Universidade Federal de São Carlos
Departamente de Computação

# NOVOS MÉTODOS PARA DETERMINAR A QUALIDADE DA REPRESENTAÇÃO QUIRAL EM MODELOS BASEADOS EM GNN

**Iago Elias de Faria Barbosa**

**Trabalho de Conclusão de Curso**

Universidade Federal de São Carlos
Departamento de Computação

# NOVOS MÉTODOS PARA DETERMINAR A QUALIDADE DA REPRESENTAÇÃO QUIRAL EM MODELOS BASEADOS EM GNN

**Iago Elias de Faria Barbosa**
**Orientador(a): Prof. Dr. Ricardo Cerri**

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Engenharia da Computação.

**São Carlos**
**Setembro de 2024**

# Agradecimentos

Gostaria de expressar minha profunda gratidão aos amigos que convivem comigo e que acompanharam todo o processo de desenvolvimento deste trabalho, Vinícius, Lucas, Gabriel, Jonathan, Matheus e Natanael, sem vocês esse processo teria sido muito mais pesado e maçante. Para minhas grandes amigas que mesmo morando longe continuam sendo grandes confidentes e pilares da minha vida, Ana e Luana. Para minha mãe eu devo o mundo, pela preocupação, pelo cuidado, pelo apoio e por me fazer o filho mais querido do mundo, para o meu pai e melhor amigo eu agradeço pelas conversas e pelas dicas, você é minha referência no mundo. Agradeço a todos os meus grandes amigos do trabalho que aliviaram minha vida e se sensibilizaram pela importância de escrita deste trabalho, em especial Matheus, Gustavo e Leonardo. E especial gostaria de agradecer ao meu irmão Igor, você é muito fera e o mundo é seu!

*"Há aqueles que crêem que o destino*
*descansa nos joelhos dos deuses, mas a verdade é que trabalha,*
*como um desafio candente, sobre as consciências dos homens"*
(Eduardo Galeano)

# Resumo

Como as redes neurais de grafos, especialmente modelos como SphereNet e ChIRo, incorporam a percepção de quiralidade molecular? O presente trabalho teve como objetivo analisar e avaliar a capacidade das redes neurais de grafos em incorporar a percepção de quiralidade molecular, focando em modelos específicos como SphereNet e ChIRo. Hipotetizamos que as redes neurais de grafos atuais, como SphereNet e ChIRo, apresentam limitações na incorporação completa da percepção de quiralidade molecular, especialmente em casos complexos e contínuos. Para testar essa concepção, realizamos testes e datasets que reflitam de maneira mais abrangente a incorporação quiral, incluindo (I) classificação RSA como alternativa à classificação RS tradicional; (II) classificação de quiralidades complexas; (III) criação e uso de um dataset de quiralidade contínua (CCM). Conclui-se que a classificação RSA demonstrou potencial para melhorar a acurácia dos modelos, bem como o dataset CCM revelou a importância da geometria como fator determinante na classificação quiral. Assim, foi possível identificar limitações nos modelos atuais para classificar quiralidades complexas. Cientificamente, o trabalho aponta para a necessidade de desenvolvimento de novas arquiteturas de redes neurais, que melhor incorporem as propriedades da quiralidade molecular. Socialmente, há a potencial melhoria na previsão e design de moléculas quirais, impactando áreas como desenvolvimento de fármacos e da ciência de materiais.

**Palavras-chave**: *redes neurais de grafo (GNN), quiralidade, medidas quirais contínuas (CCM), química computacional.*

# Abstract

How do graph neural networks, especially models such as SphereNet and ChIRo, incorporate the perception of molecular chirality? This study aimed to analyze and evaluate the capacity of graph neural networks to incorporate the perception of molecular chirality, focusing on specific models such as SphereNet and ChIRo. We hypothesized that current graph neural networks, including SphereNet and ChIRo, exhibit limitations in fully incorporating the perception of molecular chirality, particularly in complex and continuous cases. To test this concept, we conducted tests and developed datasets that more comprehensively reflect chiral incorporation, including (I) RSA classification as an alternative to traditional RS classification; (II) classification of complex chiralities; and (III) creation and use of a continuous chirality dataset (CCM). We conclude that RSA classification demonstrated potential to improve model accuracy, while the CCM dataset revealed the importance of geometry as a determining factor in chiral classification. Furthermore, we identified limitations in current models for classifying complex chiralities. Scientifically, this work points to the need for developing new neural network architectures that better incorporate the properties of molecular chirality. Socially, there is potential for improved prediction and design of chiral molecules, impacting areas such as drug development and materials science.

**Keywords**: *graph neural networks (GNN), chirality, continuous chirality measure (CCM), computational chemistry.*

# Lista de Figuras


2.1 Molécula do metanol com átomos enumerados. . . . . 25
2.2 Taxonomia dos estereoisômeros . . . . . 30
2.3 Propadieno . . . . . 32
2.4 Visualização lateral do aleno . . . . . 33
2.5 Enantiômero R de um aleno . . . . . 33
2.6 Exemplo de potencial molecular planar . . . . . 34
2.7 R-Ciclofano . . . . . 35

3.1 Distribuição do número de estereoisômeros por número de conformômero no dataset RS. . . . . 51

4.1 Distribuição do número de conformômeros por número de átomos no dataset RSA. . . . . 58
4.2 Distribuição de densidade CCM por categoria . . . . . 60
4.3 Distribuição de densidade logarítmica CCM por categoria . . . . . 61
4.4 Loss do dataset de treinamento por epoch . . . . . 62
4.5 Loss do dataset de validação por epoch . . . . . 62
4.6 Loss de treinamento do SphereNet . . . . . 63
4.7 Loss de validação do SphereNet . . . . . 64

# Lista de Tabelas


2.1 Matriz de adjacência do metanol . . . . . 25
2.2 Matriz de atributos de nó do metanol . . . . . 25

3.1 Balanço de rótulos R/S no dataset R/S . . . . . 51
3.2 Balanço de rótulos no dataset CHIRAL . . . . . 52

4.1 Balanço de rótulos no dataset RSA . . . . . 58
4.2 Divisão do dataset RSA entre treinamento, validação e teste . . . . . 58
4.3 Balanço de rótulos no dataset CCM . . . . . 59
4.4 Divisão do dataset CCM entre treinamento, validação e teste . . . . . 59
4.5 Resultados dos testes de normalidade sobre o dataset . . . . . 60
4.6 Acurácia do ChIRo na classificação RSA . . . . . 63
4.7 Acurácia do ChIRo na classificação RSA . . . . . 64
4.8 Acurácia do *Spherenet* na classificação RSA . . . . . 65
4.9 Acurácia dos modelos na classificação RS . . . . . 65

# Sumário


**1 Introdução** **19**

1.1 Problema Norteador . . . . . 20

1.2 Objetivos . . . . . 21

1.2.1 Objetivo Geral . . . . . 21

1.2.2 Objetivos Específicos . . . . . 21

**2 Referencial Teórico** **23**

2.1 Representação de um grafo . . . . . 24

2.2 Redes Neurais de Grafos (GNNs) . . . . . 26

2.2.1 GCN . . . . . 27

2.2.2 Ponto de vista da passagem de mensagens . . . . . 28

2.3 Esteoquímica e quiralidade molecular . . . . . 29

2.3.1 Contexto histórico . . . . . 29

2.3.2 Nomenclatura e definições da estereoquímica . . . . . 30

2.3.3 Limitações do modelo geométrico . . . . . 35

2.4 Teoria dos Grupos . . . . . 36

2.4.1 Definição de Grupo . . . . . 36

2.4.2 Simetria e Grupos de Simetria . . . . . 37

2.4.3 Representações de Grupos e Quiralidade . . . . . 38

2.4.4 Medidas Contínuas de Simetria e Quiralidade . . . . . 39

2.5 Modelagem de moléculas computacionalmente . . . . . 39

2.5.1 *Simplified Molecular Input Line Entry System* (SMILES) . . . . . 40

2.5.2 Representação da Quiralidade no SMILES . . . . . 40

2.6 Possíveis soluções para a quiralidade molecular . . . . . 41

2.6.1 *SphereNet* . . . . . 41

2.6.2 *Chiral InterRoto-Invariant Neural Network* ChIRo . . . . . 43

2.7 Biblioteca e métodos de modelagem molecular . . . 43
2.7.1 RDKit . . . 43
2.7.2 *Experimental-Torsion-Knowledge Distance Geometry* (ETKDG) . . . 44
2.7.3 Otimizador *Universal Force Field* (UFF) . . . 44
2.8 Análise Exploratória de Dados . . . 45
2.9 Validação de Normalidade . . . 45
2.9.1 Análise de *Outliers* . . . 46
2.9.2 Transformação de Dados . . . 46
2.10 Elementos de Machine Learning . . . 46
2.10.1 Funções de Ativação . . . 47
2.10.2 Função de Perda *(Loss Function)* . . . 47
2.10.3 Treinamento e Validação . . . 48

**3 Desenvolvimento** **49**
3.1 Desenvolvimento do dataset de três categorias (aquiral, R ou S) . . . 50
3.2 Desenvolvimento do dataset de quiralidades complexas . . . 51
3.3 Desenvolvimento do dataset do indicadores quirais contínuos . . . 52
3.4 Teste dos modelos na tarefa de distinguirem quiralidade (RSA) . . . 53
3.4.1 ChIRo *(Chiral InterRoto-Invariant Neural Network)* . . . 53
3.4.2 *SphereNet* . . . 54

**4 Análise Experimental** **57**
4.1 Discussão dos Resultados . . . 57
4.1.1 Dataset de três categorias (aquiral, R ou S) . . . 57
4.1.2 Dataset de indicadores quirais contínuos . . . 59
4.1.3 Teste dos modelos na tarefa de distinguir quiralidade e não-quiralidade (RSA) . . . 61

**5 Conclusão e trabalhos futuros** **67**

**Referências Bibliográficas** **69**

**A Código de obtenção do dataset aquiral** **73**

**B Código de para a filtragem de molecúlas sem angulos dihedrais** **75**

# Capítulo 1

# Introdução

Redes Neurais de Grafos (GNNs) emergiram como uma técnica poderosa para representar e processar dados estruturados em grafos, encontrando ampla aplicação em áreas como descoberta de antibióticos, simulações físicas, detecção de fake news e predição de propriedades moleculares. Sua versatilidade e capacidade de capturar relações complexas entre entidades tornaram-nas particularmente úteis no campo da química computacional, onde moléculas podem ser naturalmente representadas como grafos.

No entanto, um dos principais desafios enfrentados pelas GNNs é a representação precisa de propriedades estereoquímicas, especialmente a quiralidade molecular. A quiralidade, uma propriedade fundamental em química, refere-se à existência de moléculas que são imagens especulares uma da outra, mas não sobreponíveis. Esta característica tem implicações em várias áreas, notadamente na farmacologia, onde enantiômeros podem ter efeitos drasticamente diferentes no organismo humano.

Apesar dos avanços recentes, as GNNs ainda enfrentam dificuldades na ordem da representação e previsão acurada da quiralidade molecular. O problema central reside, portanto, na limitação das GNNs em captar as sutis diferenças tridimensionais próprias da quiralidade, uma vez que, de modo geral, tratam as moléculas como grafos bidimensionais. Diante deste cenário, surge a questão: Como desenvolver métodos mais eficazes para avaliar e melhorar a capacidade das Redes Neurais de Grafos em representar e prever a quiralidade molecular?

Algumas hipóteses para abordar esta problemática incluem: (1) A incorporação de simetrias e equivariâncias especiais nas GNNs pode melhorar sua capacidade de representar quiralidade; (2) A utilização de datasets mais complexos e diversificados, incluindo diferentes tipos de quiralidade e medidas contínuas, pode fornecer uma avaliação mais

precisa das capacidades das GNNs; (3) A combinação de abordagens geométricas com considerações de natureza quântica pode ser necessária para capturar completamente as características da quiralidade molecular.

O objetivo geral deste trabalho é analisar e propor novos datasets e métodos que permitam uma avaliação mais aprofundada da capacidade das GNNs em representar e compreender a quiralidade molecular. Especificamente, pretende-se: (1) Desenvolver um dataset de três categorias (aquiral, R ou S); (2) Criar um dataset de quiralidades complexas; (3) Implementar um dataset com um indicador contínuo de quiralidade; e (4) Analisar a capacidade dos modelos em distinguir diferentes formas de quiralidade.

A relevância deste trabalho reside na sua potencial contribuição para o avanço das GNNs no campo da química computacional. Ao propor métodos mais robustos para avaliar e melhorar a representação da quiralidade, este estudo pode trazer benefícios para diversas áreas, como o desenvolvimento de fármacos, onde a compreensão precisa da quiralidade auxilia na previsão da eficácia e segurança de novas moléculas. Além disso, os insights gerados podem abrir novas direções para o desenvolvimento de modelos de aprendizado de máquina mais sofisticados em química e ciência dos materiais. A metodologia deste trabalho envolverá o desenvolvimento e análise de novos datasets, incluindo moléculas com diferentes tipos de quiralidade e medidas contínuas de quiralidade. Serão realizados testes comparativos entre diferentes arquiteturas de GNNs, com foco em sua capacidade de representar e prever propriedades quirais. Análises exploratórias serão conduzidas para identificar os fatores determinantes na distinção da quiralidade pelos modelos, utilizando técnicas de visualização de dados e interpretação de modelos de aprendizado de máquina.

## 1.1 Problema Norteador

A dificuldade em desenvolver testes eficazes para avaliar a capacidade das Redes Neurais de Grafos (GNNs) em representar e prever a quiralidade molecular. Os testes atuais, como a classificação de moléculas em R ou S, são insuficientes para capturar a complexidade da quiralidade, necessitando de abordagens mais sofisticadas para avaliar e explicar as limitações desses modelos.

## 1.2 Objetivos

### 1.2.1 Objetivo Geral

Analisar (e propor) novos datasets e métodos que permitam uma avaliação mais aprofundada da capacidade das Redes Neurais de Grafos (GNNs) em representar e compreender a quiralidade molecular, buscando superar limitações dos métodos de teste atuais.

### 1.2.2 Objetivos Específicos

1. Desenvolver um dataset de três categorias (aquiral, R ou S): Criar um conjunto de dados que permita avaliar a capacidade das Redes Neurais de Grafos (GNNs) em classificar moléculas nas categorias aquiral, R ou S. Esse dataset proporcionará uma base para a análise da precisão dos modelos em tarefas de classificação quiral. Além disso, esse trabalho também passa por verificar se os modelos não apresentam vieses em relação às moléculas quirais, assegurando que a habilidade de percepção quiral adquirida pelos modelos não resulte em um tratamento inadequado de moléculas aquirais.

2. Desenvolver um dataset de quiralidades complexas: Criar conjuntos de dados que incorporem quiralidades mais complexas, como as quiralidades planar, axial e helicoidal. Esses dados permitirão uma avaliação da capacidade das GNNs em lidar com formas de quiralidade que transcendem a classificação binária tradicional (R/S). O propósito é examinar se os modelos tendem a se concentrar exclusivamente nos centros quirais como fator determinante, o que seria indesejável, exceto nos casos em que os modelos são explicitamente treinados com anotações referentes à quiralidade dos nós.

3. Desenvolver e analisar um dataset com um indicador contínuo de quiralidade: Implementar conjuntos de dados baseados em medidas contínuas de quiralidade, como a Continuous Chirality Measure (CCM) e outras métricas quirais. A meta é avaliar a capacidade dos modelos de capturar nuances contínuas da quiralidade, em vez de se limitarem a uma abordagem binária. A análise buscará identificar se, ao tratar moléculas com diferentes graus de torção molecular de forma homogênea, os modelos perdem informações cruciais sobre a quiralidade. Moleculas com máxima

quiralidade, como discutido por Vavilin e Fernandez-Corbaton (2022), podem apresentar características distintas de moléculas com baixa quiralidade, o que motiva uma análise exploratória para verificar se os indicadores quirais são efetivamente determinantes na classificação de quiralidade.

