Trabalho de Conclusão de Curso

MassFormer: Uma Abordagem de Aprendizado Profundo para a Predição de Espectros de Massa

Anderson Miguel Clemente Santos
amcs@ic.ufal.br

Orientador:

Prof. Dr. André Luiz Lins de Aquino

Maceió, Julho de 2024

Anderson Miguel Clemente Santos

MassFormer: Uma Abordagem de Aprendizado Profundo para a Predição de Espectros de Massa

Monografia apresentada como requisito parcial para obtenção do grau de Bacharel em Engenharia de Computação do Instituto de Computação da Universidade Federal de Alagoas.

Orientador:

Prof. Dr. André Luiz Lins de Aquino

Maceió, Julho de 2024

Catalogação na fonte
Universidade Federal de Alagoas
Biblioteca Central
Divisão de Tratamento Técnico
Bibliotecária: Girlaine da Silva Santos – CRB-4 – 1127

S237m Santos, Anderson Miguel Clemente.
MassFormer: uma abordagem de aprendizado profundo para a predição de espectros de massa / Anderson Miguel Clemente Santos. – 2024.
54 f. : il. color.

Orientador: André Luiz Lins de Aquino.
Monografia (Trabalho de Conclusão de Curso em Engenharia de Computação : Bacharelado) - Universidade Federal de Alagoas. Instituto de Computação, Maceió, 2024.

Bibliografia: f. 51- 54.

1. Espectrometria de massa. 2. Redes neurais (Computação). 3. Predição de espectros. 4. Modelagem molecular. 5. Deep Learning. I. Título.

CDU: 004.032.26

UNIVERSIDADE FEDERAL DE ALAGOAS/UFAL
Instituto de Computação – IC
Campus A. C. Simões - Av. Lourival de Melo Mota, BL 12
Tabuleiro do Martins, Maceió/AL - CEP: 57.072-970
Telefone: (082) 3214-1401

Trabalho de Conclusão de Curso - TCC

Formulário de Avaliação

Nome do Aluno
Anderson Miguel Clemente Santos

Nº de Matrícula
18210850

Título do TCC (Tema)
MassFormer: Uma Abordagem de Aprendizado Profundo para a Predição de Espectros de Massa

Banca Examinadora

André Luiz Lins de Aquino
Nome do Orientador

ANDRE LUIZ LINS DE AQUINO:03235015400
Assinado de forma digital por ANDRE LUIZ LINS DE AQUINO:03235015400
Dados: 2024.08.09 19:51:27 -03'00'
Assinatura

Danilo Fernandes
Nome do Professor

gov.br Documento assinado digitalmente
DANILO FERNANDES COSTA
Data: 11/08/2024 12:50:47-0300
Verifique em https://validar.iti.gov.br
Assinatura

Jordão Alvez
Nome do Professor

gov.br Documento assinado digitalmente
JOSIAS JORDAO ANDRADE ALVES
Data: 10/08/2024 15:47:13-0300
Verifique em https://validar.iti.gov.br
Assinatura

Data da Defesa		Nota Obtida
09/08/2024		10 (dez)

Observações:

Coordenador do Curso
De Acordo

gov.br Documento assinado digitalmente
JOBSON DE ARAUJO NASCIMENTO
Data: 22/08/2024 15:18:22-0300
Verifique em https://validar.iti.gov.br
Assinatura

Agradecimentos

Eu gostaria de expressar minha profunda gratidão e apreço a todos que contribuíram para a realização deste trabalho e para meu desenvolvimento pessoal e acadêmico ao longo de minha jornada.

Em primeiro lugar, um agradecimento muito especial aos meus pais, Adalberon e Joana Clemente. O amor, suporte e a força que vocês me proporcionaram foram essenciais para superar todas as adversidades que enfrentei até aqui. Vocês são a base de tudo.

Agradeço imensamente aos meus colegas de curso, que tornaram toda essa jornada acadêmica mais enriquecedora e suportável. Um agradecimento especial ao Pedro Henrique e Rafael Galhos, cujas habilidades de programação foram inestimáveis; e à Ester de Lima e Rafael Augusto, cuja experiência em pesquisa e brilhantismo acadêmico foram uma constante fonte de inspiração. Vocês são mais que colegas, são amigos que levarei para a vida.

Meu sincero agradecimento ao professor Dr. André Aquino, por abrir portas para o conhecimento e colaborar diretamente no desenvolvimento deste trabalho junto ao Laccan/Orion. Sua orientação foi fundamental para a realização deste projeto.

Agradeço profundamente à banca examinadora, que gentilmente aceitou o convite para avaliar este trabalho. Sua disposição em participar deste processo é grandemente apreciada e valorizada.

Um agradecimento especial também aos professores Tiago Vieira e Thales, que me motivaram e ajudaram a encontrar minha paixão na computação, especialmente no campo da inteligência artificial. Seu entusiasmo e dedicação foram cruciais para minha formação.

Agradeço ao Instituto de Computação e ao corpo docente altamente qualificado, que oferece um ambiente de excelência acadêmica. Agradeço especialmente aos professores da engenharia, Thiago Cordeiro, Ícaro Araújo e Erick Barbosa, cuja dedicação e capacidade de inspirar os estudantes são essenciais, especialmente diante dos desafios e da complexidade dos cursos de engenharia.

Estendo meus agradecimentos ao corpo secretariado do Instituto de Computação, especialmente ao Marcelo, por toda a agilidade e eficiência na condução dos processos administrativos e burocráticos. Seu trabalho facilitou muitas das etapas essenciais para a conclusão deste curso.

Por último, mas não menos importante, agradeço a todos que passaram pela minha vida e deixaram sua marca de alguma forma. Cada conversa, cada gesto de apoio e cada momento compartilhado contribuíram significativamente para meu crescimento social, acadêmico e indi-

vidual.

A todos vocês, meu sincero obrigado.

