo de execução dos VAEs e das GNNs, enquanto que em conjuntos de dados menores, como o CheMBL e o PubChem, seu tempo

de execução estava razoavelmente similar à média de tempo obtida por todos os métodos.

Com isso, é possível concluir que, em um geral, as GANs são uma ótima opção de algoritmo a ser usado para a geração de conjuntos de moléculas e representam o melhor algoritmo deste trabalho para essa finalidade, pois performam muito bem em determinados parâmetros e métricas de avaliação de qualidade estabelecidos, e mesmo naqueles parâmetros que os VAEs ou as GNNs têm performances melhores que a da GAN, seu desempenho não fica tão abaixo do desempenho deles e pode ser considerado como uma ferramenta viável para aquela determinada métrica da mesma forma.

6.1 Limitações

Houve algumas limitações durante a realização deste trabalho, as principais que podem ser destacadas são:

- Tamanho dos dados: Houve limitações relacionadas às quantidades de amostras dos dados, pois algumas bibliotecas não conseguiam carregar uma carga muito grande de moléculas presentes nos conjuntos de dados, o que refletiu em recortes não muito grandes dos dados e que pode ter tido um impacto na análise de desempenho.
- Parâmetros e Treinamento: Os hiperparâmetros (taxa de aprendizado, número de épocas) foram padronizados em números fixos para a realização das execuções baseado na primeira execução do algoritmo, então pode ser que os hiperparâmetros usados não fossem os valores ideais para otimizar o desempenho dos modelos.
- Generalização dos dados: O modelo foi treinado em um conjunto específico de dados, então pode ser que as aplicações de dados no mundo real não tenham o mesmo desempenho, pois as distribuições de dados dos conjuntos selecionados pode ser diferente das distribuições no contexto aplicado em vida real, dificultando a generalização da análise de desempenho.
- Capacidade do Modelo: Atualmente, existem arquiteturas híbridas utilizando mais de uma rede neural ou unindo diferentes algoritmos para melhorar os resultados. Porém, a análise foi primeiramente definida para testar apenas um algoritmo por vez e como cada um performava na tarefa definida.
- Captação dos dados do conjunto ZINC: Funções que exportavam recortes de dados do conjunto de dados ZINC apresentavam problemas de compatibilidade quando aplicados para um número grande de moléculas. Logo, uma função geral que exportava uma grande quantidade de dados não balanceados em treinamento, validação e teste teve que ser usada para esta base, causando alterações nos resultados e desempenho nulo nas métricas de unicidade e novidade.

6.2 Trabalhos Futuros

Para trabalhos futuros, existem algumas questões que não foram abordadas neste trabalho que podem ser possivelmente relevantes:

- Desempenho com base no recorte de conjuntos de dados: Os conjuntos de dados presentes nesse estudo tiveram um número fixo de amostras previamente estabelecido. É possível que aumentar ou diminuir os conjuntos de dados presentes nesse estudo pode afetar o desempenho dos algoritmos nas métricas de avaliação utilizadas.

- Conjuntos de dados com aplicações específicas: O foco deste trabalho foi analisar o desempenho da geração de moléculas usando algoritmos de IA generativa, porém esses mesmos algoritmos podem ser aplicados em diferentes contextos para ter seu desempenho avaliado, como na geração de textos, imagens, músicas, aúdios ou vídeos. Pode ser que determinados algoritmos obtenham desempenhos diferentes dependendo do contexto em que serão utilizados.

- Otimização de hiperparâmetros: Os modelos utilizados nesse trabalho seguiram os valores padrão para cada implementação. A alteração desses valores e a busca pelos melhores valores possíveis para esses parâmetros pode melhorar ainda mais os resultados obtidos neste trabalho.

- Análise técnica e de complexidade: É possível que os algoritmos utilizados neste trabalho possam ser melhorados ou implementados de forma diferente. Logo, um estudo aprofundado nos detalhes da implementação desses algoritmos pode proporcionar novas perspectivas sobre cada método.

- Exploração de novas técnicas: Existe uma enorme variedade de algoritmos de aprendizado de máquina que vão cada vez mais sendo atualizados e otimizados para melhor obtenção de desempenho e resultados. Conforme surjam novos algoritmos ou conforme os algoritmos atuais sejam melhorados ou combinados com outros algoritmos para aumento de desempenho, novos estudos de análise comparativa de desempenho podem ser realizados para conferir como o desempenho geral desses métodos pode evoluir.