4. Analisar a capacidade dos modelos de distinguirem quiralidade: Avaliar a eficácia das GNNs em diferenciar as diversas formas de quiralidade, bem como sua habilidade de distinguir entre moléculas quirais e aquirais. Esta análise será realizada utilizando os datasets desenvolvidos, com o intuito de identificar as limitações e os pontos fortes dos modelos na tarefa de classificação quiral.

# Capítulo 2

# Referencial Teórico

Com o objetivo prover um panorama dos conceitos utilizados no desenvolvimento, nesse capítulo são apresentados fundamentos da base teórica utilizados posteriormente.

Conforme supracitado, as Redes Neurais de Grafos (GNNs) surgiram como uma técnica para representar e processar dados estruturados em forma de grafos, permitindo a aplicação em diversas áreas.

Introduzidas por Scarselli et al. (2009), as GNNs possibilitam a criação de representações detalhadas de grafos, capturando informações de nós, arestas e subestruturas, facilitando sua integração em modelos preditivos para uma ampla gama de aplicações.

No campo molecular, a representação de moléculas como grafos é especialmente eficaz, dada a natureza intrinsecamente conectada das moléculas. Diversos estudos têm demonstrado o sucesso das GNNs na predição de propriedades moleculares, como posição atômica, cargas parciais, solubilidade, ponto de ebulição e afinidade de ligação.

No entanto, um dos maiores desafios enfrentados por essas redes é a correta representação das propriedades estereoquímicas, especialmente a quiralidade – uma característica fundamental que descreve como certas moléculas possuem versões espelhadas não superponíveis, os chamados enantiômeros.

A quiralidade desempenha um papel particularmente relevante na farmacologia, onde diferentes enantiômeros de uma mesma molécula podem ter efeitos opostos no organismo. Exemplos como a talidomida, onde um enantiômero tem efeito anti-inflamatório, enquanto outro causa malformações, ilustram a importância de uma previsão precisa da quiralidade molecular. Entretanto, as GNNs, em suas formas mais simples, tratam moléculas como grafos bidimensionais, não capturando adequadamente as sutis diferenças tridimensionais cruciais para distinguir entre enantiômeros.

Atualmente, diversas abordagens tentam solucionar essa limitação, como a incorporação de simetrias geométricas nas redes. Modelos como SchNet e DimeNet focam em simetrias Euclidianas (E(3)), enquanto modelos como SphereNet e ChIRo incorporam simetrias mais avançadas, como a sensibilidade a reflexões e rotações, essenciais para a representação quiral. Outras abordagens incluem o uso de tags explícitos para identificar centros quirais ou redes que são sensíveis à ordem das ligações.

Apesar de progressos, essas soluções ainda não capturam completamente as complexidades da quiralidade, especialmente em tarefas que envolvem interações com luz polarizada ou afinidade com aminoácidos quirais. Para melhorar a representação da quiralidade em GNNs, é necessário explorar métodos que considerem não apenas simetrias geométricas, mas também simetrias de natureza quântica, abrindo caminho para soluções mais robustas e holísticas.

## 2.1 Representação de um grafo

Utilizando a notação apresentada por White (2022), um grafo $G$ é um conjunto de nós $V$ e arestas $E$, que descrevem as conexões entre os vértices.

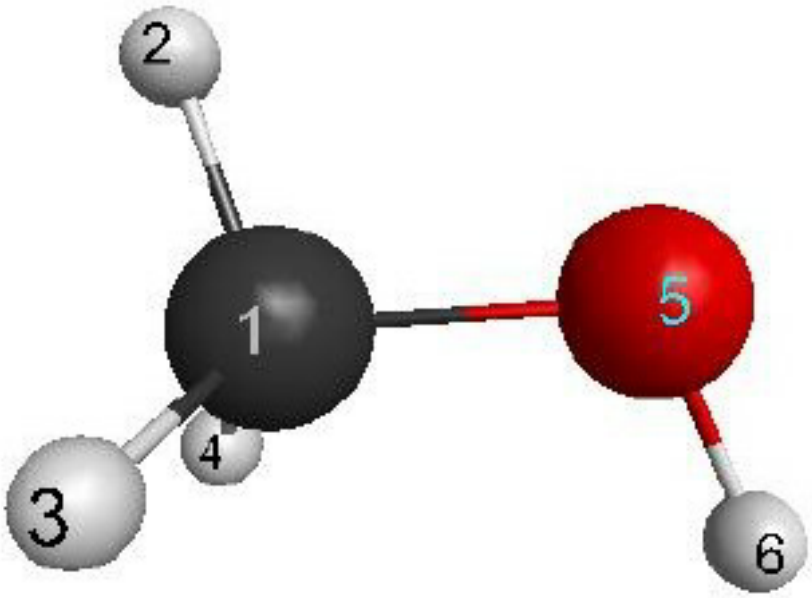
Cada nó $i$ no grafo é representado por um vetor $\vec{v}_i.$, que pode conter diversas informações, como propriedades atômicas, no caso de moléculas. As arestas são geralmente descritas por uma matriz de adjacência $E$ na qual $e_{ij} = 1$ indica que os nós $i$ e $j$ estão conectados por uma aresta.

No contexto molecular, os grafos são simétricos ($e_{ij} = e_{ji}$, já que as ligações entre átomos não possuem direção.

A representação de moléculas como grafos permite que a estrutura atômica e as conexões entre os átomos sejam remodeladas de maneira natural, com vértices representando os átomos e as arestas representando as ligações. Um exemplo clássico dessa abordagem pode ser visto a partir da molécula do metanol (Figura 2.1). Na figura a seguir, a molécula é representada com os seus átomos numerados, e a matriz de adjacência correspondente, mostrada na Tabela 2.2, descreve como os átomos estão conectados.

Além da matriz de adjacência, é possível associar a cada nó uma matriz de atributos, que descreve propriedades específicas de cada átomo. Por exemplo, pode-se utilizar uma codificação *one-hot*, que permite identificar o tipo de átomo (carbono, hidrogênio ou oxigênio).

Figura 2.1: Molécula do metanol com átomos enumerados.




Fonte: George Pitsevich (2012)

Tabela 2.1: Matriz de adjacência do metanol

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 1 | 0 | 0 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 |
| 3 | 1 | 0 | 1 | 0 | 0 | 0 |
| 4 | 1 | 0 | 0 | 1 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 1 | 1 |
| 6 | 0 | 0 | 0 | 0 | 1 | 1 |

Tabela 2.2: Matriz de atributos de nó do metanol

| | C | H | O |
|---|---|---|---|
| 1 | 1 | 0 | 0 |
| 2 | 0 | 1 | 0 |
| 3 | 0 | 1 | 0 |
| 4 | 0 | 1 | 0 |
| 5 | 0 | 0 | 1 |
| 6 | 0 | 1 | 0 |

Conforme visto, esse processo permite a construção de representações ricas em informações, que podem ser utilizadas em modelos preditivos para diversas aplicações, como predição de propriedades moleculares.

A principal vantagem dessa abordagem é que ela captura as interações locais entre os átomos de uma maneira que pode ser facilmente integrada a modelos de aprendizado de máquina, especialmente as Redes Neurais de Grafos, que veremos a seguir, e que são particularmente adequadas para dados com essa estrutura.

## 2.2 Redes Neurais de Grafos (GNNs)

Nesse sentido, conforme White (2022) *Graph Neural Networks* (GNNs) são uma classe de modelos de aprendizado profundo desenvolvidos para lidar com dados estruturados em grafos. Esses modelos se destacam por sua capacidade de capturar informações sobre as conexões entre nós e as suas características, permitindo a aplicação em diversos domínios.

Nas GNNs, a informação flui entre os nós por meio de um processo conhecido como passagem de mensagem (*message passing*), onde cada nó recebe informações de seus vizinhos conectados. Esse processo permite que as GNNs atualizem as representações dos nós com base nas características dos vizinhos, mantendo a estrutura do grafo intacta. A formulação mais comum desse processo foi introduzida por Kipf e Welling (2017), com o modelo Graph Convolutional Network (GCN). Nesse modelo, a informação sobre as conexões entre os nós é agregada e processada de forma eficiente.

A característica central das GNNs é sua equivariança a permutações. Isso significa que, ao reorganizar a ordem dos nós ou das arestas, o modelo manterá sua capacidade de representar corretamente as relações, e também as estruturas internas do grafo. Ao lidar com moléculas, a equivariança ganha protagonismo, pois trata-se de um espaço onde a ordem dos átomos pode ser permutada sem alterar sua estrutura física ou química.

Por isso, as GNNs são amplamente utilizadas no contexto molecular, devido a essa habilidade de 'modelar a natureza conectada' e tridimensional das moléculas. Cada átomo e ligação pode ser representado em nós e arestas, com atributos como suas propriedades atômicas, tipos de ligação e até mesmo coordenadas especiais, permitindo que o modelo capture as interações moleculares. Além disso, as GNNs são capazes de incorporar informações globais e locais, permitindo prever propriedades como a quiralidade molecular.

Porém, as GNNs tradicionais enfrentam algumas limitações acerca dessa previsão, de-

vido às propriedades tridimensionais mais sutis. A passagem de mensagem convencional é, por vezes, insuficiente para distinguir entre enantiômeros, uma vez que a disposição especial dos átomos pode ser perdida no processo de agregação de informação. Por isso, modelos mais avançados como *SphereNet* e *ChIRO* foram desenvolvidos para lidar especificamente com essa limitação, incorporando informações sobre simetrias geométricas e torções moleculares, para melhorar a precisão na previsão de propriedades quirais.

As GNNs possuem diversas implementações e variações que foram desenvolvidas para atender a diferentes necessidades, sendo a *Graph Convolutional Network* (GCN) uma das mais populares e amplamente utilizadas, se destacando pela simplicidade de sua formulação e pela eficiência na agregação de informações entre os nós de um grafo, sendo particularmente útil para tarefas de classificação. A seguir, exploraremos o funcionamento das GCNs.

### 2.2.1 GCN

Uma das primeiras implementações de GNN que ganhou ampla popularidade é a *Graph Convolutional Network* (GCN), proposta por Kipf e Welling (2017). As GCNs operam sobre a estrutura do grafo, representada pela matriz de adjacência $E$ e os atributos dos nós, codificados na matriz de características $V$. O objetivo principal da GCN é produzir uma nova matriz de características $V'$, que incorpora tanto a informação estrutural do grafo quanto as características originais dos nós. A equação principal em uma camada GCN é:

$$v_{il} = \sigma\left(\frac{1}{d_i} e_{ij} v_{jk} w_{kl}\right)$$

Aqui, $v_{il}$ representa a característica atualizada do nó $i$, $e_{ij}$ denota o elemento correspondente na matriz de adjacência, $\sigma$ é a função de ativação não-linear e $w_{kl}$ são os parâmetros aprendíveis. Essa expressão captura a essência da operação de convolução em grafos, onde cada nó agrega informações de sua vizinhança imediata.

Aqui, cada nó atualiza suas características com base nas características de seus vizinhos conectados (o que é garantido pela multiplicação por $e_{ij}$), ponderadas por uma matriz de pesos. Essa operação é intrinsecamente local, focando nas conexões entre nós diretamente adjacentes.

A equivariância permutacional surge justamente por esse motivo, a localidade, uma vez

que o cálculo de cada nó depende dos vizinhos conectados, independentemente da ordem em que os nós ou arestas são enumerados no grafo. Consequentemente, o processo de atualização de características é invariante à permutação dos índices dos nós, garantindo que a rede capture a estrutura do grafo de forma consistente, independentemente da representação gráfica escolhida.

Embora a formulação matemática das GCNs traga consigo essa base, uma interpretação igualmente elucidativa surge quando consideramos a passagem de mensagens, que abordaremos agora.

### 2.2.2 Ponto de vista da passagem de mensagens

No ponto de vista de passagem de mensagem, as Graph Convolutional Networks (GCNs) podem ser vistas como redes nas quais a informação é transmitida entre os nós de um grafo através de suas arestas. Este processo pode ser interpretado como uma troca de mensagens, pois cada nó 'dissemina suas características' para os nós adjacentes, e, concomitamente, agrega as características vizinhas em sua própria representação.

Esta abordagem reflete como as GCNs capturam a estrutura local dos grafos. Ao propagar informações ao longo das camadas da rede, as GCNs conseguem incorporar progressivamente informações de vizinhanças cada vez mais amplas, permitindo a captura de padrões estruturais em diferentes escalas.

É importante denotar que esta interpretação baseada em passagem de mensagens não se limita apenas às GCNs. De fato, ela é tão importante que uma classe mais ampla de modelos é denominada de *Message Passing Neural Networks* (MPNNs). As MPNNs generalizam o conceito de passagem de mensagens, permitindo uma variedade de esquemas de agregação e atualização, dos quais a GCN representa um caso específico.

A compreensão das Graph Convolutional Networks (GCNs) e seu funcionamento fornece uma base para abordar desafios complexos. Dentre esses, o presente trabalho, conforme supracitado, trata da representação e previsão precisa da quiralidade molecular, um conceito fundamental em estereoquímica.

A quiralidade, uma propriedade geométrica que não pode ser facilmente capturada por representações bidimensionais simples, apresenta um teste rigoroso para as capacidades das GCNs e outras arquiteturas de redes neurais em grafos. A natureza tridimensional e as sutis diferenças estruturais que caracterizam as moléculas quirais exigem que os modelos de aprendizado de máquina sejam capazes de capturar e processar informações

geométricas complexas.

Na próxima exploraremos este tema em profundidade, estruturando nossa discussão em três partes principais: (I) examinaremos o desenvolvimento histórico do conceito de quiralidade, desde suas origens até sua importância atual na química e nas ciências biomédicas; (II) introduziremos os termos e conceitos-chave necessários para discutir o tema, incluindo as convenções de nomenclatura utilizadas para descrever diferentes configurações estereoquímicas; (III) por fim, exploraremos as diversas formas de quiralidade, desde a quiralidade central clássica até formas mais complexas como quiralidade axial e planar.

Esta base em estereoquímica nos permitirá, posteriormente, analisar de forma mais informada os desafios que a quiralidade apresenta para as redes neurais de grafos e considerar abordagens para superar essas limitações.

## 2.3 Esteoquímica e quiralidade molecular

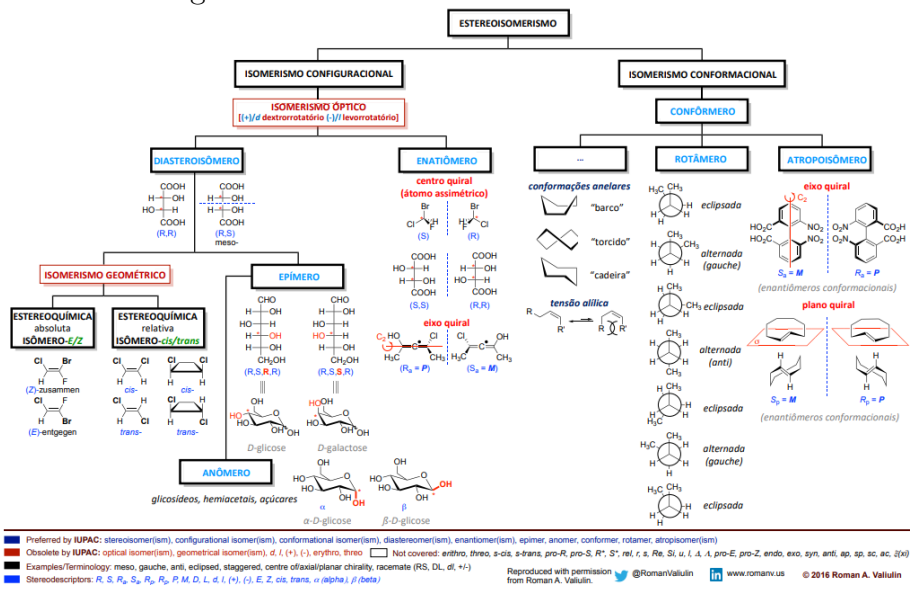
### 2.3.1 Contexto histórico

A definição da quiralidade nasce do problema observado inicialmente por Pasteur (1848), onde cristais de ácido tartárico apresentavam duas formas distintas que eram imagens especulares uma da outra, mas não sobreponíveis. Pasteur percebeu que essas duas formas de cristal causavam a rotação da luz polarizada em direções opostas. Esse fenômeno levou à compreensão de que certas moléculas possuem uma assimetria intrínseca, sendo capazes de existir em duas formas, chamadas enantiômeros, que não podem ser sobrepostas, semelhante à relação entre a mão direita e a mão esquerda.