Anderson Clemente

"Deus,
Conceda-me a serenidade
Para aceitar aquilo que não posso mudar,
A coragem para mudar o que me for possível
E a sabedoria para discernir entre as duas."

– Reinhold Niebuhr

Resumo

Este trabalho apresenta uma investigação detalhada do uso do MassFormer, um modelo baseado em transformer, para a predição de espectros de massa a partir de estruturas moleculares. Utilizando o conjunto de dados público MassBank of North America (MoNA), este estudo explora a eficácia do modelo MassFormer, que é refinado a partir de um modelo Graphormer pré-treinado. O Graphormer, conhecido por sua capacidade de modelar eficientemente dados estruturados como grafos, fornece uma base sólida para o MassFormer ao capturar as interações complexas entre os nós representativos dos átomos em uma molécula.

No decorrer deste trabalho, foram implementadas e comparadas duas abordagens de divisão dos dados: uma divisão aleatória simples e uma mais complexa baseada em Scaffold. Ambas as estratégias produziram resultados similares, demonstrando a robustez do modelo em condições variadas. A performance do MassFormer, embora inferior à dos modelos comerciais devido às limitações nos dados públicos disponíveis, mostrou-se promissora, especialmente considerando a possibilidade de melhoria contínua através da contribuição comunitária para o banco de dados MoNA.

Este estudo não apenas demonstra a aplicabilidade do MassFormer em contextos acadêmicos e de pesquisa, mas também destaca o potencial para futuras melhorias e expansões na modelagem de espectros de massa, promovendo uma compreensão mais profunda e a aplicação prática em espectrometria de massa.

Palavras-chave: Espectrometria de Massa; Predição de Espectros; Modelagem Molecular; Deep Learning

Abstract

This work presents a detailed investigation into the use of the MassFormer, a transformer-based model, for predicting mass spectra from molecular structures. Utilizing the public dataset MassBank of North America (MoNA), this study explores the efficacy of the MassFormer model, which is refined from a pre-trained Graphormer model. Graphormer, known for its ability to efficiently model structured data like graphs, provides a solid foundation for the MassFormer by capturing the complex interactions between nodes representing atoms in a molecule.

Throughout this work, two approaches to data splitting were implemented and compared: a simple random split and a more complex Scaffold-based split. Both strategies produced similar results, demonstrating the robustness of the model under varied conditions. The performance of the MassFormer, although inferior to commercial models due to the limitations in publicly available data, proved promising, especially considering the potential for continuous improvement through community contributions to the MoNA database.

This study not only demonstrates the applicability of the MassFormer in academic and research contexts but also highlights the potential for future improvements and expansions in mass spectrum modeling, promoting a deeper understanding and practical application in mass spectrometry.

Keywords: Mass Spectrometry; Spectrum Prediction; Molecular Modeling; Deep Learning

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

MS	Espectrometria de Massa
m/z	Relação Massa/Carga
MS/MS	Espectrometria de Massa em Tandem
MoNA	MassBank of North America
GNN	Redes Neurais Baseadas em Grafos (Graph Neural Networks)
ESI	Electrospray ionization (Ionização por Electrospray)
TOF	Time of Flight (Tempo de Voo)
ML	Machine Learning (Aprendizado de Máquina)
DL	Deep Learning (Aprendizado Profundo)
4IR	Quarta Revolução Industrial
NLP	Natural Language Processing (Processamento de Linguagem Natural)
GNN	Graph Neural Network (Rede Neural de Grafos)
MLP	Multilayer Perceptron (Perceptron Multicamadas)
DGL	Deep Graph Library
GCN	Graph Convolutional Network (Rede de Convolução de Grafos)
GAT	Graph Attention Network (Rede de Atenção de Grafos)
FAIR	Facebook AI Research
AI	Artificial Intelligence (Inteligência Artificial)
MACCS	Molecular ACCess System
ECFP	Extended-Connectivity Fingerprints
MHA	Multi-Head Self-Attention
MLP	Multi-Layer Perceptrons
NIST	National Institute of Standards and Technology

ChEMBL	Chemical Entities of Biological Interest Database
GrAFF	Graph neural network for Approximation via Fixed Formulas of Mass Spectra
3DMolMS	3D Molecular Network for Mass Spectra Prediction
OGB-LSC	Open Graph Benchmark Large-Scale Challenge

Conteúdo

1

Introdução

1.1 Contextualização e Motivação

A espectrometria de massa (MS) é uma técnica analítica indispensável para a identificação e quantificação de compostos químicos em amostras complexas. Através da MS, moléculas são ionizadas e analisadas com base na relação massa/carga (m/z), fornecendo espectros que são essenciais para o estudo em áreas como proteômica, metabolômica, forense e química ambiental. Em particular, a espectrometria de massa em tandem (MS/MS), que inclui uma etapa de fragmentação molecular, é uma ferramenta versátil para inferir estruturas moleculares a partir dos espectros de fragmentos (41).

No entanto, um dos principais desafios no uso da MS/MS é a capacidade de simular com precisão a fragmentação de moléculas dentro do espectrômetro.

As simulações atuais baseadas em princípios fundamentais da física teórica, embora ofereçam ferramentas para compreender o processo de fragmentação no espectrômetro de massas, são demasiadamente lentas e dependem de aproximações que limitam sua precisão. Isso representa um problema fundamental para o campo da espectrometria de massa, restringindo a capacidade de analisar eficientemente os dados coletados. Este é um problema fundamental na espectrometria de massa, pois restringe nossa habilidade de analisar e entender completamente os espectros gerados, essencial para aplicações em áreas como a descoberta de biomarcadores e o desenvolvimento de novos medicamentos (41).