Referências

AGGARWAL, A.; MITTAL, M.; BATTINENI, G. Generative adversarial network: An overview of theory and applications. *International Journal of Information Management Data Insights*, v. 1, n. 1, p. 100004, 2021. ISSN 2667-0968. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2667096820300045>. Citado na página 21.

ALEXANDER, S. T.; ALEXANDER, S. T. The mean square error (mse) performance criteria. *Adaptive Signal Processing: Theory and Applications*, p. 8–33, 1986. Citado na página 33.

ARúS-POUS, J. et al. Randomized smiles strings improve the quality of molecular generative models. *Journal of Cheminformatics*, 2020. Citado na página 26.

BLANCHARD, A. E.; STANLEY, C.; BHOWMIK, D. Using gans with adaptive training data to search for new molecules. *Journal of cheminformatics*, v. 13, p. 1–8, 2021. Citado na página 27.

BLASCHKE, T. et al. Application of generative autoencoder in de novo molecular design. *Molecular Informatics*, v. 37, n. 1-2, p. 1700123, 2018. Citado na página 18.

BOCK, H.-H. A history of k-means algorithms. *Institute of Statistics, RWTH Aachen University*, D-52056 Aachen, Germany, 2007. Citado na página 19.

BONGINI, P.; BIANCHINI, M.; SCARSELLI, F. Molecular generative graph neural networks for drug discovery. *Neurocomputing*, Elsevier, v. 453, p. 157–167, 2021. Disponível em: <https://www.sciencedirect.com/science/article/abs/pii/S0925231221005737>. Citado na página 31.

CAO, N. D.; KIPF, T. Molgan: An implicit generative model for small molecular graphs. In: . [S.l.: s.n.], 2022. Citado na página 25.

CINELLI, L. P. et al. Variational autoencoder. In: *Variational Methods for Machine Learning with Applications to Deep Networks*. [S.l.]: Springer, Cham, 2021. Citado na página 22.

GANGWAL, A. et al. Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities. *Frontiers in Pharmacology*, v. 15, p. 1331062, 2024. Citado na página 26.

GAULTON, A. et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, v. 40, n. D1, p. D1100–D1107, 2012. Citado na página 30.

GEEKSFORGEEKS. *What are Graph Neural Networks?* 2024. Available at: <https://www.geeksforgeeks.org/what-are-graph-neural-networks/>. Citado na página 24.

GOODFELLOW, I. et al. Generative adversarial nets. In: *Advances in Neural Information Processing Systems*. [S.l.: s.n.], 2014. v. 27. Citado 4 vezes nas páginas 17, 20, 25 e 26.

Google Developers. *GAN Structure*. 2022. Available at: <https://developers.google.com/machine-learning/gan/gan_structure?hl=pt-br>. Citado na página 21.

GóMEZ-BOMBARELLI, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 2018. Citado na página 31.

IRWIN, J. J.; SHOICHET, B. K. Zinc - a free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, v. 45, n. 1, p. 177–182, 2005. Citado na página 30.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, v. 349, n. 6245, p. 255–260, 2015. Citado na página 19.

KADURIN, A. et al. Drugan: An advanced generative adversarial autoencoder model for de-novo generation of new molecules. *Molecular Pharmaceutics*, v. 14, n. 9, p. 3098–3104, 2017. Citado na página 17.

KARAGIANNAKOS, S. *Graph Neural Networks - An overview*. 2020. Disponível em: <https://theaisummer.com/Graph_Neural_Networks/>. Citado na página 23.

KIM, S.; BOLTON, E. E. PubChem: A Large-Scale Public Chemical Database for Drug Discovery. In: DAINA, A.; PRZEWOSNY, M.; ZOETE, V. (Ed.). *Open Access Databases and Datasets for Drug Discovery*. Wiley-VCH, 2024. Disponível em: <https://doi.org/10.1002/9783527830497.ch2>. Citado na página 30.

KIM, S. et al. Pubchem substance and compound databases. *Nucleic Acids Research*, v. 44, n. D1, p. D1202–D1213, 2016. Citado na página 30.

KINGMA, D. P.; WELLING, M. An introduction to variational autoencoders. *Foundations and Trends in Machine Learning*, v. 12, n. 4, p. 307–392, 2019. Citado na página 22.

MAZIARKA, . et al. Mol-cyclegan: A generative model for molecular optimization. *Journal of Cheminformatics*, 2020. Citado na página 26.