A definição de quiralidade foi posteriormente formalizada por William Thomson, conhecido como Lord Kelvin, em 1904 em suas aulas (Knudsen, 2005). Kelvin definiu um objeto como "quiral"se ele não puder ser sobreposto à sua imagem especular, estabelecendo um conceito mais abrangente que vai além dos cristais. Essa ideia foi associada à descoberta de Hoff (1874) e Bel (1874), que propuseram, de forma independente, que a quiralidade molecular surge da presença de um átomo de carbono ligado a quatro grupos diferentes, formando um centro quiral tetrahédrico. Essas descobertas tiveram um papel fundamental na compreensão da estrutura tridimensional das moléculas e como essa estrutura influencia as propriedades químicas e biológicas dos compostos quirais.

Com a formulação do modelo clássico da quiralidade, que surgiu como uma genera-

Figura 2.2: Taxonomia dos estereoisômeros




Fonte: Roman A. Valiulin (2016)

lização geométrica baseada em observações experimentais, tornou-se essencial estabelecer uma terminologia para descrever as diversas formas de isomeria espacial, que veremos na próxima seção. Nesse sentido, a nomenclatura e as definições da estereoquímica atuam na categorização e entendimento das diferenças entre moléculas quirais, especialmente na distinção entre os diversos tipos de estereoisômeros e enantiômeros.

### 2.3.2 Nomenclatura e definições da estereoquímica

A estereoquímica é o ramo da química que estuda a disposição espacial dos átomos em moléculas e como essa disposição influencia suas propriedades e reações. Nesta, a isometria espacial ocupa um lugar especial, onde as moléculas compartilham a mesma fórmula molecular e conectividade, mas diferem na organização tridimensional dos seus átomos.

Dentre os isômeros espaciais, os mais relevantes para o estudo da quiralidade são os chamados estereoisômeros - as múltiplas formas que uma molécula definida por um mesmo grafo 3D podem se manifestar espacialmente. A figura 2.2 mostra a taxonomia dos estereoisômeros.

Estes podem ser divididos em duas classes principais, os enantiômetros e os diastereoisômeros. Para o estudo da quiralidade, os que mais nos interessam são os enantiômeros, imagens especulares não sobreponíveis uma da outra, semelhantes à relação entre as mãos direita e esquerda. Já os diastereoisômeros não são imagens especulares e podem ter

propriedades físicas e químicas significativamente diferentes entre si.

A notação para diferenciar os dois tipos é determinada pelas regras desenvolvidas por Cahn-Ingold-Prelog (Cahn *et al.*, 1966), que estabelece uma hierarquia de prioridade para os ligantes ao redor de um centro quiral, geralmente um átomo de carbono com quatro substituintes diferentes.

Com base nestas, os enantiômeros são classificados como R (*rectus*), se a disposição dos substituintes ao redor do centro quiral segue o sentido horário, ou S (*sinister*), se segue o sentido anti-horário.

Outra nomenclatura importante é a designação de compostos quirais como dextrógiros ou levógiros, que indica a direção na qual o enantiômero rotaciona a luz polarizada. Um composto dextrógiro rotaciona a luz para a direita, enquanto um composto levógiro a rotaciona para a esquerda. Essa propriedade da distinção experimental de enantiômeros é amplamente utilizada na indústria farmacêutica e em outras áreas da química.

Os enantiômeros mais comuns surgem da presença de um estereocentro, um átomo na molécula ligado a pelo menos três ligantes distintos, conferindo assimetria à molécula. No entanto, existe outra categoria de enantiômeros, conhecidos como atropoisômeros, ou enantiômeros conformacionais. Estes derivam da assimetria inerente à sua conformação tridimensional. Tal configuração estabelece dois tipos de quiralidade distintos da quiralidade central definida pela presença de um estereocentro: a quiralidade axial e a planar.

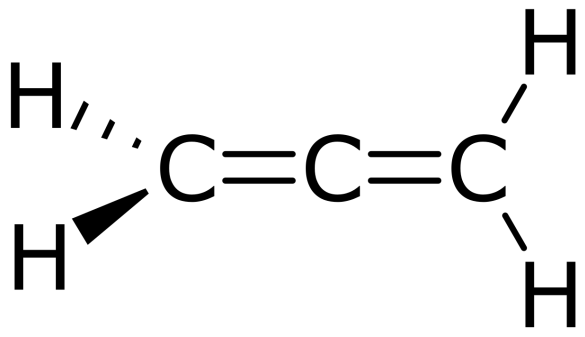
Portanto, a correta utilização da nomenclatura estereoquímica é essencial para a identificação, estudo e aplicação das propriedades quirais, uma vez que pequenas diferenças na configuração espacial de uma molécula podem resultar em efeitos biológicos e químicos completamente distintos.

A quiralidade central é a mais comum e ocorre quando um átomo, geralmente de carbono, está ligado a quatro grupos diferentes. Esse átomo, chamado de centro quiral, dá origem a duas formas não superponíveis (enantiômeros), que são imagens especulares uma da outra. Realizado este preâmbulo, podemos passar ao estudo dos tipos de quiralidade.

### Quiralidade Axial

A quiralidade axial surge em moléculas que possuem um eixo em torno do qual diferentes grupos se organizam de forma assimétrica. Isso geralmente ocorre em compostos onde dois átomos ou grupos funcionais estão conectados por uma ligação dupla ou tripla, como nos alenos e bifenilos, criando um eixo quiral. A distinção entre enantiômeros nesses

Figura 2.3: Propadieno




Fonte: Warraich Sahib (2014)

sistemas se baseia na disposição dos grupos ao redor desse eixo, resultando em quiralidade mesmo na ausência de um centro quiral tradicional.

Os alenos são um grupo de hidrocarbonetos caracterizado por um eixo de três carbonos, onde o carbono central está ligado aos outros dois por meio de ligações duplas. Além disso, quatro ligantes estão conectados aos carbonos nas extremidades desse eixo. O carbono central do aleno forma duas ligações *sigma* e duas ligações *pi*, indicando uma geometria linear para os carbonos do aleno. Este arranjo faz com que os substituintes dos carbonos terminais fiquem em planos perpendiculares.

A figura 2.3 mostra o propadieno, um exemplo de aleno.

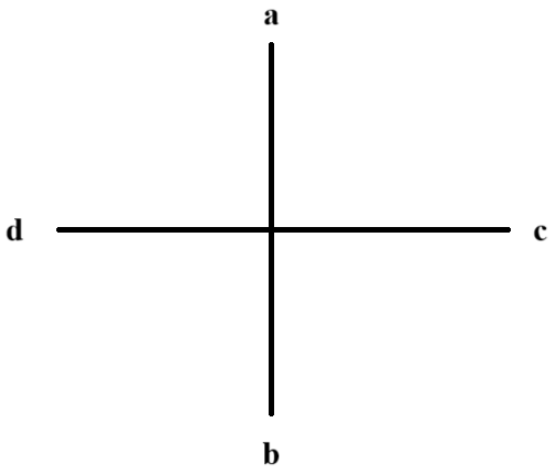
Para explicar como a quiralidade se manifesta nos alenos, é necessário compreender a geometria desse grupo. Nesse caso, os dois ligantes ligados ao carbono esquerdo e o próprio carbono formam um plano ortogonal ao plano formado pelo carbono direito e seus dois ligantes.

Considerando que o ligante superior do carbono esquerdo é chamado de "$a$", o inferior de "$b$", o ligante superior do carbono direito de "$c$"e o inferior de "$d$", e posicionando-se um ponto de observação à esquerda da molécula de forma que os carbonos das extremidades esquerda e direita se sobreponham, é possível observar que a disposição tridimensional desses ligantes cria uma quiralidade inerente à estrutura do aleno.

Essa configuração espacial faz com que as imagens especulares dos alenos não sejam sobreponíveis, caracterizando sua propriedade quiral. Portanto, a geometria particular dos alenos, com seus carbonos centrais ligados por duplas ligações e os substituintes em planos perpendiculares, é responsável pela manifestação da quiralidade nessa classe de compostos.

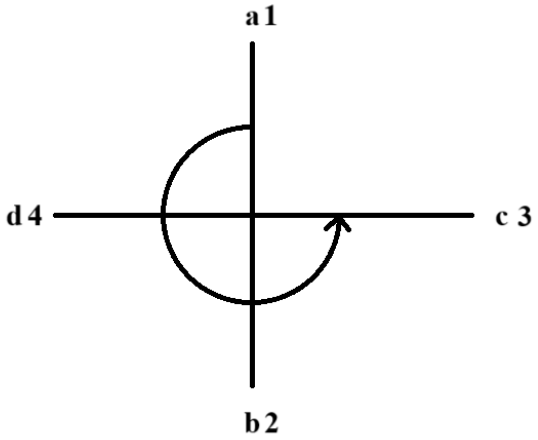
Assim, obtemos a visualização vista na figura 2.3.2.

Figura 2.4: Visualização lateral do aleno




Fonte: Autor (2024)

Figura 2.5: Enantiômero R de um aleno



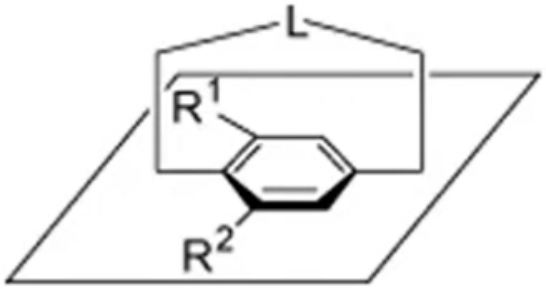

Fonte: Autor (2024)

A assimetria desse grupo, e consequentemente a sua quiralidade e conferência de ativação ótica acontece somente na situação onde o ligando $a$ é diferente do ligando $b$ e o ligando $c$ é diferente do ligando $d$. Neste caso, utilizamos a regra de prioridade de Cahn-Ingold-Prelog (Cahn *et al.*, 1966) para assinalar os índices, primeiro de um lado (escolhido de maneira arbitrária) e posteriormente de outro.

Com esse procedimento, é verificado o sentido da rotação 1-2-3. Caso seja horário, se trata do enantiômero R. Caso contrário, do enantiômero S.

Na figura 2.3.2 simulamos um enantiômero R de um aleno. (Favre e Powell, 2014)

Figura 2.6: Exemplo de potencial molecular planar




Fonte: Dr. Ani Deepthi (2021)

**Quiralidade Planar**

A quiralidade planar ocorre quando um plano em uma molécula é assimétrico em relação aos seus ligantes ou grupos funcionais. Esse tipo de quiralidade é comum em compostos organometálicos e em sistemas moleculares com estruturas em anel ou planos rígidos.

A figura 2.6 exemplifica este fenômeno, onde caso os ligandos $R_1$ $R_2$ sejam diferentes ou um deles exista o plano definido pelo ligando $L$ deixa de ser perpendicular ao plano, adquirindo uma torção que lhe confere a propriedade planar.

Para descobrir qual a categoria do enantiômero em uma quiralidade planar, é necessário primeiramente selecionar um átomo pivô. Este deve estar conectado ao plano quiral (mas sem pertencer ao mesmo) e também a algum plano perpendicular de molécula.

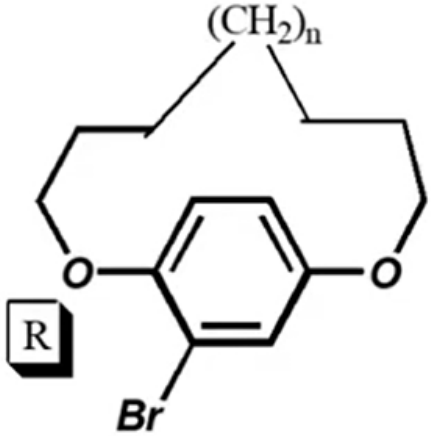
Dos átomos possíveis, o átomo escolhido deve ser mais próximo ao ligando do plano quiral com maior prioridade, de acordo com a regra de Cahn-Ingold-Prelog (Cahn *et al.*, 1966).

Em seguida, deve ser observado se a direção de rotação do átomo pivô ao ligando de maior quiralidade é horária, categorizando-a entre $R$ e $S$.

Esse processo pode ser visto no R-ciclofano da figura 2.7, onde o átomo pivô é o oxigênio esquerdo e o ligando de maior prioridade é o bromo.

Embora os diferentes tipos de quiralidade ofereçam uma compreensão das formas com que esta pode se manifestar nas moléculas, as abordagens geométricas tradicionais utilizadas para representá-las apresentam algumas limitações. Em muitos casos, essas definições baseadas na geometria espacial são insuficientes para capturar a complexidade e as nuances presentes em sistemas moleculares reais, especialmente aqueles com alta flexibilidade conformacional. A seguir, discutiremos as limitações inerentes ao modelo geométrico e a

Figura 2.7: R-Ciclofano




Fonte: Dr. Ani Deepthi (2021)

necessidade de abordagens mais robustas para descrever e quantificar a quiralidade em sistemas dinâmicos.

### 2.3.3 Limitações do modelo geométrico

Embora a definição geométrica de quiralidade seja amplamente utilizada, ela pode ser limitada quando aplicada a sistemas reais.

Tradicionalmente, a quiralidade é tratada como uma característica binária: um objeto é considerado quiral se não puder ser superposto à sua imagem especular, e aquiral caso contrário. No entanto, essa definição simples, baseada na ausência de um centro ou plano de simetria, não captura completamente a complexidade da quiralidade em sistemas dinâmicos e flexíveis.

Como discutido por Vavilin e Fernandez-Corbaton (2022), a limitação principal da definição geométrica reside no fato de que ela trata a quiralidade como uma propriedade que simplesmente existe ou não, sem considerar um espectro ou escala para sua presença.

A busca atual na ciência e na tecnologia vai além de simplesmente determinar se um objeto é quiral, mas sim de quantificá-la. Para isso, em vez de se basear apenas na geometria, há um esforço para definir quiralidade em termos de interações com campos eletromagnéticos de diferentes helicidades, o que permite medir o grau de quiralidade de forma mais precisa e em um contexto físico mais específico.

Essa abordagem reconhece que, em sistemas complexos, a quiralidade pode variar em intensidade e não apenas na presença ou ausência, abrindo caminho para a definição de objetos maximamente quirais eletromagneticamente, por exemplo, o que não é capturado

pela abordagem geométrica tradicional.

Além disso, ao utilizar medidas contínuas de quiralidade, como o *Continuous Chirality Measure* (CCM) (Zabrodsky e Avnir, 1995), surgem problemas adicionais, especialmente ao lidar com conformômeros flexíveis.

Conformômeros aquirais podem, paradoxalmente, exibir valores de quiralidade menores do que conformômeros quirais, devido à flexibilidade e às torções moleculares (Abraham e Nitzan, 2024).