Uma abordagem promissora para superar essas limitações é o uso de modelos preditivos de espectros, que podem aumentar significativamente a cobertura das bibliotecas de referência usadas para identificação de compostos. Tradicionalmente, as bibliotecas de espectros são limitadas e contêm uma fração pequena dos metabolitos conhecidos, o que dificulta a identificação de novas moléculas. A geração in silico de espectros a partir de estruturas moleculares conhecidas se apresenta como uma solução viável para ampliar o escopo das bibliotecas existentes (23).

Recentemente, o avanço das técnicas de aprendizado profundo, especialmente as que utilizam redes neurais baseadas em grafos, oferece novas possibilidades para abordar a predição de espectros de MS/MS de forma mais eficiente. O "MassFormer", um modelo que utiliza grafo transformers, é capaz de modelar interações entre todos os nós do grafo molecular, capturando as propriedades topológicas essenciais para prever com precisão os espectros de massa. Esta abordagem é capaz de superar muitas das limitações dos métodos anteriores, que lutavam para modelar as interações globais entre átomos distantes na molécula (41) resultando em predições imprecisas dos espectros de massa.

Neste contexto, o trabalho desenvolvido visa explorar a capacidade do "MassFormer"para traduzir estruturas moleculares em espectros de massa preditos, oferecendo uma ferramenta valiosa para aprimorar a análise espectrométrica. Esta pesquisa contribui para o desenvolvimento tecnológico na espectrometria de massa, potencializando a identificação de moléculas complexas em amostras biológicas, abrindo novos caminhos para a descoberta científica e aplicação prática em diversas áreas da ciência (41; 23).

1.2 Objetivos

Este trabalho visa principal investigar a eficácia e aplicabilidade do *MassFormer*, um modelo avançado baseado em grafo transformer, para a predição de espectros de massa. Pretende-se avaliar o desempenho do modelo em um cenário realista, expandindo a utilidade deste para enriquecer bibliotecas de espectros existentes. Além disso, almeja-se explorar a potencialidade deste software como uma ferramenta de identificação de compostos químicos acessível, ideal para implementação em universidades públicas. Dessa forma, o estudo busca contribuir para o avanço das técnicas de análise espectrométrica, fornecendo uma solução eficiente para pesquisadores e acadêmicos no campo da química analítica.

1.2.1 Objetivos específicos

Este trabalho possui os seguintes objetivos específicos:

- Identificar o espectro de massa correspondente ou aproximado;
- Avaliar o modelo após a identificação;

1.3 Estrutura do Trabalho

Este trabalho está estruturado em cinco capítulos principais, cada um abordando aspectos distintos do estudo realizado.

- **Capítulo 2 - Fundamentação Teórica:**Este capítulo fornece os conceitos básicos e essenciais tanto da espectrometria de massa quanto das técnicas de aprendizado de máquina e aprendizado profundo relevantes para este estudo. Aborda-se também conceitos de transformers e redes neurais de Grafo (GNNs), assim como o Graphormer. Também são introduzidas as ferramentas e bibliotecas fundamentais como a Deep Graph Library e o PyTorch, que suportam o desenvolvimento e a implementação do modelo proposto.

- **Capítulo 3 - Trabalhos Relacionados:** Revisa a literatura existente sobre a predição de espectros de massa, explorando diferentes abordagens e modelos utilizados anteriormente na área. Este capítulo destaca como o *MassFormer* se situa no contexto dos avanços recentes em aprendizado de máquina para espectrometria de massa.

- **Capítulo 4 - Metodologia:** Descreve detalhadamente a metodologia utilizada neste estudo, incluindo a preparação dos dados, a arquitetura do modelo *MassFormer*, e as estratégias de treinamento e validação. Este capítulo também detalha o uso do MoNA como fonte de dados e como os dados foram processados para treinar o modelo.

- **Capítulo 5 - Resultados & Discussão:** Apresenta os resultados obtidos com a aplicação do modelo *MassFormer*, analisando sua desempenho em diversas métricas de avaliação. Discussões são feitas comparando os resultados das predições com espectros reais e explorando as implicações dos achados no contexto da identificação de compostos químicos.

- **Capítulo 6 - Conclusão:** Conclui o trabalho, resumindo os principais achados e contribuições da pesquisa. Discute as limitações do estudo atual e sugere direções para pesquisas futuras, visando aprimoramentos no modelo e sua aplicação em contextos práticos mais amplos.

Cada capítulo foi estruturado de maneira a fornecer uma compreensão clara e aprofundada sobre cada etapa do processo de pesquisa, desde a teoria até a aplicação prática e avaliação do modelo desenvolvido.

2

Fundamentação Teórica

2.1 Espectrometria de Massa

A Espectrometria de Massa (MS) é uma técnica analítica de identificação, caracterização e quantificação de compostos químicos Watson and Sparkman (37). Seu desenvolvimento remonta ao trabalho de J.J. Thomson no final do século XIX, que demonstrou a separação de partículas com base em massas e cargas elétricas por campos elétricos e magnéticos (33). A partir dos anos 1940 e 1950, a MS avançou significativamente com a introdução de analisadores de massas de setor magnético e de tempo de voo, permitindo a determinação precisa de massas atômicas e a análise isotópica de gases nobres.

Este processo começa com a conversão das moléculas da amostra em partículas carregadas. Essas partículas são então aceleradas em um campo elétrico e passam por um analisador de massa, que as separa com base em suas razões m/z. O resultado é um espectro de massa que exibe a abundância dessas partículas em função de seu m/z, com o eixo x frequentemente rotulado como 'massa' em lugar de m/z (13).

O espectro de massa, representado em termos de Daltons (Da) por unidade de carga, é fundamental para identificar a composição química de uma amostra. A análise começa com a geração de partículas carregadas, onde, por exemplo, a técnica de ionização por electrospray (ESI) cria íons em pressão atmosférica ao passar uma amostra líquida através de um pequeno capilar que possui um potencial elétrico significativo em relação a um eletrodo oposto. Essa técnica produz um aerossol de gotículas carregadas que, eventualmente, perde seu solvente, liberando as partículas que são direcionadas para o analisador de massa (13).