MOGHADAM, M. M. et al. Game of gans: Game-theoretical models for generative adversarial networks. January 2022. Citado na página 21.

MORALES, E. F.; ESCALANTE, H. J. A brief introduction to supervised, unsupervised, and reinforcement learning. In: TORRES-GARCíA, A. A. et al. (Ed.). *Biosignal Processing and Classification Using Computational Learning and Intelligence*. Academic Press, 2022. p. 111–129. ISBN 9780128201251. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780128201251000178>. Citado na página 19.

NAEEM, S. et al. An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*, v. 13, n. 1, p. 1–, abr. 2023. ISSN 2210-142X. Disponível em: <http://dx.doi.org/10.12785/ijcds/130172>. Citado na página 19.

PAN, Z. et al. Loss functions of generative adversarial networks (gans): Opportunities and challenges. *IEEE Transactions on Emerging Topics in Computational Intelligence*, v. 4, n. 4, p. 500–522, ago. 2020. Citado na página 35.

PAPADIMITRIOU, C. H. Computational complexity. In: *Encyclopedia of Computer Science*. [S.l.]: John Wiley & Sons, 2003. p. 260–265. Citado na página 34.

POLYKOVSKIY, D. et al. Molecular sets (moses): A benchmarking platform for molecular generation models. 2018. Citado na página 25.

RAMAKRISHNAN, R. et al. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, v. 1, p. 140022, 2014. Citado na página 25.

RAZAVI-FAR, R. et al. (Ed.). *Generative adversarial learning: Architectures and applications*. [S.l.]: Springer, 2022. v. 217. (Intelligent Systems Reference Library, v. 217). Citado 3 vezes nas páginas 18, 25 e 31.

RENDO, A. S. *Molecule generation with GANs*. Dissertação (Mestrado) — Universitat Politècnica de Catalunya, 2022. Citado na página 27.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. Learning representations by back-propagating errors. *Nature*, v. 323, n. 6088, p. 533–536, 1986. Citado na página 20.

SINGH, A.; OGUNFUNMI, T. An overview of variational autoencoders for source separation, finance, and bio-signal applications. *Entropy*, v. 24, n. 1, p. 55, 2022. Disponível em: <https://doi.org/10.3390/e24010055>. Citado na página 23.

SOUSA, R. *Variational Autoencoders (VAEs): Definição, redução de dimensionalidade, espaço latente e regularização*. 2022. Available at: <https://www.deeplearningbook.com.br/variational-autoencoders-vaes-definicao-reducao-de-dimensionalidade-espaco-latente-e-regularizacao/>. Citado na página 22.

TINGLE, B. I. et al. ZINC-22A Free Multi-Billion-Scale Database of Tangible Compounds for Ligand Discovery. *Journal of Chemical Information and Modeling*, American Chemical Society, v. 63, n. 4, p. 1166–1176, 2023. ISSN 1549-9596. Disponível em: <https://doi.org/10.1021/acs.jcim.2c01253>. Citado na página 30.

WAIKHOM, L.; PATGIRI, R. Graph neural networks: Methods, applications, and opportunities. *arXiv preprint*, 2021. Citado na página 24.

WANG, Q. et al. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, Springer, v. 9, p. 187–212, 2022. Citado na página 33.

WU, Y. chen; FENG, J. wen. Development and application of artificial neural network. *Wireless Personal Communications*, Springer, v. 102, p. 1645–1656, 2018. Citado na página 20.

WU, Z. et al. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, v. 32, n. 1, p. 4–24, 2021. Citado 2 vezes nas páginas 23 e 31.

XU, Y. et al. Deep learning for molecular generation. *Future Medicinal Chemistry*, v. 11, n. 6, p. 567–597, 2019. Disponível em: <https://doi.org/10.4155/fmc-2018-0358>. Citado na página 17.

ZDRAZIL, B. et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, v. 52, n. D1, p. D1180–D1192, January 2024. Disponível em: <https://doi.org/10.1093/nar/gkad1004>. Citado na página 30.

ZHANG, Z. et al. Gans for molecule generation in drug design and discovery. In: RAZAVI-FAR, R. et al. (Ed.). *Generative Adversarial Learning: Architectures and Applications*. Cham: Springer, 2022, (Intelligent Systems Reference Library, v. 217). Disponível em: <https://doi.org/10.1007/978-3-030-91390-8_11>. Citado 2 vezes nas páginas 17 e 27.