Esse fenômeno ocorre porque as variações conformacionais podem induzir distorções que afetam a simetria percebida, resultando em uma aparente quiralidade que não é representativa da verdadeira natureza quiral ou aquiral da molécula. Isso evidencia ainda mais as limitações da abordagem geométrica e a necessidade de métodos mais robustos para quantificar a quiralidade.

Nesse contexto, a Teoria dos Grupos surge como uma ferramenta matemática poderosa para lidar com simetrias, trazendo os fundamentos necessários para descrever não apenas a presença ou ausência de quiralidade, mas também como diferentes simetrias se manifestam em sistemas moleculares, permitindo entender a quiralidade de maneira quantitativa e formal. A seguir, exploraremos os princípios da Teoria dos Grupos e seu papel na modelagem de simetrias moleculares e quirais.

## 2.4 Teoria dos Grupos

A teoria dos grupos, proposta por Klein (1893), sugere um tipo de estrutura algebraica abstrata, de maneira análoga a como espaços vetoriais descrevem operações genéricas para vários tipo de algébra. Nela, os grupos fornecem um meio de trabalhar a ideia de simetria para todas as álgebras que se encaixam em suas regras.

### 2.4.1 Definição de Grupo

Na matemática, um grupo é uma estrutura algébrica formada por um conjunto $G$ e uma operação binária $*$ que combina dois elementos do conjunto para formar um terceiro elemento, também pertencente a $G$. Para que $(G, *)$ seja considerado um grupo, ele deve satisfazer as seguintes propriedades fundamentais:

1. Fechamento: Para todos $a, b \in G$, o resultado da operação $a * b$ também pertence

a $G$.

$$a * b \in G$$

2. Associatividade: A operação $*$ é associativa, ou seja, para todos $a, b, c \in G$, temos:

$$(a * b) * c = a * (b * c)$$

3. Elemento Neutro: Existe um elemento $e \in G$ tal que para todo $a \in G$, a operação com $e$ deixa $a$ inalterado:

$$e * a = a * e = a$$

Este elemento $e$ é chamado de identidade do grupo.

4. Elemento Inverso: Para cada elemento $a \in G$, existe um elemento $a^{-1} \in G$ tal que:

$$a * a^{-1} = a^{-1} * a = e$$

Um exemplo simples é o grupo dos números inteiros $\mathbb{Z}$ com a operação de adição $+$, onde o elemento neutro é 0 e o inverso de qualquer número $n$ é $-n$.

### 2.4.2 Simetria e Grupos de Simetria

A teoria dos grupos é uma ferramenta matemática amplamente usada para descrever as simetrias de objetos, em particular de moléculas.

Parafraseando Weyl (1952), dada uma configuração espacial $\mathcal{F}$, os automorfismos do espaço que deixam $\mathcal{F}$ inalterada formam um grupo $\Gamma$, que descrevem precisamente uma simetria do grupo $\mathcal{F}$. Um automorfismo é uma transformação que preserva as distâncias (isometria), como rotações e reflexões.

Matematicamente, considere um objeto $O$ em um espaço Euclidiano tridimensional $\mathbb{R}^3$. Uma transformação $T : \mathbb{R}^3 \to \mathbb{R}^3$ é uma simetria de $O$ se, para todos os pontos $p$ de $O$, temos $T(p) \in O$. O conjunto de todas as transformações que satisfazem essa condição forma o grupo de simetria $\Gamma$ de $O$.

Compreender as simetrias moleculares por meio da Teoria dos Grupos é fundamental para descrever como as moléculas quirais diferem das aquirais. No entanto, é necessário avançar para o conceito de representações de grupos, que permite expressar matematicamente como essas simetrias atuam sobre os diferentes elementos de uma molécula.

A seguir, discutiremos como as representações de grupos contribuem para a análise da quiralidade em moléculas.

### 2.4.3 Representações de Grupos e Quiralidade

Uma das aplicações fundamentais da teoria dos grupos é na modelagem da quiralidade molecular.

Em um contexto de grupo de simetria, a quiralidade surge quando o grupo $\Gamma$ de uma molécula não contém operações que possam mapear a molécula em sua imagem especular.

Por exemplo, uma molécula com um centro quiral tetrahédrico pode ser representada por um grupo $C_1$ (grupo trivial, que consiste apenas na identidade), indicando que a molécula não possui simetrias, e, portanto, é quiral. Moléculas que pertencem a grupos mais complexos, como $C_n$, ainda podem ser quirais se não houver operações de simetria que façam a superposição com sua imagem especular.

Na literatura de redes neurais de grafos, comumente são referenciados dois grupos de simetria, $E(3)$ e $SE(3)$, eles são grupos fundamentais na descrição de simetrias e transformações no espaço tridimensional.

O Grupo de Euler $E(3)$ é o grupo das isometrias no espaço tridimensional $\mathbb{R}^3$, ou seja, o conjunto de todas as transformações que preservam as distâncias. Esse grupo inclui rotações, translações e reflexões. Matematicamente, pode ser descrito como o produto semidireto $E(3) = O(3) \ltimes \mathbb{R}^3$, onde $O(3)$ é o grupo ortogonal que inclui todas as rotações e reflexões em três dimensões.

Já o Grupo Especial de Euler $SE(3)$ é um subgrupo de $E(3)$ que consiste apenas de rotações e translações, excluindo reflexões. Este grupo é usado para descrever movimentos rígidos no espaço tridimensional.

O grupo $SE(3)$ é dado por $SE(3) = SO(3) \ltimes \mathbb{R}^3$, onde $SO(3)$ é o grupo especial ortogonal que inclui apenas rotações, sem reflexões, o grupo $SE(3)$ é importante para a literatura de redes neurais sensíveis a quiralidade uma vez que se a GNN for equivariante ao $SE(3)$, a rede é sensível à definição espacial de quiralidade.

No entanto, em muitos casos, a quiralidade não pode ser tratada apenas como uma propriedade binária, mas como algo que varia em intensidade. Nesse sentido, as medidas contínuas de simetria e quiralidade surgem como uma ferramenta para quantificar o grau de quiralidade em diferentes sistemas moleculares.

### 2.4.4 Medidas Contínuas de Simetria e Quiralidade

Além de determinar a presença ou ausência de quiralidade, podemos quantificar o grau de quiralidade utilizando medidas contínuas, como a *Continuous Chirality Measure* ($CCM$) (Zabrodsky e Avnir, 1995).

Essa medida se baseia no conceito de sobreposição mínima entre uma molécula e sua imagem especular:

$$CCM(R) = 100 \times \min\left(1 - \frac{\langle R|S\rangle}{\langle R|R\rangle}\right)$$

Aqui, $\langle R|S\rangle$ é o produto interno entre a forma original $R$ e sua imagem especular $S$. O $CCM$ fornece uma quantificação do quão longe uma molécula está de ser aquiral: um $CCM$ de 0 indica quiralidade máxima, enquanto valores mais altos indicam aproximação à aquiralidade.

Esse formalismo matemático permite uma análise precisa das simetrias moleculares, facilitando a classificação e quantificação da quiralidade em moléculas complexas, principalmente em casos onde flexibilidade e rotação de ligações moleculares desempenham algum papel.

A análise detalhada das simetrias moleculares e a quantificação da quiralidade são facilitadas pela modelagem computacional de moléculas, uma abordagem que utiliza algoritmos e simulações para prever e visualizar a estrutura tridimensional das moléculas. Esses modelos computacionais permitem não apenas a representação precisa das simetrias e da quiralidade, mas também a análise dinâmica de como as moléculas interagem e se comportam em diferentes condições, incluindo a flexibilidade e rotação das ligações.

## 2.5 Modelagem de moléculas computacionalmente

Para problemas moleculares computacionais, é essencial estabelecer padrões e especificações que possibilitem a entrada de moleculas nos algorítmos e modelos.

Nessa seção abordamos o SMILES, representação que será utilizada no desenvolvimento do trabalho.

### 2.5.1 *Simplified Molecular Input Line Entry System* (SMILES)

Uma das formas mais comuns de representar moléculas em sistemas computacionais, o *Simplified Molecular Input Line Entry System* (SMILES) é uma notação que traduz a estrutura bidimensional de uma molécula em uma string de caracteres, que pode ser facilmente manipulada por computadores. A notação é compacta, legível por humanos e permite a representação de uma ampla variedade de estruturas químicas.

Por exemplo, a molécula de etanol, $C_2H_5OH$, pode ser representada no SMILES como "CCO". Nesta string, cada átomo de carbono é representado por "C", o átomo de oxigênio por "O", e as ligações simples são implicitamente representadas entre os átomos adjacentes.

### 2.5.2 Representação da Quiralidade no SMILES

No SMILES, a quiralidade é representada utilizando uma notação específica para indicar a configuração dos átomos em torno de um centro quiral (tipicamente um átomo de carbono). Esta notação usa os símbolos "@"e "@@"para designar a configuração quiral dos átomos. O @ indica que os átomos estão conectados em uma ordem específica que corresponde à configuração (R) ou (S) de acordo com as regras da IUPAC, e o @@ indica que a configuração dos átomos é oposta àquela especificada por "@".

Por exemplo, a molécula de ácido lático, que possui um centro quiral no carbono central, pode ser representada por duas strings SMILES diferentes, $C[C@H](O)C(=O)O$ representa a configuração (R)-ácido lático e a $C[C@@H](O)C(=O)O$ representa a configuração (S)-ácido lático.

Vemos que embora o SMILES ofereça uma forma eficaz de representar a estrutura e a quiralidade das moléculas, ele não está isento de limitações. A notação SMILES pode simplificar a complexidade da quiralidade, mas tem problemas em captar a informação tridimensional necessária para uma análise. Esse desafio é particularmente relevante quando se trabalha com redes neurais gráficas (GNNs), que frequentemente enfrentam dificuldades em interpretar e integrar a quiralidade molecular. Possíveis soluções para a quiralidade molecular envolvem o desenvolvimento de técnicas mais sofisticadas que podem melhorar a capacidade das GNNs de lidar com essas complexidades, oferecendo uma representação mais fiel da estrutura molecular e aprimorando a precisão das previsões e análises realizadas por esses modelos computacionais.

## 2.6 Possíveis soluções para a quiralidade molecular

Abaixo são brevemente explicados duas tentativas de solução do problema da insensibilidade das GNN's em relação a quiralidade.

### 2.6.1 *SphereNet*

Liu *et al.* (2022) define uma abordagem baseada em GNNs que utiliza *Spherical Message Passing (SMP)* para aprendizado de grafos moleculares 3D.

Essa abordagem representa a estrutura tridimensional das moléculas utilizando um sistema de coordenadas esféricas, o sistema final é SE(3)-invariante.

A sua diferença em relação a MPNNs tradicionais é que com o grafo molecular em mãos, é utilizada uma representação tridimensional adjacente em coordenadas esféricas ao invés da representação cartesiana tradicional o seja, ao invés de coordenas $(x, y, z)$, o SMP, representa a posição de cada átomo em termos de $(d, \theta, \phi)$, onde:

1. $d$ é a distância radial, ou seja, a distância entre dois átomos.
2. $\theta$ é o ângulo polar, que mede o ângulo em relação ao eixo z.
3. $\phi$ é o ângulo azimutal, que mede o ângulo no plano xy em relação ao eixo x.

A ideia fundamental dessa mudança representativa é que a geometria relevante para questões como quiralidade fiquem mais claras para o modelo pelos seguintes motivos:

1. Rotações em coordenadas cartesianas exigem que a rede adquira alguma noção a respeito de múltiplicações matriciais, enquanto em coordenadas esféricas, essa operação exige simplesmente mudanças dos ângulos $\theta$ e $\phi$ de maneira análoga a como a translação é feita em sistemas cartesianos.
2. Coordenadas esféricas são automaticamente invariantes à translação, uma vez que todas as variáveis de ângulo e as distâncias $d$ são invariantes a esse processo.
3. Ângulos de ligação e torção entre as moleculas são diretamente representados a partir das coordenadas polares ao invés de inferidos através de cálculos complexos como é o caso em coordenadas cartesianas.

Sobre o funcionamento da SMP, a ideia é realizar o *message passing* (passagem de mensagem) em termos das coordenadas esféricas.

Em um grafo molecular 3D, para cada átomo $i$, a posição relativa de cada átomo vizinho $j$ pode ser expressa como uma tupla $(d_{ij}, \theta_{ij}, \phi_{ij})$. O SMP utiliza essas informações para passar mensagens entre os nós do grafo, considerando não apenas as distâncias, mas também os ângulos e torções.

Assim, fazemos a passagem de mensagem no SMP com as seguintes etapas:

1. Para cada par de átomos $i$ e $j$ conectados por uma ligação, a distância $d_{ij}$ é calculada como:

$$d_{ij} = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2}$$

   Essa distância é então usada para calcular uma representação de distância esférica, $\Psi(d_{ij})$, que pode ser uma função base, como a função Bessel esférica.

2. Os ângulos $\theta_{ij}$ e $\phi_{ij}$ são calculados a partir das coordenadas cartesianas transformadas em coordenadas esféricas:

$$\theta_{ij} = \arccos\left(\frac{z_j - z_i}{d_{ij}}\right)$$

$$\phi_{ij} = \arctan\left(\frac{y_j - y_i}{x_j - x_i}\right)$$

   Esses ângulos são usados para calcular representações angulares $\Psi(\theta_{ij}, \phi_{ij})$.

3. A torção é o ângulo formado entre dois planos definidos por quatro átomos consecutivos. No SMP, a torção é considerada para capturar a quiralidade, ou seja, a assimetria espacial da molécula, que é essencial para distinguir moléculas quirais (como R e S). O ângulo de torção $\tau$ pode ser definido usando os vetores normais dos planos formados pelos átomos.

4. A mensagem $e'_k$ para cada ligação $k$ (entre átomos $i$ e $j$) é atualizada utilizando as representações de distância, ângulo e torção:

$$e'_k = \phi_e\left(e_k, v_i, v_j, \{e_h\}_{h \in \text{vizinhos de } j}, \Psi(d, \theta, \phi)\right)$$

   onde $\phi_e$ é uma função de atualização que agrega informações de vizinhança (utilizando uma rede neural, por exemplo), e $\Psi(d, \theta, \phi)$ representa a informação 3D completa.

5. Após a passagem de mensagem, as representações dos átomos $v_i'$ são atualizadas considerando todas as mensagens recebidas das suas ligações:

$$v_i' = \phi_v \left( v_i, \sum_k e_k' \right)$$

onde $\phi_v$ é uma função de agregação e atualização, como uma rede neural.

O SMP foi aplicado em várias tarefas de previsão molecular, como energia de interação, propriedades eletrônicas, etc., demonstrando melhorias em comparação com métodos anteriores, especialmente na tarefa de identificar corretamente a quiralidade molecular.

### 2.6.2 *Chiral InterRoto-Invariant Neural Network* ChIRo

Já a ChIRo (*Chiral InterRoto-Invariant Neural Network*) (Adams *et al.*, 2021) é um modelo projetado para ser capaz de diferenciar enantiômeros. Enquanto outras implementações de GNNs 3D são projetadas para serem sensíveis à quiralidade incorporando somente invariância SE(3) (Liu *et al.*, 2022), a ChIRo permite a diferenciação de estereoisômeros adicionando invariância a torções de ligações.

Essa sensibilidade reside em seu codificador de torções, que alcança a invariância a torções codificando os ângulos de torção de uma maneira que respeita o acoplamento inerente entre eles, mantendo a invariância a rotações sobre ligações internas sem perder a capacidade de distinguir enantiômeros.

O resultado é que desta forma o modelo capaz de prever com maior precisão propriedades moleculares sensíveis à quiralidade, como a atividade óptica de enantiômeros ou a afinidade de ligação em contextos de docking molecular.