Após essa etapa, as partículas carregadas são aceleradas por um campo elétrico. A velocidade com que cada partícula atravessa o analisador de massa depende de sua relação massa/carga. Diferentes tipos de analisadores, como o tempo de voo (TOF) e o quadrupolo, são utilizados para medir essa relação com precisão, permitindo a identificação de compostos com

base em suas propriedades espectrométricas únicas. O detector então registra a abundância de cada partícula, produzindo um espectro que pode ser usado para inferir a estrutura química dos analitos (13).

Essa tecnologia facilitam a identificação de proteínas e outros biomarcadores em misturas complexas, mas também é crucial para a triagem rápida de interações droga-alvo. Por exemplo, a MS pode ajudar a determinar a força e os locais de ligação entre um medicamento e seu alvo, uma aplicação crítica na fase de design e otimização de novos medicamentos. A MS é particularmente valiosa por sua habilidade em realizar análises rápidas e detalhadas, tornando-se uma ferramenta indispensável em laboratórios de pesquisa e desenvolvimento farmacêutico (13).

A figura 1 ilustra o processo básico da espectrometria de massa. Inicialmente, a amostra é introduzida através do injetor e vaporizada. Um feixe de elétrons ioniza as moléculas da amostra, convertendo-as em íons. Um campo elétrico acelera esses íons, que então passam por uma série de lentes que focam o feixe. Ao atravessar um campo magnético, os íons sofrem desvios proporcionais às suas razões massa/carga (m/z), permitindo sua separação antes de atingirem o detector. O detector mede a abundância de cada íon, gerando um espectro que possibilita a determinação da composição química da amostra. Através desse processo, é possível realizar análises detalhadas da estrutura molecular e identificar as substâncias presentes na amostra.

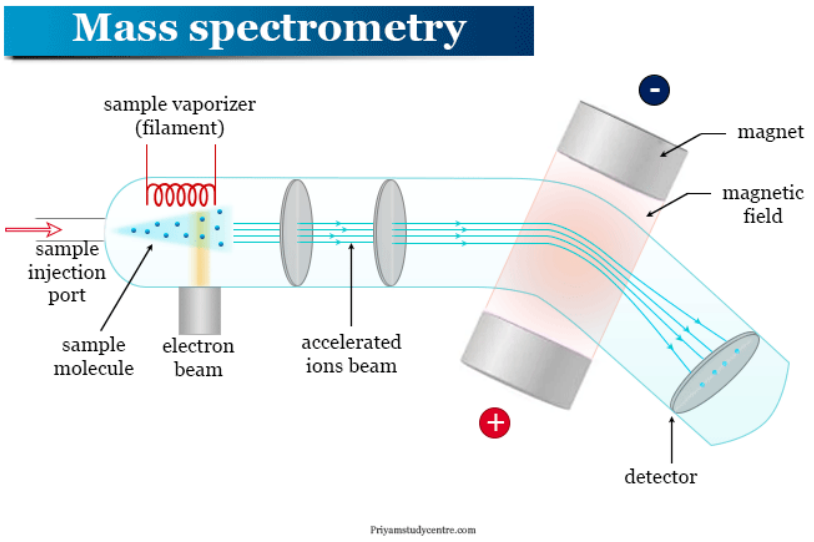

Figura 1: Espectrometria de massa. Fonte: Priyam Study Centre (28).

A figura 2 ilustra um espectro de massa típico obtido por espectrometria de massa. Neste gráfico, o eixo horizontal (x) representa a razão massa/carga (m/z) dos íons detectados, enquanto o eixo vertical (y) mostra a abundância relativa desses íons. Cada pico no espectro corresponde a um íon específico, onde a posição do pico indica a razão m/z do íon e a altura do pico reflete a abundância desse íon na amostra analisada. Os rótulos em cada pico, como "z=2"ou "z=3", indicam o estado de carga do íon, que é crucial para a determinação da massa molecular real

das moléculas na amostra. Este tipo de análise é fundamental para identificar e caracterizar compostos em misturas complexas, permitindo uma análise detalhada da composição química de amostras biológicas, ambientais ou químicas.

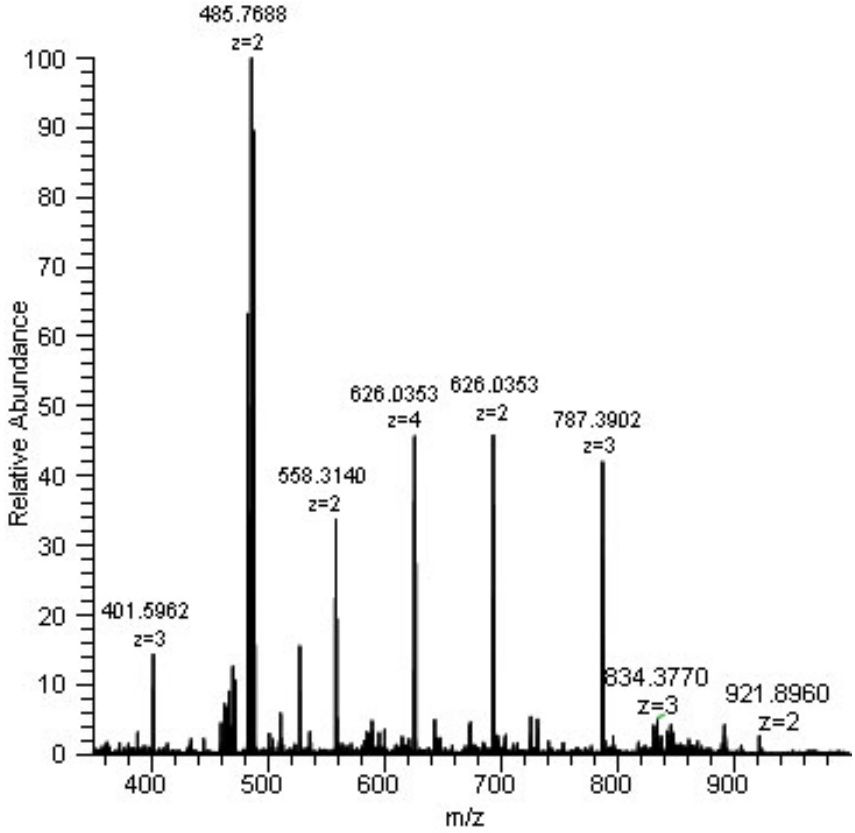

Figura 2: Espectro de massa. **(author?)** (Broad Institute).