## 2.7 Biblioteca e métodos de modelagem molecular

### 2.7.1 RDKit

Por sua vez, o RDKit é uma biblioteca *open-source* amplamente utilizada para tarefas de química computacional e informatica química.

Esta biblioteca oferece um conjunto de ferramentas para o processamento e manipulação de estruturas químicas, permitindo desde a geração e manipulação de moléculas até a execução de cálculos químicos simples e complexos. Com RDKit, os usuários podem

realizar diversas operações, como a conversão entre diferentes formatos de arquivo molecular, a busca por subestruturas, a geração de descritores moleculares, e a construção de modelos QSAR (*Quantitative Structure-Activity Relationship*).

Além disso, RDKit inclui funcionalidades avançadas para a construção e otimização de conformações moleculares, utilizando métodos como a geometria de distância, o que é essencial para prever as possíveis conformações que uma molécula pode assumir. As seguintes seções abordam o ETKDG e UFF, métodos de geração e otimização conformacional que estão diretamente acessíveis através do RDKit.

### 2.7.2 *Experimental-Torsion-Knowledge Distance Geometry* (ETKDG)

O *Experimental-Torsion-Knowledge Distance Geometry* (ETKDG) é uma abordagem híbrida que combina informações experimentais sobre torções de ligações químicas com métodos de geometria de distância para gerar conformações moleculares mais realistas.

O método ETKDG está disponível no RDKit como parte das suas funcionalidades de geração de conformações moleculares, é possível aplicar o ETKDG para gerar conformações iniciais de uma molécula.

### 2.7.3 Otimizador *Universal Force Field* (UFF)

O *Universal Force Field* (UFF) é um campo de força generalizado que pode ser aplicado a praticamente todos os elementos da tabela periódica. Desenvolvido para fornecer uma ferramenta universal na otimização de geometria molecular, o UFF utiliza parâmetros de ligação baseados na tabela periódica, permitindo que ele seja usado para uma ampla gama de moléculas, incluindo aquelas que contêm metais de transição.

O UFF pode ser utilizado diretamente dentro do RDKit para a otimização de conformações moleculares. Após gerar as conformações iniciais utilizando, por exemplo, o método ETKDG, o RDKit permite que essas conformações sejam otimizadas aplicando o UFF, minimizando a energia para encontrar a geometria mais estável.

## 2.8 Análise Exploratória de Dados

A Análise Exploratória de Dados (AED) é um conjunto de técnicas e métodos usados para examinar, visualizar e resumir as características principais de um conjunto de dados, frequentemente com o auxílio de representações gráficas.

Introduzida por Tukey (1962) na década de 1970, a AED não segue uma abordagem rígida e estruturada, como as inferências estatísticas tradicionais, mas é uma prática investigativa destinada a revelar padrões, detectar anomalias, testar hipóteses e verificar suposições.

O objetivo principal da AED é obter *insights* iniciais sobre os dados antes de aplicar modelos mais complexos, como redes neurais ou outras técnicas de *machine learning*. Abaixo, são explicados alguns elementos desta análise.

## 2.9 Validação de Normalidade

A validação da normalidade dos dados é um passo essencial na AED, especialmente quando se pretende utilizar técnicas estatísticas que assumem uma distribuição normal dos dados. A normalidade implica que os dados seguem uma distribuição simétrica em torno da média, com a maioria dos valores próximos à média e caudas que se afinam gradualmente. Abaixo estão as principais ferramentas e testes utilizados para validar a normalidade dos dados:

1. Histograma: Um histograma é uma representação gráfica que exibe a distribuição de uma variável contínua. Para avaliar a normalidade, a forma do histograma deve ser aproximadamente simétrica e em forma de sino.

2. Teste de Shapiro-Wilk: Desenvolvida por Shapiro e Wilk (1965), este é um dos testes mais poderosos para pequenas amostras e é amplamente utilizado para validar a normalidade. O teste avalia a hipótese nula de que os dados seguem uma distribuição normal. Um p-valor alto (tipicamente ¿ 0,05) indica que não há evidência suficiente para rejeitar a normalidade.

3. Teste de Kolmogorov-Smirnov (K-S): Desenvolvido por Kolmogorov (1933) e embora menos poderoso que o Shapiro-Wilk, o K-S é útil para amostras maiores. Ele compara a distribuição cumulativa dos dados com a distribuição cumulativa esperada de uma distribuição normal.

### 2.9.1 Análise de *Outliers*

*Outliers* são observações que se desviam significativamente das outras observações no conjunto de dados. A identificação e tratamento de *outliers* são críticos, pois eles podem distorcer a análise estatística e afetar o desempenho dos modelos de *machine learning*, alguns métodos de análise de *outliers* seguem:

1. *Boxplot*: O *boxplot* é uma ferramenta gráfica simples, mas eficaz, para identificar *outliers*. Ele mostra a mediana, quartis, e possíveis *outliers* (pontos fora dos "bigodes"do *boxplot*. *Outliers* são definidos como valores que estão além de 1,5 vezes o intervalo interquartil (IQR) acima do terceiro quartil ou abaixo do primeiro quartil.

2. *Z-Score*: O Z-score mede quantos desvios padrão um dado ponto está da média. Valores de *Z-score* maiores que 3 (ou menores que -3) são tipicamente considerados *outliers* em uma distribuição normal.

3. Média e Desvio Padrão: Em uma distribuição normal, aproximadamente 68% dos dados caem dentro de um desvio padrão da média, 95% dentro de dois desvios padrão, e 99,7% dentro de três desvios padrão. Valores fora desses intervalos são potenciais outliers.

### 2.9.2 Transformação de Dados

Se os dados não seguem uma distribuição normal, pode ser necessário aplicar transformações para normalizá-los. Isso é especialmente relevante se os modelos que serão aplicados posteriormente assumirem normalidade. Durante o desenvolvimento do trabalho foi a transformação logarítmica, aplicada principalmente a dados com uma distribuição assimétrica à direita (positivamente enviesada), com essa transformação é possível suavizar a variabilidade dos dados e aproximá-los de uma distribuição normal.

## 2.10 Elementos de Machine Learning

Nas sub-seções abaixo são explicados alguns elementos de machine learning com foco nas ferramentas citadas durante o desenvolvimento do projeto.

### 2.10.1 Funções de Ativação

As funções de ativação são componentes essenciais em redes neurais, responsáveis por introduzir não-linearidades nos modelos. Sem essas funções, as redes neurais se comportariam como simples modelos lineares, incapazes de capturar relações complexas nos dados, algumas destas são:

1. ReLU (Rectified Linear Unit): É uma das funções de ativação mais utilizadas devido à sua simplicidade e eficácia. Ela transforma as entradas negativas em zero, mantendo as positivas, o que ajuda a rede a aprender representações úteis sem saturar o gradiente.

2. *Leaky ReLU*: Uma variação da ReLU, onde em vez de zero para valores negativos, é aplicada uma pequena inclinação, permitindo que a rede continue a aprender mesmo com valores negativos. Isso evita o problema conhecido como "neurônios mortos".

3. *Softmax*: Usada comumente na camada de saída para problemas de classificação multiclasse, a função softmax transforma os valores de saída em probabilidades, permitindo que a rede atribua uma probabilidade a cada classe.

### 2.10.2 Função de Perda *(Loss Function)*

A função de perda é uma medida que quantifica a diferença entre as previsões do modelo e os valores reais. Para problemas de classificação, a função de perda mais comum é a *cross-entropy loss*.

1. *Cross-Entropy Loss*: Essa função avalia a distância entre a distribuição de probabilidades prevista pelo modelo (após o *softmax*) e a distribuição verdadeira (*one-hot encoded*). Em termos matemáticos, ela é definida como:

$$L = -\sum_{i=1}^{n} y_i \log(\hat{y}_i)$$

Onde $y_i$ representa a classe verdadeira e $\hat{y}_i$ é a probabilidade prevista para essa classe. Essa função penaliza previsões erradas, especialmente aquelas feitas com alta confiança (ou seja, quando $\hat{y}_i$ é muito alto para a classe incorreta).

2. *Binary Cross-Entropy Loss*: Essa é uma função de perda amplamente utilizada em problemas de classificação binária. Ela mede a diferença entre as previsões do modelo (como probabilidades) e os rótulos verdadeiros (0 ou 1), penalizando previsões incorretas com alta confiança. Matematicamente, ela é definida como:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)]$$

Aqui, $y_i$ é o rótulo verdadeiro e $\hat{y}_i$ é a probabilidade prevista para a classe 1. A função incentiva o modelo a aumentar a probabilidade da classe correta e reduzir a da classe incorreta.

### 2.10.3 Treinamento e Validação

O processo de treinamento de modelos como SphereNet e ChIRo envolve a propagação do erro (*backpropagation*) e a atualização dos pesos por meio de algoritmos de otimização, como o *gradient descent* e suas variantes, incluindo Adam. O modelo é treinado para minimizar a função de perda nos dados de treinamento e é avaliado periodicamente em dados de validação para verificar seu desempenho fora da amostra de treinamento.

# Capítulo 3

# Desenvolvimento

Este capítulo detalha o processo metodológico empregado no desenvolvimento e avaliação de modelos de aprendizado de máquina para a classificação de quiralidade molecular. O trabalho é estruturado em quatro seções principais, cada uma abordando um aspecto do projeto.

Uma consideração importante no processo de desenvolvimento foi o tratamento e armazenamento dos dados moleculares. Os arquivos foram salvos como objetos *Chem* do *RDKIT*, utilizando o formato de serialização *pickle*. Este método transforma um objeto em execução na memória em uma série de caracteres transmissíveis, permitindo a preservação da conformação tridimensional das moléculas.

Cada linha dos arquivos de dados contém uma instância serializada desses objetos, que já incluem a conformação tridimensional da molécula especificada.

Tradicionalmente, é mais comum armazenar descritores ou características que permitam a reconstrução dos objetos, em vez de serializar os objetos diretamente. Isso se deve ao fato de que a transformação de um objeto da memória em uma série de strings pode ocasionar problemas na sua execução posterior e potencialmente introduzir vulnerabilidades.

É importante notar que esta abordagem, embora não seja a mais comum na prática geral devido a potenciais riscos de segurança, foi escolhida por estar alinhada com os estudos de caso analisados durante a realização deste trabalho.

Nessa seara, o capítulo está organizado de forma a refletir o fluxo de trabalho do projeto, começando com a preparação dos dados e culminando nos testes dos modelos:

A primeira seção detalha o processo de criação do dataset principal, que classifica moléculas em três categorias: aquiral, R (rectus) ou S (sinister).

A segunda seção aborda a tentativa de desenvolvimento de um dataset para quiralidades complexas, discutindo os desafios encontrados neste processo.

A terceira seção foca no desenvolvimento de um dataset para indicadores quirais contínuos, especificamente a medida quiral contínua (CCM).

A quarta e última seção apresenta os métodos e resultados dos testes realizados com dois modelos de aprendizado de máquina (ChIRo e SphereNet) na tarefa de classificação de quiralidade.

Cada seção inclui detalhes sobre as fontes de dados, métodos de processamento, desafios encontrados e soluções implementadas. O capítulo conclui com uma análise dos resultados obtidos e suas implicações para o campo de estudo da quiralidade molecular.

## 3.1 Desenvolvimento do dataset de três categorias (aquiral, R ou S)

O primeiro passo foi a obtenção do dataset de moléculas quirais utilizados por Adams *et al.* (2021).

Este dataset, disponível por um link no artigo original, foi criado para a classificação dos centros quirais em R ou S. A criação do dataset envolveu os seguintes passos:

1. Seleção aleatória de 50 dos 6199 arquivos SDF disponíveis no servidor FTP do projeto PubChem3D (Bolton *et al.*, 2011), especificamente do diretório "*10 conformers per compound*". Estes dados encontram-se disponíveis no servidor FTP 2. Filtragem dos dados para incluir apenas grafos 2D com pelo menos dois estereocentros, onde cada estereoisômero possuísse no mínimo dois conformômeros. 3. Separação de 20

Posteriormente, o dataset R/S foi dividido na proporção 70/15/15, com pares de enantiômeros atribuídos à mesma partição de dados. Destacamos os seguintes conjuntos:

- Conjunto de treinamento: 326.865 confômeros de 55.084 estereoisômeros (27.542 pares de enantiômeros); - Conjunto de validação: 70.099 confômeros de 11.748 estereoisômeros (5.874 pares de enantiômeros); - Conjunto de teste: 69.719 confômeros de 11.680 estereoisômeros (5.840 pares de enantiômeros).

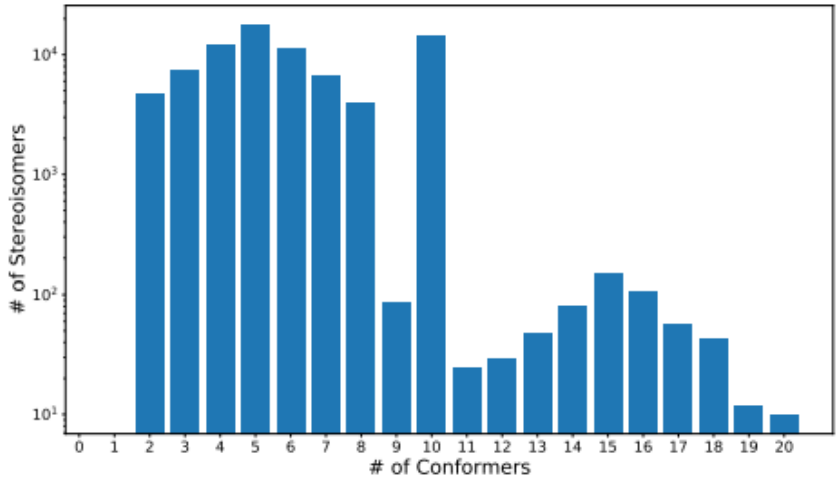
A distribuição do dataset em relação ao número de estereoisômeros por número de conformômero é ilustrada na Figura 3.1. Já a tabelaRS mostra a quantidade de confórmômeros e enantiômeros no dataset de origem.

Para complementar o dataset com moléculas aquirais, utilizou-se o *ChemBL* (Gaulton

Tabela 3.1: Balanço de rótulos R/S no dataset R/S

| rótulos R/S | # de conformômeros | # de enantiômeros |
|---|---|---|
| R | 236222 | 39256 |
| S | 230461 | 39256 |

Figura 3.1: Distribuição do número de estereoisômeros por número de conformômero no dataset RS.




Fonte: Adams *et al.* (2021)

*et al.*, 2011). O processo envolveu a busca por *datasets* públicos com marcadores quirais, a utilização do *ChemBL*, que indica se a molécula é quiral ou aquiral, a geração de múltiplos conformômeros para cada molécula aquiral usando o método ETKDG para o *embedding* inicial, seguido pela aplicação do otimizador UFF.

Por fim, foi feito um balanceamento final do dataset completo, apresentado no apêndice A. O resultado final pode ser visto na seção 4.1.1.

## 3.2 Desenvolvimento do dataset de quiralidades complexas

Já na tentativa de criar um dataset que distinguisse variedades quirais como planar e quiral, ressaltamos que, apesar de extensa busca bibliográfica, não foram encontrados dados suficientes ou coerentes para essa tarefa.

O único dataset disponível online que distingue as variedades quirais entre central, planar e axial é o CHIRAL (Kok *et al.*, 2022), coletado para predizer a categoria quiral baseado em desenhos lineares de moléculas. Porém, alguns problemas surgiram ao analisar

Tabela 3.2: Balanço de rótulos no dataset CHIRAL

| Categoria | # de moléculas |
|---|---|
| Aquiral | 17 |
| Quiral central | 116 |
| Quiral planar | 61 |
| Quiral axial | 11 |
| Total | 205 |

o dataset.

Dentre estes, destacamos o número de moléculas repetidas, resultando num número total de moléculas anotadas que apresentavam inconsistências na indicação de quiralidade. Devido a essas limitações, o desenvolvimento de um dataset de quiralidades complexas não foi incluído nos resultados finais deste trabalho.

Essas limitações são apresentadas na tabela3.2.

## 3.3 Desenvolvimento do dataset do indicadores quirais contínuos

Devido às questões ressaltadas, foi selecionado somente um indicador quiral contínuo: a medida quiral contínua (CCM) (Zabrodsky e Avnir, 1995).

Essa escolha se deve à proximidade desta métrica com o referencial utilizado pelos modelos a serem testados, como o ChIRo e o SphereNet, no sentido de metrificar a simetria que esses modelos buscam incorporar. Dentre os testes de modelos na literatura, destacamos o ChIRo por Adams *et al.* (2021) e o SphereNet por Liu *et al.* (2022).

O dataset foi criado utilizando a biblioteca `cosymlib` (Alemany *et al.*, 2021), buscando a simetria $C_s$ no *embedding* molecular.

Por se tratar de um algorítmo custoso computacionalmente e que difere muito em tempo de execução por molécula, foi desenvolvido um algorítmo de processamento paralelo para otimizar a utilização dos 4 núcleos disponíveis na máquina e o código, disponível no apêndice C.

Além disso, como o algorítmo demorava algumas horas, em algumas moléculas específicas foi desenvolvido um método que impedia o comportamento bloqueador da biblioteca, baseando-se na execução com *timeout* progressivo sobre o dataset RSA sem filtragens de forma que as moléculas que entregavam resultados mais rapidamente eram

executadas primeiro.

Assim, foi possível obter a simetria da maioria das moléculas evitando as mais demoradas. O resultado final também encontra-se no capítulo 5.

## 3.4 Teste dos modelos na tarefa de distinguirem quiralidade (RSA)

Abaixo são apresentados os métodos e resultados obtidos ao testar os 2 modelos na tarefa de classificação de conformômeros entre R (*rectus* direito), S (*sinister* esquerdo) e A (aquiral),

Os modelos foram selecionados por serem os melhores modelos disponíveis de acordo com Adams *et al.* (2021) nas tarefas que dizem respeito a quiralidade.

Nas seções abaixo são explicados os procedimentos necessários para adaptar os modelos e seus códigos para a tarefa de classificação RS.

### 3.4.1 ChIRo *(Chiral InterRoto-Invariant Neural Network)*

Para testar o ChiRo, algumas modificações foram necessárias, disponíveis no fork do ChIRo na conta do autor `https://github.com/barelias/ChIRo`.

A primeira modificação foi feita filtrando o dataset para conter somente moléculas que possuissem angulos dihedrais, pois o funcionamento do encoding de angulos de rotação, essenciais para a rede, dependem desta característica. Sendo assim, a função disponível no anexo B foi desenvolvida para filtrar essas moléculas.

A próxima modificação foi feita na saída da rede para ter 3 resultados (R, S, A) em vez de classificação binária.

A modificação foi feita na classe `Encoder` do módulo `model.alpha_encoder` para incluir um parâmetro extra, `is_binary` - que controla a dimensão da saída do *Multi-Layer Perceptron*, a última camada do *head* do modelo.

No caso do valor parâmetro `is_binary` tenha valor booleano verdadeiro, a saída tem dimensão 1 e, caso contrário, 3, que é referente à quantidade de classes do dataset RSA.

Sendo assim, a última camada entrega os 3 *logits* que entrega a probabilidade de entrada ser de cada classe. No algorítmo de validação da rede foi adicionada uma softmax para retornar apenas um inteiro que corresponde ao índice da classe da entrada.

O tipo de *loss* utilizado no treinamento da rede também foi modificado de *Binary Entropy Loss* para *Categorial Entropy Loss* para que a rede interpretasse corretamente a saída do modelo e o *backpropagation* funcionasse.

Os resultados desse treinamento podem ser vistos na seção 4.1.3

### 3.4.2 *SphereNet*

Já para executar a classificação RSA sobre o modelo *SphereNet* (Liu *et al.*, 2022) foram utilizadas em primeira instância modificações semelhantes às feitas para o ChIRo.

As classes de modelo e as funções auxiliares de treinamento e avaliação foram modificadas para terem 3 *logits* de saída e o loss foi modificado para *Cross Entropy Loss*.

Além disso, foi necessário modificar as funções de treinamento para que *batches* que retornavam saídas $NaN$ não fossem contabilizadas.

Adams *et al.* (2021) evitaram esse problema mapeando todas as moléculas que não retornavam valores númericos que fazem sentido, esse esforço não foi executado no contexto deste trabalho.

Os resultados do treinamento podem ser verificados na seção 4.1.3.

Este capítulo detalhou o processo metodológico empregado no desenvolvimento de datasets e na adaptação de modelos de aprendizado de máquina para a classificação de quiralidade molecular. Abordamos os aspectos cruciais do projeto, como a criação de um dataset de três categorias (aquiral, R ou S), combinando dados de moléculas quirais e aquirais de fontes como PubChem3D e ChemBL.

Além disso, a tentativa de desenvolvimento de um dataset para quiralidades complexas, embora não tenha sido bem-sucedida devido à escassez de dados coerentes, proporcionou avanços valiosos sobre as limitações atuais neste campo e sobre os aspectos metodológicos na consecução desse próprio trabalho.

Contribuiu, assim, para o desenvolvimento de um dataset de indicadores quirais contínuos, focando na medida quiral contínua (CCM), que oferece uma métrica mais refinada para a análise de quiralidade. Por fim, houve a adaptação e preparação de dois modelos de ponta, ChIRo e SphereNet, para a tarefa específica de classificação de quiralidade em três categorias.

Ao longo deste processo, enfrentamos e superamos diversos desafios, desde a manipulação de grandes volumes de dados moleculares até a modificação de arquiteturas de redes neurais complexas. O tratamento cuidadoso dos dados, utilizando objetos *Chem* do

*RDKIT* e serialização *pickle*, embora não convencional, permitiu preservar informações sobre a conformação tridimensional das moléculas.

No próximo capítulo, exploraremos em detalhes os resultados obtidos a partir destes desenvolvimentos. Examinaremos o desempenho dos modelos ChIRo e SphereNet no dataset RSA, analisaremos a eficácia da medida quiral contínua (CCM) como um indicador de quiralidade, e discutiremos as implicações destes resultados para o campo da química computacional e do aprendizado de máquina aplicado a problemas moleculares.

# Capítulo 4

# Análise Experimental

Na seção anterior, foram apresentados os métodos utilizados para desenvolver e testar modelos de aprendizado de máquina capazes de lidar com a classificação de moléculas quirais e aquirais, com foco na aplicação de redes neurais gráficas (GNNs). Além disso, discutimos a construção e a preparação dos datasets utilizados nos experimentos, destacando a importância da simetria molecular e sua representação computacional.

Agora, nesta seção de resultados, serão analisados os principais achados experimentais, com ênfase na performance dos modelos ChIRo e SphereNet na tarefa de distinção entre moléculas quirais e não-quirais.

As limitações e desafios enfrentados pelos modelos, bem como possíveis soluções, serão discutidos com base nos dados obtidos, destacando caminhos para melhorias futuras na classificação de quiralidade.

## 4.1 Discussão dos Resultados

### 4.1.1 Dataset de três categorias (aquiral, R ou S)

Utilizando os métodos descritos no capítulo de desenvolvimento, obtivemos um dataset de classificação RSA contendo 100.382 conformômeros, o que representa 25,56

O histograma resultante do número de conformômeros por número de átomos segue na figura 4.1.

Por sua vez, a tabela 4.1 possui a divisão do dataset por classes e a tabela 4.2 possui a divisão do dataset entre treinamento, validação e teste.

É importante notar que a proporção do dataset RSA em relação ao dataset RS original

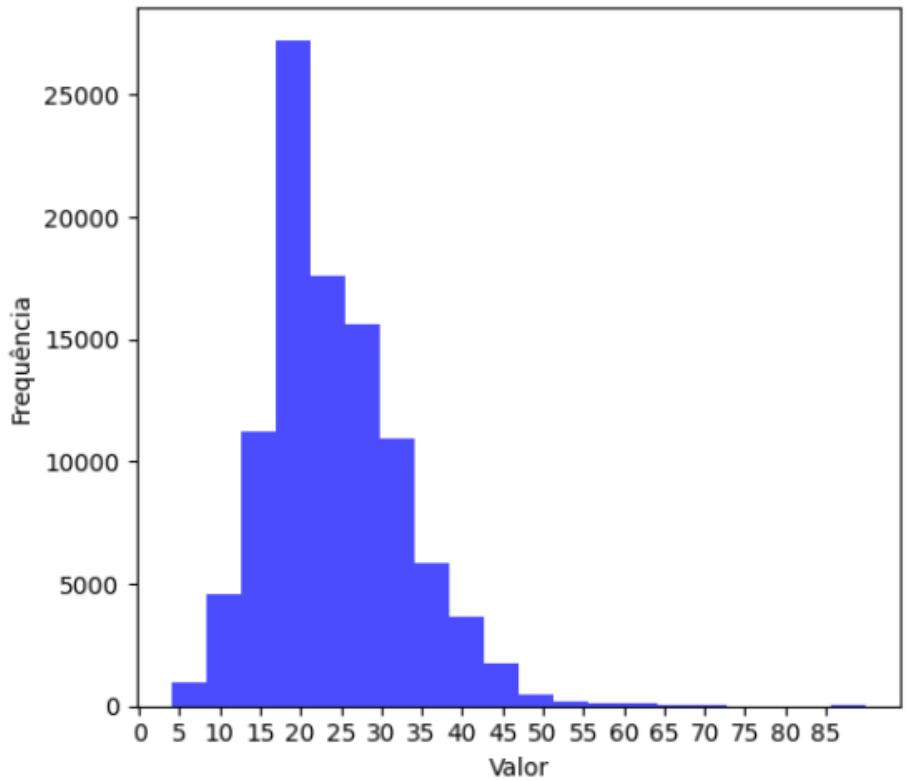
Tabela 4.1: Balanço de rótulos no dataset RSA

| rótulos RSA | # de conformômeros | # de enantiômeros |
|---|---|---|
| R | 33695 | 5717 |
| S | 32791 | 5720 |
| A | 33894 | 5649 |

Tabela 4.2: Divisão do dataset RSA entre treinamento, validação e teste

| Categoria | # de conformômeros | # de enantiômeros |
|---|---|---|
| Treinamento | 70181 | 11965 |
| Validação | 15054 | 15145 |
| Teste | 15145 | 2565 |

Figura 4.1: Distribuição do número de conformômeros por número de átomos no dataset RSA.




Fonte: Autor

Tabela 4.3: Balanço de rótulos no dataset CCM

| rótulos RSA | # de conformômeros | # de enantiômeros |
|---|---|---|
| R | 33695 | 5717 |
| S | 32791 | 5720 |
| A | 34326 | 5721 |

Tabela 4.4: Divisão do dataset CCM entre treinamento, validação e teste

| Categoria | # de conformômeros | # de enantiômeros |
|---|---|---|
| Treinamento | 70565 | 12029 |
| Validação | 15078 | 2560 |
| Teste | 15169 | 2569 |

pode limitar a capacidade de fazer afirmações mais categóricas sobre a relação entre o desempenho dos modelos nas tarefas de classificação RS e RSA. Portanto, mesmo com os resultados apresentados neste trabalho, experimentos futuros ainda serão necessários.

### 4.1.2 Dataset de indicadores quirais contínuos

Já acerca dos datasets de indicadores quirais contínuos, o dataset CCM obtido foi levemente maior que o RSA.

Isso ocorreu porque, embora algumas moléculas tenham sido filtradas devido ao longo tempo de processamento, as filtragens aplicadas no dataset RSA para compatibilidade com os modelos testados o reduziram significativamente.

O tamanho do dataset CCM por categoria RSA está disponível na tabela 4.3 e a divisão entre datasets de treinamento, validação e teste pode ser observada na tabela 4.4.

Uma análise exploratória foi executada sobre o dataset para verificar como o comportamento do CCM se dá, e em especial foi verificado se há uma diferença estatística do CCM quiral e aquiral .

Ainda mais importante, foi possível verificar se é possível dizer se o CCM, em média, é maior para moléculas quirais, o que seria esperado, no gráfico de densidade CCM por categoria.

Na figura 4.2, pode ser verificado que pelo menos visualmente a métrica de simetria aparenta ser menor para moléculas quirais, apesar de haver muita intersecção entre as duas distribuições. Desse modo, não é possível de maneira categórica afirmar observando apenas o valor do CCM se uma molécula é quiral ou não.

Figura 4.2: Distribuição de densidade CCM por categoria

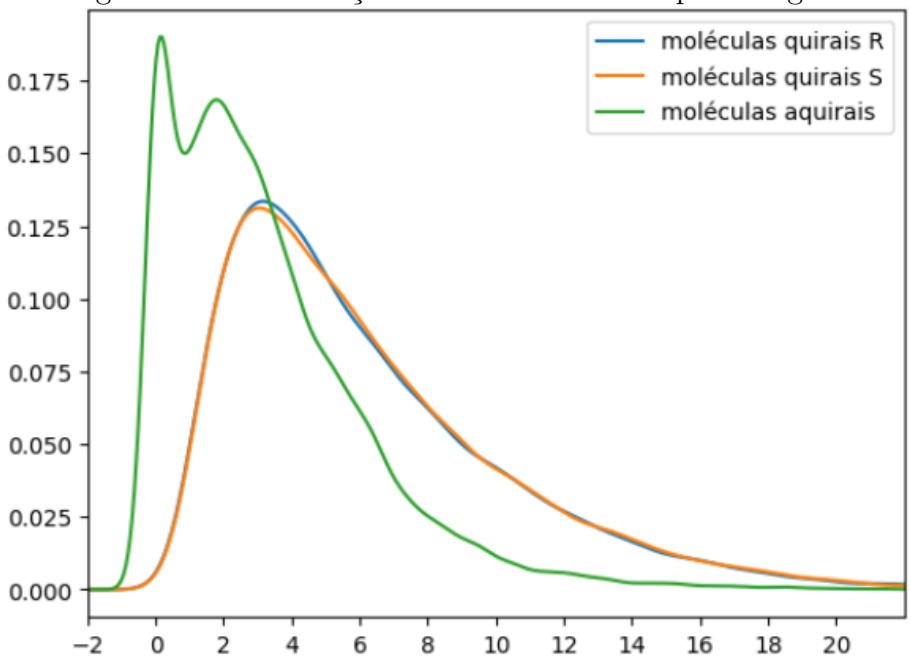



Tabela 4.5: Resultados dos testes de normalidade sobre o dataset

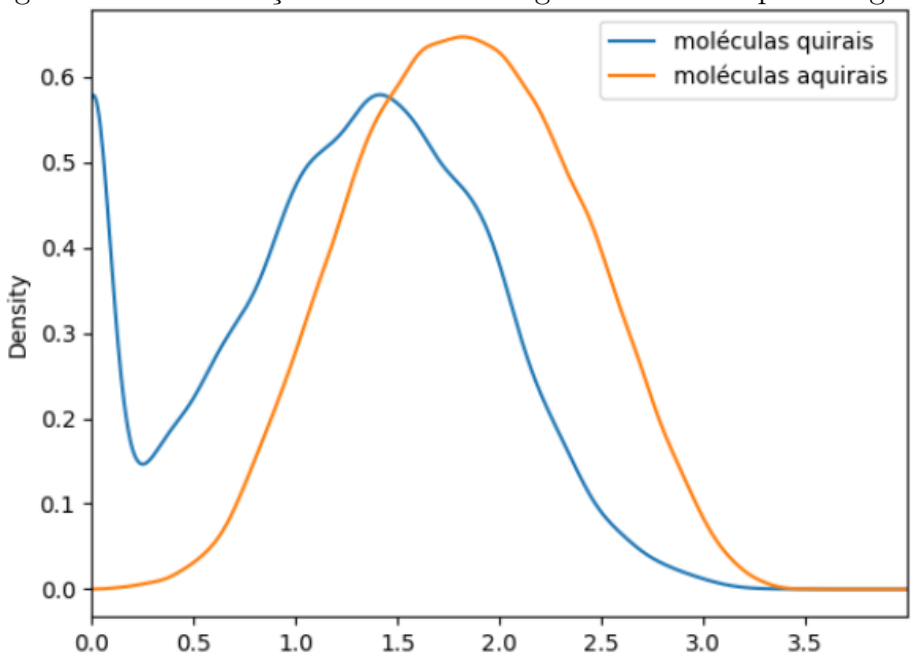
| Categoria | p-valor Shapiro-Wilk | p-valor Kolmogorov-Smirnov | p-valor Jarque–Bera |
|---|---|---|---|
| RSA | 0.0 | 0.0009 | 0.0 |
| Quiral | 0.0 | 0.0009 | 0.0 |
| Aquiral | 0.0 | 0.0009 | 0.0 |

Foram então aplicados vários testes de normalidade, com o objetivo final de avaliar se é possível verificar a diferença entre as distribuições somente utilizando a média.