2.2 Aprendizado de Máquina

2.2.1 Visão geral

Aprendizado de máquina é uma área da ciência da computação focada no desenvolvimento de algoritmos que aprendem a partir de dados e fazem previsões ou tomam decisões baseadas nesses dados. Como destacado por Mohri et al. (24), esses métodos computacionais utilizam experiências passadas, geralmente na forma de dados eletrônicos, para melhorar o desempenho ou realizar previsões acuradas. A eficácia desses algoritmos depende fortemente da qualidade e do volume dos dados disponíveis, integrando conceitos de estatística, probabilidade e otimização para analisar e interpretar as informações coletadas.

As tarefas de aprendizado de máquina são variadas e abrangem diversos aspectos de análise de dados, incluindo, segundo Mohri et al. (24):

- Classificação: Consiste em atribuir uma categoria a cada item. Exemplos incluem a classificação de documentos em categorias como política ou esportes e a classificação de imagens em categorias como carros ou trens.
- Regressão: Focada em prever um valor real para cada item, como o preço de ações ou a variação de variáveis econômicas. A precisão é crítica aqui, pois o custo de uma previsão incorreta depende da magnitude do erro.
- Ranking: Esta tarefa envolve ordenar itens de acordo com algum critério. Um exemplo clássico é o ranking de páginas da web em resultados de busca.
- Agrupamento (Clustering): Visa particionar um conjunto de itens em subconjuntos homogêneos. É amplamente usado na análise de redes sociais para identificar comunidades ou grupos naturais.
- Redução de Dimensionalidade (Manifold Learning): Esta tarefa transforma a representação de itens de uma forma de alta dimensão para uma de menor dimensão, preservando certas propriedades essenciais, útil em visão computacional e outras áreas que lidam com grandes volumes de dados.

2.2.2 Tipos de aprendizado de máquina

Géron (12) categoriza os sistemas de aprendizado de máquina com base em diversos critérios, incluindo o tipo de supervisão que recebem durante o treinamento. O aprendizado supervisionado envolve treinar modelos com dados rotulados onde cada entrada de dados vem com a resposta correta, facilitando tarefas como classificação e regressão. Por outro lado, o aprendizado não supervisionado trabalha com dados não rotulados, explorando a estrutura intrínseca dos dados para realizar tarefas como agrupamento e redução de dimensionalidade. O aprendizado semissupervisionado combina os dois métodos anteriores usando grandes volumes de dados não rotulados com um pequeno conjunto de dados rotulados para melhorar a aprendizagem. Por fim, o aprendizado por reforço é uma abordagem distinta na qual o sistema, chamado de agente, aprende a tomar decisões observando as consequências de suas ações, ajustando suas estratégias para maximizar a recompensa ao longo do tempo (12). Cada tipo de aprendizado tem aplicações específicas e é comumente escolhido com base na natureza dos dados e dos objetivos do problema.

2.3 Aprendizado Profundo (Deep Learning)

Deep Learning (DL) é um subcampo avançado do aprendizado de máquina (ML) que se concentra na aprendizagem de representações de dados em várias camadas de abstração. Segundo Sarker (32), deep learning foi revitalizado e ganhou destaque com os trabalhos de Hinton

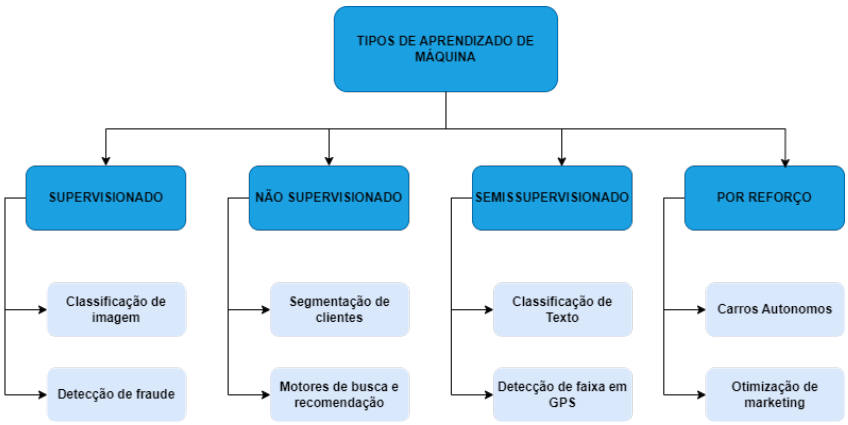

Figura 3: Tipos de aprendizado de máquina:.

et al. (17), levando a uma nova geração de redes neurais conhecidas por seu sucesso em uma ampla variedade de desafios de classificação e regressão. Essas redes são capazes de capturar relações complexas nos dados, graças às múltiplas camadas de processamento que imitam uma forma de processamento hierárquico de informações (32).

Deep learning diferencia-se significativamente dos métodos tradicionais de aprendizado de máquina em sua capacidade de operar eficientemente com grandes volumes de dados. Como destaca Francois (9), enquanto métodos tradicionais de aprendizado de máquina muitas vezes se concentram em aprender uma ou duas camadas de representação dos dados, o deep learning explora várias camadas, o que permite uma modelagem mais profunda e complexa da informação.

No contexto da Quarta Revolução Industrial (4IR), que se concentra em automação e sistemas inteligentes, o deep learning emerge como uma tecnologia central. Sua aplicabilidade se estende por várias áreas, incluindo saúde, reconhecimento visual, processamento de linguagem natural, e análise de sentimentos. Estas aplicações mostram como o deep learning pode transformar indústrias ao fornecer soluções que além de automatizarem processos, também proporcionam insights profundos e análises preditivas baseadas em grandes conjuntos de dados (32).

No entanto, a complexidade dos modelos de deep learning também apresenta desafios, especialmente no que diz respeito à sua natureza "caixa-preta". Isso significa que muitas vezes é difícil entender completamente como as decisões são feitas dentro do modelo, o que pode ser problemático em aplicações que exigem transparência e responsabilidade. Além disso, o treinamento de modelos de deep learning é computacionalmente intensivo, exigindo abundância de dados e poder de processamento, o que pode limitar sua implementação em dispositivos com recursos limitados.

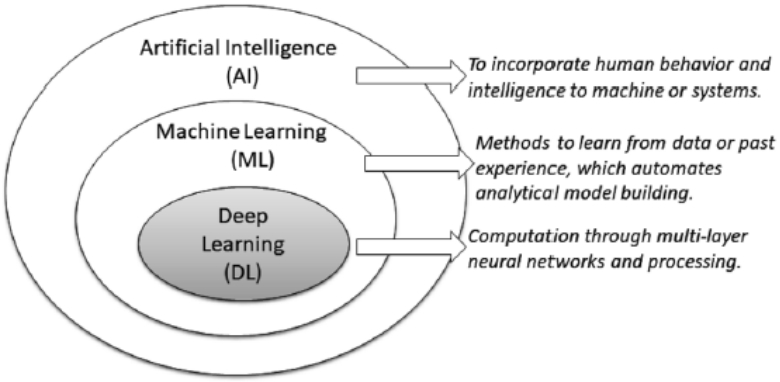

Figura 4: Uma ilustração da posição do aprendizado profundo (DL), comparando com o aprendizado de máquina (ML) e a inteligência artificial (IA). Fonte: Sarker (32)

2.3.1 Transformers

Os Transformers, introduzidos por Vaswani et al. (34), representam uma inovação significativa no campo do processamento de linguagem natural e aprendizado de máquina, ao substituírem modelos recorrentes tradicionais por uma arquitetura baseada inteiramente em mecanismos de atenção. Essa abordagem permite que o modelo processe dados de entrada em paralelo e capture dependências complexas, independentemente da distância entre os elementos na sequência de entrada. A arquitetura do Transformer é composta por dois componentes principais, o codificador (encoder) e o decodificador (decoder), cada um formado por múltiplas camadas que utilizam a atenção multi-cabeça e redes totalmente conectadas. Esse design facilita o aprendizado de representações profundas sem a necessidade de recorrência, possibilitando melhorias significativas em eficiência e desempenho em tarefas como tradução automática e modelagem de linguagem Vaswani et al. (34). Este modelo estabeleceu um novo padrão para modelos de sequência, influenciando uma ampla gama de aplicações e pesquisas subsequentes na área.

A figura 5 apresenta a arquitetura do Transformer, um modelo avançado usado principalmente para tarefas de processamento de linguagem natural. Essencialmente, o Transformer é dividido em duas partes principais: o codificador e o decodificador, cada um consistindo de múltiplas camadas idênticas. No codificador, a entrada é primeiro incorporada e somada com codificação posicional, passando por camadas que incluem atenção multi-cabeça e redes feed-forward, todas normalizadas para estabilizar o aprendizado. O decodificador opera de maneira similar, mas adiciona uma camada de atenção que usa as saídas anteriores para prever a próxima palavra, processo que também é normalizado e depois transformado em probabilidades através de uma camada linear e softmax. Este design permite processamento paralelo eficiente, superando modelos recorrentes tradicionais em eficiência e eficácia.

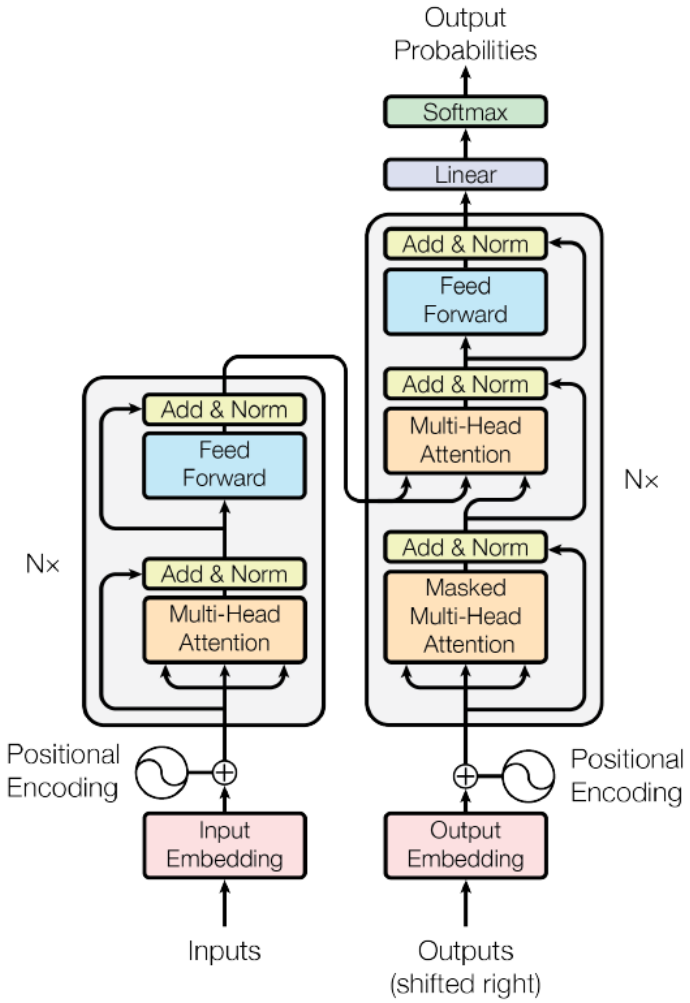

Figura 5: Arquitetura do Transformer. Fonte: Vaswani et al. (34)