Como pode ser observado na tabela 4.5, utilizando os método de Shapiro e Wilk (1965), o de Kolmogorov (1933) e Smirnov (1948) e o de Jarque e Bera (1980), os testes de normalidade indicam que, sob nenhum desses parâmetros, os datasets divididos por categoria podem ser considerados normais. O teste de Levene (valor 2816,5277 e p-value 0) revela que as distribuições diferem extremamente em termos estatísticos, mas estes testes ainda não são suficientes para concluir que uma distribuição é maior que a outra.

Para tentar normalizar as distribuições, aplicou-se uma função logarítmica aos dados, na figura ig:distlogCCM. Porém, mesmo após esta transformação, os testes ainda rejeitaram a hipótese de normalidade.

Figura 4.3: Distribuição de densidade logarítmica CCM por categoria




### 4.1.3 Teste dos modelos na tarefa de distinguir quiralidade e não-quiralidade (RSA)

**ChIRo *(Chiral InterRoto-Invariant Neural Network)***

Durante o treinamento do ChIRo, observou-se uma diminuição estável do *loss* no conjunto do treinamento, conforme 4.4.

Porém, o *loss* da validação diminuiu de maneira instável, como mostra a figura 4.5. É possível interpretar essa situação de algumas formas:

Possíveis interpretações para este comportamento incluem:

1. Overfitting: O modelo pode estar aprendendo muito bem os detalhes e o ruído dos dados de treinamento, mas não generalizando bem para novos dados. 2. Representatividade do dataset: O dataset RSA pode não ser totalmente representativo do problema real. 3. Learning rate inadequado: O learning rate pode estar muito alto ou muito baixo, afetando a capacidade do modelo de minimizar o *loss* de validação de maneira consistente.

Desse modo, a acurária geral do modelo e por categoria pode ser verificada na tabela 4.7 e mostra uma tendência.

Apesar do resultado geral ter sido baixo (64,41

Uma possível solução é incluir mais representatividade das moleculas quirais, uma vez que nessas categorias o dataset RSA é somente um subconjunto do dataset RS. O modelo também se mostra mais perto de RSA-completo, uma vez que é possível combinar um

Figura 4.4: Loss do dataset de treinamento por epoch

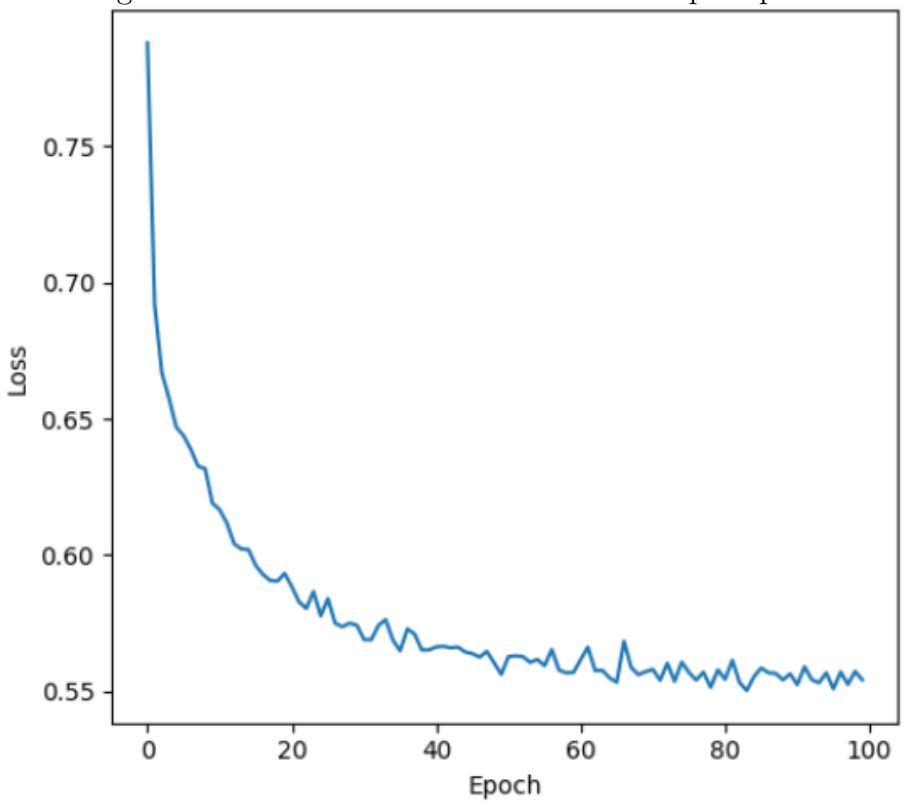



Figura 4.5: Loss do dataset de validação por epoch

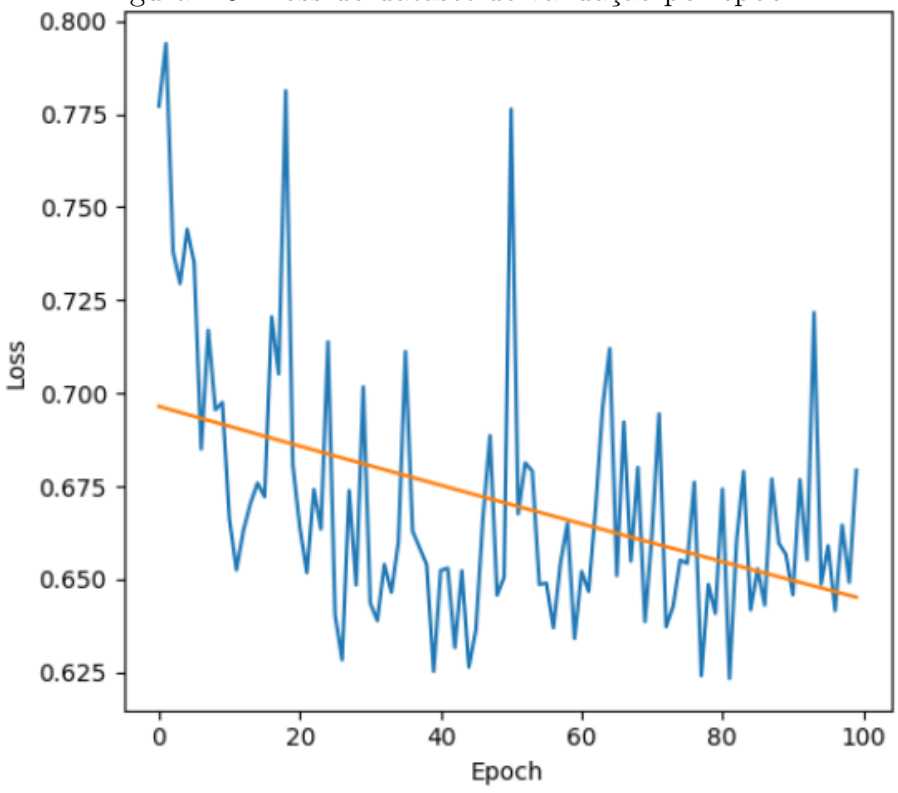
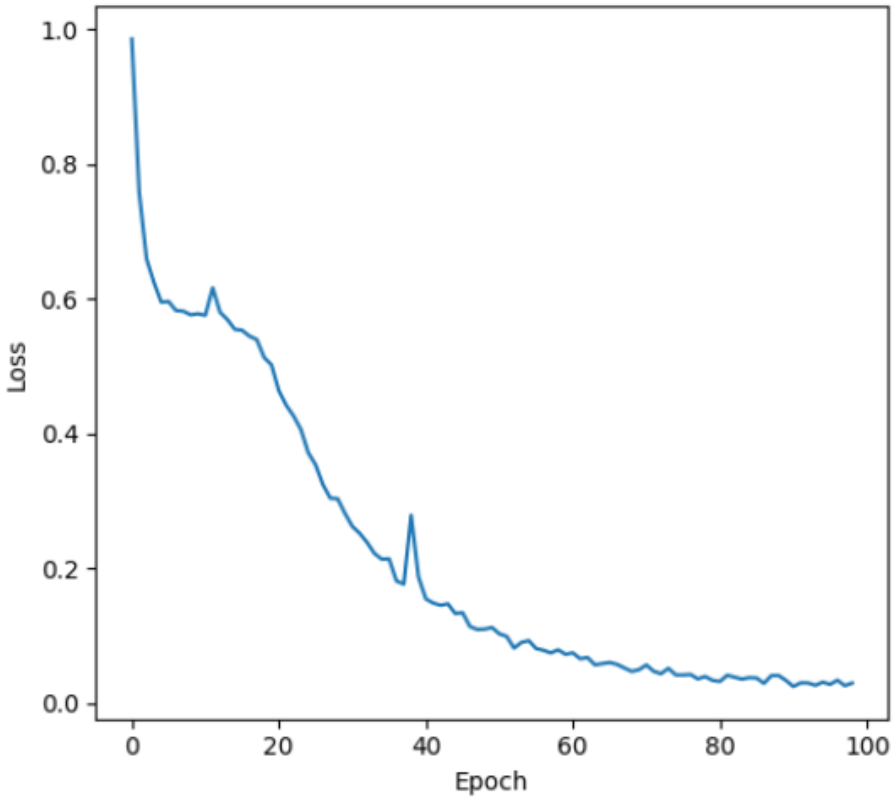


Tabela 4.6: Acurácia do ChIRo na classificação RSA

| Categoria | # Acurácia |
|---|---|
| RSA | 64,41% |
| R | 6,15% |
| S | 89,12% |
| A | 97,29% |

Figura 4.6: Loss de treinamento do SphereNet




classificador RS, muito acurado em classificar os tipos de quiralidade com o classificador RSA, muito acurado em distinguir quiralidade de aquiralidade.

### *SphereNet*

De maneira análoga ao ChIRo, o *loss* de treinamento, visto na figura 4.6 foi mais estável que o de validação, visto na figura 4.7, levando às mesmas possíveis interpretações vistas na seção anterior, ou seja, que o treinamento tenha chegado em um estado de *overfitting*, que o dataset talvez não seja representativo do problema real ou que o *learning rate* talvez esteja em um valor não apropriado.

A acurária geral do modelo e por categoria pode ser verificada na tabela 4.8.

Os resultados do SphereNet diferem significativamente dos encontrados no ChIRo.

Figura 4.7: Loss de validação do SphereNet

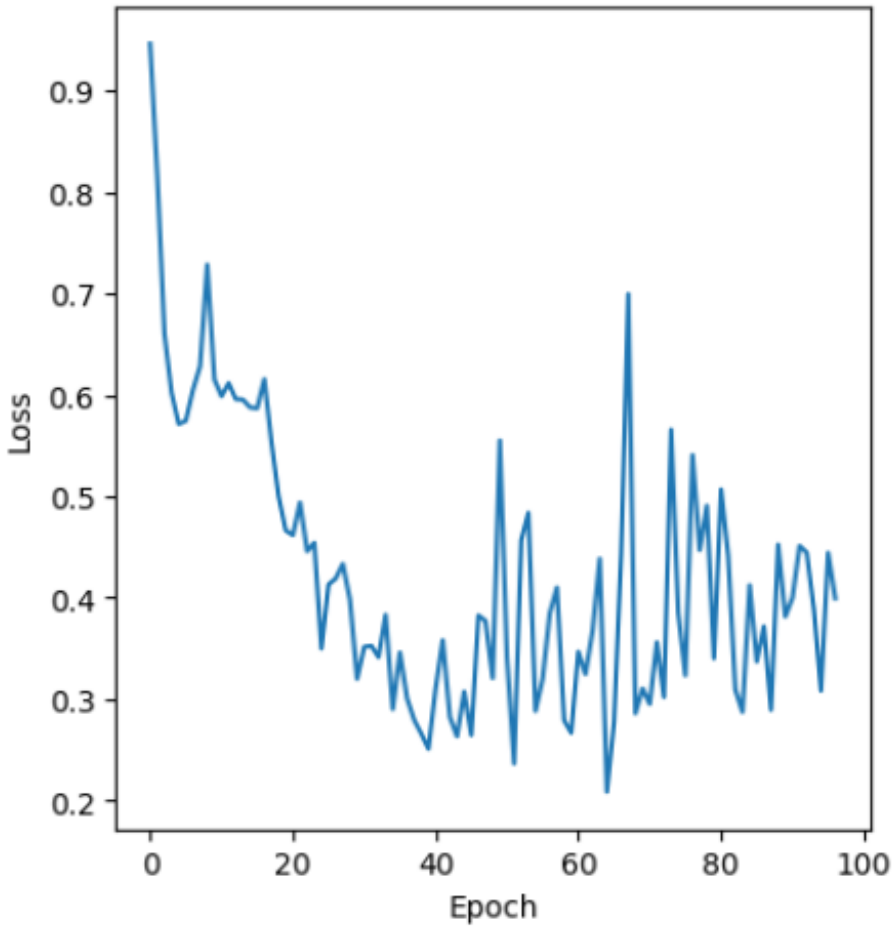



Tabela 4.7: Acurácia do ChIRo na classificação RSA

| Categoria | # Classificado como R |
|---|---|
| R | 96,91% |
| S | 50,63% |
| A | 92,67% |

Tabela 4.8: Acurácia do *Spherenet* na classificação RSA

| Categoria | # Acurácia |
|---|---|
| RSA | 50,14% |
| R | 96,91% |
| S | 47,37% |
| A | 7,32% |

Tabela 4.9: Acurácia dos modelos na classificação RS

| Modelo | # Acurácia |
|---|---|
| ChIRo | 98,5% |
| SphereNet | 98,2% |

Enquanto a classificação de R (96,91

No restante dos casos, os exemplares que deveriam ter sido rotulados com S o modelo classificou como R. O mesmo acontece para o A, onde em 92,67% dos dados o modelo classifica estas entradas também como R.

Algumas hipóteses podem ser formuladas a partir desse resultado. Inicialmente, talvez o modelo necessite de mais dados. Além disso, a falta de filtragem das moléculas pode ter contribuído para que o modelo não possa processar, subsequentemente excluindo *batches* que continham essas moléculas e desfavorecendo os rótulos aquirais, que é o único rótulo cujas entrada continham estas moléculas não processáveis.

É possível também que dada a comparação com o resultado do ChIRo, a invariância $SE(3)$ combinada à invariância à torção de ligações seja mais robusta não somente na classificação RS como também na classificação RSA.