2.3.2 Redes Neurais de Grafos (GNN)

As Redes Neurais de Grafos (GNNs) aproveitam a capacidade de manter as simetrias inerentes ao grafo, preservando as invariâncias de permutação enquanto processam atributos de nós, arestas e contexto global. As GNNs são projetadas para modificar os atributos internos dos elementos do grafo sem alterar sua estrutura de conectividade. A abordagem "graph-in, graph-out"das GNNs permite que um grafo seja inserido como entrada e que suas características sejam progressivamente transformadas, mantendo a integridade estrutural do grafo original (31).

Na prática, as GNNs aplicam modelos diferenciáveis, como o perceptron multicamada (MLP), a cada componente do grafo, incluindo nós, arestas e o vetor de contexto global. Este processo envolve a aprendizagem de novos embeddings para cada elemento do grafo, que são atualizados iterativamente através de múltiplas camadas de processamento. As camadas de GNN são empilhadas, similarmente às camadas em redes neurais convencionais, permitindo uma aprendizagem profunda e complexa das características dos dados. Consequentemente, o grafo de saída reflete a mesma estrutura de conectividade que o grafo de entrada, mas com

embeddings enriquecidos que incorporam uma compreensão mais profunda das interações e relações nos dados, conforme destacado por Sanchez-Lengeling et al. (31).

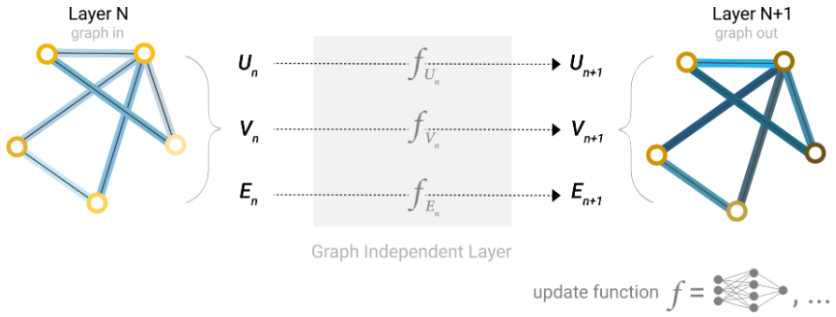

Figura 6: Camada única de uma GNN simples. Fonte: Sanchez-Lengeling et al. (31)

2.3.3 Graphormer

O Graphormer é uma abordagem inovadora que expande o uso do modelo Transformer, conhecido por sua eficácia em dados sequenciais como linguagem natural e visão computacional, para o domínio dos grafos (34). Embora os Transformers tenham se mostrado poderosos em diversas áreas, sua aplicação em grafos não era considerada padrão nos principais rankings de representação gráfica (40).

O Graphormer foi concebido para incorporar eficazmente a informação estrutural dos grafos, um aspecto crítico que os modelos tradicionais de Transformer não abordam diretamente. A estrutura de um grafo, que destaca a relação entre os nós e suas conexões, não é naturalmente adaptada pela arquitetura original do Transformer, que se concentra mais na semântica do que na estrutura.

Para resolver isso, o Graphormer introduz codificações estruturais que melhoram a capacidade do modelo de entender e processar grafos:

- **Codificação de Centralidade:** Esta codificação visa capturar a importância relativa de cada nó no grafo. Utiliza-se a centralidade de grau, atribuindo um vetor aprendível a cada nó com base em seu grau, que é adicionado às características do nó na camada de entrada.

- **Codificação Espacial:** Para modelar a relação estrutural entre os nós, o Graphormer emprega uma codificação espacial que considera a distância do caminho mais curto entre os nós. Esta distância é incorporada como um termo de viés no mecanismo de atenção softmax, facilitando a captura da dependência espacial no grafo.

- **Codificação de Arestas:** Além das relações nodais, o Graphormer também leva em conta as características das arestas, como o tipo de ligação em um grafo molecular. Para cada par de nós, calcula-se a média dos produtos internos das características das arestas e

embeddings aprendíveis ao longo do caminho mais curto, utilizando essas informações no módulo de atenção.

Essas codificações permitem que o Graphormer modele com precisão tanto as relações locais quanto as globais dentro de um grafo, superando muitas variantes populares de GNNs, que são consideradas casos especiais do Graphormer. Sua capacidade de incorporar de maneira eficaz a informação estrutural e semântica leva a um desempenho de ponta em uma ampla gama de tarefas de predição em nível de grafo, demonstrado em desafios como o Open Graph Benchmark Large-Scale Challenge (OGB-LSC) (20) e outros rankings populares de aprendizado de representação gráfica (21; 7).

2.4 Bibliotecas

2.4.1 Deep Graph Library (DGL)

A Deep Graph Library (DGL) é uma biblioteca de código aberto para Python que facilita o trabalho com dados em forma de grafos, especialmente no contexto do aprendizado profundo. Ela é desenvolvida para fornecer uma maneira eficiente, flexível e escalável de trabalhar com redes neurais de grafos (GNNs) e outras estruturas de dados baseadas em grafos (36).

1. Suporte a Múltiplos Backends: Uma das principais vantagens da DGL é sua compatibilidade com várias plataformas de aprendizado profundo, incluindo TensorFlow, PyTorch e MXNet. Isso oferece flexibilidade aos usuários para escolherem o framework de sua preferência, permitindo a integração fácil com projetos existentes e o aproveitamento de recursos específicos de cada plataforma.