Estes resultados sugerem que a classificação RSA pode ser uma maneira mais clara de avaliar o desempenho na interpretação da quiralidade em modelos que se propõem a interpretá-la, uma vez que as diferenças entre os modelos são mais evidentes nesta tarefa do que na classificação RS original.

Corrobora com essa visão apontar que, uma vez que comparado ao resultado da classificação RS vista na tabela 4.9 as diferenças entre os modelos são salientadas.

Nesta seção, foram discutidos os resultados obtidos nos experimentos com os modelos ChIRo e SphereNet, avaliando suas capacidades de distinguir entre moléculas quirais e aquirais. Embora ambos os modelos apresentem limitações, especialmente na classificação de enantiômeros, observou-se que o ChIRo teve uma performance ligeiramente superior ao SphereNet na distinção de quiralidade, possivelmente devido à sua maior robustez na

interpretação de simetrias moleculares.

Além disso, os resultados destacaram a necessidade de datasets mais representativos e ajustes em hiperparâmetros, como o learning rate, para evitar problemas. Essas descobertas nos indicam a necessidade de aprimoramentos nos métodos de treinamento e na representatividade dos dados, estabelecendo as bases para as considerações finais e futuras direções de pesquisa a serem abordadas na última seção do trabalho.

# Capítulo 5

# Conclusão e trabalhos futuros

Os testes realizados com os modelos SphereNet e ChIRo destacaram contribuições à literatura e à pesquisa empírica no que tange a classificação molecular no contexto da quiralidade. Em particular, a introdução da classificação RSA demonstrou maior precisão na segregação de moléculas quirais e aquirais, quando comparada à abordagem RS tradicional. Isso evidencia o impacto positivo da incorporação de simetrias e equivariâncias mais completas nos modelos de redes neurais gráficas aplicados à química computacional.

A abordagem computacional aqui proposta representa um avanço importante, sugerindo que a classificação de moléculas pode ser otimizada através do uso de arquiteturas mais sofisticadas, capazes de lidar com a complexidade da quiralidade molecular.

Entre as principais conclusões, destaca-se que, embora o modelo ChIRo tenha mostrado resultados promissores, o desempenho da SphereNet ainda requer melhorias substanciais, especialmente em termos de sua capacidade de generalização e robustez em conjuntos de dados de maior escala.

Outro ponto relevante foi a constatação de que a conformação geométrica, por si só, pode ser insuficiente para distinguir quiralidade de forma eficaz, conforme observado nos resultados do dataset CCM.

Essa descoberta abre um novo campo de investigação para a formulação de novas regras matemáticas que permitam uma segregação mais precisa das variedades quirais. Além disso, a proposta de criar um dataset de quiralidade contínua quântica e avaliar a capacidade dos modelos de identificar tal atributo, se realizada com sucesso representa uma contribuição para o campo da química computacional, com potencial para transformar as abordagens de modelagem molecular.

Em síntese, as contribuições deste trabalho se manifestam na ampliação das capa-

cidades das redes neurais gráficas para representar de maneira mais fiel a quiralidade molecular, trazendo avanços tanto para a ciência dos materiais quanto para a química computacional. A continuidade dessas pesquisas poderá consolidar novas arquiteturas de modelos mais eficientes e precisos, com implicações significativas para o desenvolvimento de tecnologias baseadas em propriedades moleculares complexas.

As futuras direções de pesquisa, como a exploração de diferentes hiperparâmetros e a criação de um dataset mais rigoroso e representativo, são essenciais para o refinamento dos modelos propostos.

Em suma, o trabalho futuro neste estudo sobre redes neurais gráficas (GNNs) e quiralidade molecular deve se concentrar na criação de conjuntos de dados mais diversificados e extensos, incorporando várias categorias de quiralidade, para avaliar completamente e aprimorar a precisão do modelo.

Além disso, existe a possibilidade de desenvolver novas arquiteturas de GNN que capturem melhor as complexidades da quiralidade molecular, aproveitando representações geométricas e simétricas avançadas. As limitações atuais incluem a necessidade de testes mais abrangentes dos modelos propostos, validação e avaliação comparativa em relação aos métodos existentes, além de garantir a viabilidade computacional ao lidar com conjuntos de dados moleculares em grande escala.

Será fundamental abordar as complexidades computacionais, os problemas de escalabilidade e as possíveis tendências em diversos conjuntos de dados. Além disso, as implementações práticas devem envolver análises comparativas de diferentes modelos de aprendizado de máquina, filtragem rigorosa de dados e exploração de regras matemáticas subjacentes à criação de conjuntos de dados, para solidificar os aprimoramentos propostos na modelagem precisa da quiralidade molecular.

# Referências Bibliográficas


Abraham, E. e Nitzan, A. (2024). Molecular chirality quantification: Tools and benchmarks. *The Journal of Chemical Physics*, **160**(16), 104–164.

Adams, K., Pattanaik, L. e Coley, C. W. (2021). Learning 3d representations of molecular chirality with invariance to bond rotations.

Alemany, P., Bernuz, E., Carreras, A. e Llunell, M. (2021). Cosymlib.

Bel, L. (1874). *Sur les relations qui existent entre les formules atomiques des corps organiques et le pouvoir rotatoire de leurs dissolutions*, páginas 337–347. Bulletin de la Société Chimique de Paris.

Bolton, E., Chen, J., Kim, S., Han, L., He, S., Shi, W., Simonyan, V., Sun, Y., Thiessen, P., Wang, J., Yu, B., Zhang, J. e Bryant, S. (2011). Pubchem3d: A new resource for scientists. *Journal of cheminformatics*, **3**, 32.

Cahn, R. S., Ingold, C. e Prelog, V. (1966). Specification of molecular chirality. *Angewandte Chemie International Edition in English*, **5**(4), 385–415.

Favre, H. A. e Powell, W. H. (2014). *Nomenclature of Organic Chemistry: IUPAC Recommendations and Preferred Names 2013*. 2011-2015 organic chemistry subject collection. RSC Publishing. ISBN 9780854041824.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B. e Overington, J. P. (2011). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, **40**(D1), D1100–D1107.

Hoff, V. (1874). *Archives Néerlandaises des Sciences Exactes et Naturelles*, páginas 445–454. Bulletin de la Société Chimique de Paris.

Jarque, C. M. e Bera, A. K. (1980). Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Economics Letters*, **6**(3), 255–259.

Kipf, T. N. e Welling, M. (2017). Semi-supervised classification with graph convolutional networks.

Klein, F. (1893). Vergleichende betrachtungen über neuere geometrische forschungen. *Mathematische Annalen*, **43**, 63–100.

Knudsen, O. (2005). *Lord Kelvin, Baltimore lectures on mathematical physics ((1884), 1904)*, páginas 748–756. ISBN 9780444508713.

Kok, Y. E., Woodward, S., Özcan, E. e Torres Torres, M. (2022). Identifying chirality in line drawings of molecules using imbalanced dataset sampler for a multilabel classification task. *Molecular Informatics*, **41**(12), 2200068.

Kolmogorov, A. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari*, **4**, 83–91.

Liu, Y., Wang, L., Liu, M., Zhang, X., Oztekin, B. e Ji, S. (2022). Spherical message passing for 3d graph networks.

Pasteur, L. (1848). Memoires sur la relation qui peut exister entre la forme crystalline et al composition chimique, et sur la cause de la polarization rotatoire. *Compt. rend.*, **26**, 535–538.

Shapiro, S. S. e Wilk, M. B. (1965). An analysis of variance test for normality (complete samples)†. *Biometrika*, **52**(3-4), 591–611.

Smirnov, N. (1948). Table for Estimating the Goodness of Fit of Empirical Distributions. *The Annals of Mathematical Statistics*, **19**(2), 279 – 281.

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, **33**(1), 1 – 67.

Vavilin, M. e Fernandez-Corbaton, I. (2022). Multidimensional measures of electromagnetic chirality and their conformal invariance. *New Journal of Physics*, **24**(3), 033022.

Weyl, H. (1952). *Symmetry*. Princeton Science Library. Princeton University Press. ISBN 9780691023748.

White, A. D. (2022). Deep learning for molecules and materials. *Living Journal of Computational Molecular Science*, **3**(1), 1499.

Zabrodsky, H. e Avnir, D. (1995). Continuous symmetry measures. 4. chirality. *Journal of the American Chemical Society*, **117**, 462–473.

# Apêndice A

# Código de obtenção do dataset aquiral

```
import pandas as pd
import requests
import xml.etree.ElementTree as ET
from alive_progress import alive_bar

def get_chembl_data(offset: int) -> ET:
    try:
        data = requests.get(f'https://www.ebi.ac.uk/chembl/api/data/molecule?limit=
        data.raise_for_status()
        tree = ET.fromstring(data.text)
        return tree
    except Exception as e:
        return None

def get_smiles_chirality_from_xml_document(xml_document: ET):
    try:
        # molecule/molecule_structures/canonical_smiles
        return (
            xml_document.find('molecule_structures/canonical_smiles').text,
            xml_document.find('chirality').text,
            xml_document.find('molecule_chembl_id').text
```

```
        )
    except Exception as e:
        return None, -1, -1

if __name__ == '__main__':
    molecules = []
    i = 0
    with alive_bar(int(6360)) as bar:
        while len(molecules) < 6360:
            xml_tree = get_chembl_data(i)
            '''<response>
<molecules>
<molecule>'''
            xml_tree = xml_tree.findall('molecules/molecule')
            for xml_data in xml_tree:
                smiles, chirality_tag, molecule_chembl_id = get_smiles_chirality_from_xm
                if smiles is None:
                    continue
                molecules.append((smiles, molecule_chembl_id))
                bar()
                if i % 100 == 0:
                    dt = pd.DataFrame(molecules, columns=['smiles', 'molecule_chembl_id'
                    dt.to_csv('achiral_dataset.csv', index=False)
            i+=20

    dt = pd.DataFrame(molecules, columns=['smiles'])
    dt.to_csv('achiral_dataset.csv', index=False)
```

# Apêndice B

# Código de para a filtragem de molecúlas sem angulos dihedrais

```
import pandas as pd
import networkx as nx
import rdkit

def get_all_paths(G, N = 3):
    # adapted from: https://stackoverflow.com/questions/28095646/finding-all-paths-

    def findPaths(G,u,n):
        if n==0:
            return [[u]]
        paths = [[u]+path for neighbor in G.neighbors(u) for path in findPaths(G,ne
        return paths

    allpaths = []
    for node in G:
        allpaths.extend(findPaths(G,node,N))

    return allpaths

def verify_if_has_dihedral_angle_or_more(row):
    adj = rdkit.Chem.GetAdjacencyMatrix(row['rdkit_mol_cistrans_stereo'])
```

```
    graph = nx.from_numpy_array(adj, parallel_edges=False, create_using=None)
    distance_paths, angle_paths, dihedral_paths = get_all_paths(graph, N = 1), get_all_p

    return len(dihedral_paths) != 0


test_final_RSA = pd.read_pickle("test_final_RSA_class.pkl")
train_final_RSA = pd.read_pickle("train_final_RSA_class.pkl")
validation_final_RSA = pd.read_pickle("validation_final_RSA_class.pkl")

filtered_test_final_RSA = test_final_RSA[test_final_RSA.apply(verify_if_has_dihedral_ang
filtered_train_final_RSA = train_final_RSA[train_final_RSA.apply(verify_if_has_dihedral_
filtered_validation_final_RSA = validation_final_RSA[validation_final_RSA.apply(verify_i

print(filtered_test_final_RSA.shape[0])
print(filtered_train_final_RSA.shape[0])
print(filtered_validation_final_RSA.shape[0])

filtered_test_final_RSA = filtered_test_final_RSA.reset_index(drop=True)
filtered_train_final_RSA = filtered_train_final_RSA.reset_index(drop=True)
filtered_validation_final_RSA = filtered_validation_final_RSA.reset_index(drop=True)

filtered_test_final_RSA.to_pickle('test_final_RSA_class.pkl')
filtered_train_final_RSA.to_pickle('train_final_RSA_class.pkl')
filtered_validation_final_RSA.to_pickle('validation_final_RSA_class.pkl')
```

# Apêndice C

# Código de obtenção do dataset CCM

```
from functools import wraps
import signal
import traceback
import pandas as pd
import numpy as np
from typing import Callable, List
from rdkit import Chem
import numpy as np
from rdkit.Chem import Mol, AllChem
from cosymlib import Molecule, Geometry
from multiprocessing import Pool
from multiprocessing.context import TimeoutError
from pymongo import MongoClient

def get_geo_symmetry(symmetry: str, coords: np.array, atom_ids: List[str], bonds: L
    # Define geometry
    geometry = Geometry(positions=coords.tolist(),
                        symbols=atom_ids.tolist(),
                        connectivity=bonds)

    # Geometrical symmetry measure
    sym_geom_measure = geometry.get_symmetry_measure(symmetry, central_atom=1)
    return sym_geom_measure
```

```
def get_3d_geometry_with_atom_type_identifiers(mol):
    conf = mol.GetConformer()
    coords = np.array([conf.GetAtomPosition(i) for i in range(mol.GetNumAtoms())])
    atom_ids = np.array([atom.GetSymbol() for atom in mol.GetAtoms()])
    bonds = set()
    for i, atom in enumerate(Mol.GetAtoms(mol)):
        for bond in atom.GetBonds():
            indexes = [bond.GetBeginAtomIdx(), bond.GetEndAtomIdx()]
            bonds.add((min(indexes)+1, max(indexes)+1))
    return coords, atom_ids, list(bonds)


def get_cs_symmetry_from_row(row):
    mol = row['rdkit_mol_cistrans_stereo']
    print (Chem.MolToSmiles(mol))
    try:
        symetry = get_geo_symmetry('Cs', *get_3d_geometry_with_atom_type_identifiers(mol
        return symetry
    except Exception as exc:
        traceback.print_exc()
        return None

client = MongoClient(host="localhost", port=27017, username="root", password="MongoDB201
collection = client.ttc.ccm
num_partitions = 1  # Number of partitions to split dataframe

test_final_RSA = pd.read_pickle("./test_final_RSA.pkl")
train_final_RSA = pd.read_pickle("./train_final_RSA.pkl")
validation_final_RSA = pd.read_pickle("./validation_final_RSA.pkl")

total_number_of_conformers = test_final_RSA.shape[0] + train_final_RSA.shape[0] + valida
```

```
test_final_RSA['Index'] = test_final_RSA.index
train_final_RSA['Index'] = train_final_RSA.index
validation_final_RSA['Index'] = validation_final_RSA.index

def process_df(df, pool_size, timeout, collection, df_type):

    for i in range(0, len(df), pool_size):
        with Pool(processes=pool_size) as pool:
            df_block = df.iloc[i:i + pool_size]
            to_insert_pre_result = []

            futures = []
            for _, rows in df_block.iterrows():
                mol = collection.find_one({
                    'Index': rows['Index'],
                    'df_type': df_type
                })
                if mol is None:
                    futures.append(pool.apply_async(get_cs_symmetry_from_row, (rows
                    to_insert_pre_result.append({
                        'Index': rows['Index'],
                        'ID': rows['ID'],
                        'df_type': df_type,
                        'SMILES_nostereo': rows['SMILES_nostereo'],
                        'CCM': None
                    })
                else:
                    print ('exists')
            for idx, future in enumerate(futures):
                try:
                    if timeout != -1:
                        sym = future.get(timeout=timeout)
                    else:
```

```
                        sym = future.get()
                    to_insert = to_insert_pre_result[idx]
                    to_insert['CCM'] = sym
                    collection.insert_one(to_insert)
                except TimeoutError:
                    print (f'timeout {timeout}')


for timeout in (1, 3, 5, 10, -1):
    process_df(test_final_RSA, 4, timeout, collection, 'test')
    process_df(validation_final_RSA, 4, timeout, collection, 'validation')
    process_df(train_final_RSA, 4, timeout, collection, 'train')
```