2. Eficiência e Escalabilidade: A DGL é otimizada para manipular grafos grandes e complexos de maneira eficiente. Ela utiliza técnicas avançadas para minimizar o uso de memória e acelerar os cálculos, o que é essencial ao lidar com datasets volumosos e grafos esparsos, comuns em muitos problemas do mundo real.

3. Abstrações de Alto Nível: A biblioteca simplifica significativamente o processo de implementação de GNNs, oferecendo abstrações de alto nível para uma variedade de modelos de GNN, como Graph Convolutional Networks (GCNs) e Graph Attention Networks (GATs). Essas abstrações ajudam os usuários a se concentrarem mais nos aspectos experimentais dos modelos, reduzindo a complexidade do código e os erros de implementação.

4. Modularidade e Reusabilidade: Com módulos reutilizáveis e componentes pré-definidos, a DGL permite aos usuários construir e experimentar diferentes arquiteturas de GNN de maneira modular. Isso não só acelera o desenvolvimento de novos modelos como também facilita a experimentação e o teste de novas ideias de pesquisa em um ambiente controlado.

5. Aplicações Diversificadas: A DGL encontra aplicações em uma série de áreas, incluindo sistemas de recomendação, biologia computacional, detecção de fraudes e análise de redes sociais. Seu design flexível e capacidade de processar informações complexas tornam-na adequada para uma ampla gama de tarefas que envolvem a análise de dados relacionais.

6. Comunidade e Suporte: A biblioteca possui uma comunidade ativa e uma documentação extensiva, que oferece suporte abrangente através de tutoriais, exemplos de código e guias práticos. Isso é particularmente valioso para novos usuários e para aqueles que estão explorando o campo emergente das GNNs.

2.4.2 Pytorch

PyTorch é uma biblioteca de aprendizado de máquina de código aberto amplamente utilizada para aplicações de aprendizado profundo, como visão computacional e processamento de linguagem natural. Desenvolvida principalmente pela Facebook AI Research (FAIR), a biblioteca é conhecida por sua flexibilidade, velocidade e facilidade de uso. Com seu suporte robusto para computação dinâmica através de grafos de computação autogradáveis, o PyTorch permite aos usuários ajustar e experimentar arquiteturas de rede de maneira intuitiva e flexível. Sua integração nativa com Python e compatibilidade com CUDA para processamento em GPU o torna extremamente rápido para treinamento de modelos complexos. Além disso, PyTorch tem um ecossistema rico e ativo, com muitas bibliotecas auxiliares e uma comunidade grande que continua a contribuir com ferramentas e técnicas inovadoras.

Os princípios de design do PyTorch refletem um compromisso com a simplicidade, desempenho pragmático e uma forte orientação para atender às necessidades dos pesquisadores em machine learning. Primeiramente, o PyTorch é desenvolvido para ser "Pythonic", integrando-se perfeitamente com as ferramentas de programação, depuração e visualização de dados de Python, o que facilita sua adoção por cientistas de dados já familiarizados com o ambiente Python.

Além disso, o PyTorch coloca os pesquisadores em primeiro lugar, oferecendo APIs intuitivas que escondem a complexidade inerente ao aprendizado de máquina, permitindo que os usuários se concentrem na experimentação e inovação. Em termos de desempenho, o PyTorch adota uma abordagem pragmática, buscando um equilíbrio entre velocidade e facilidade de uso, aceitando complexidades adicionais internamente para fornecer desempenho robusto sem comprometer a simplicidade do usuário. Este compromisso é evidenciado pela disposição em trocar uma pequena quantidade de velocidade por uma significativa simplificação no uso, como descrito por Paszke et al. (27). Por fim, o princípio de "Worse is better"reflete a filosofia de manter a implementação interna simples, mesmo que isso signifique que algumas funcionalidades sejam menos abrangentes, permitindo uma adaptação mais rápida e eficaz às novas demandas e

avanços na IA. Esses princípios de design fazem do PyTorch uma escolha poderosa para projetos de aprendizado profundo que valorizam tanto a produtividade do desenvolvedor quanto a performance do modelo.

2.5 Revisão de Literatura

2.5.1 Método Quebra de Ligações (Bond-breaking)

A abordagem de quebra de ligações é considerada um dos métodos mais estudados para a previsão de espectros (1; 35; 30; 4). Esse método aborda o desafio de representar o espaço de saída enumerando as estruturas de todos os íons produto prováveis, considerando-os subgrafos conectados do precursor, gerados por sequências de remoções de arestas. Cada estrutura de íon produto é avaliada quanto à sua probabilidade de formação, e um espectro é gerado associando esta probabilidade ao m/z teórico de cada estrutura. A abordagem de quebra de ligações alcança uma resolução perfeita de m/z, mas enfrenta duas limitações principais: a primeira é que a enumeração de subestruturas escala de forma inadequada com o tamanho da molécula e não é compatível com implementações massivamente paralelas em GPUs. Wang et al. (35) demonstram que a previsão de um único espectro de massa leva, em média, cerca de 5 segundos, o que inviabiliza o treinamento em grandes conjuntos de dados, estimando-se três meses para treinar em aproximadamente 300 mil espectros do NIST-20 em uma máquina de 64 núcleos. A segunda limitação é decorrente de uma suposição modelar restritiva: reações de rearranjo, frequentemente, produzem íons produto que não podem ser alcançados a partir do precursor por sequências de remoções de arestas, o que impõe sérias restrições durante o teste, especialmente quando se requer a previsão de milhões de espectros, como no caso do uso de uma grande base de dados estruturais como a ChEMBL (11).

2.5.2 Método mass-binning

O método de "mass-binning"é empregado para a previsão de espectros por Wei et al. (38) e utilizado posteriormente em trabalhos mais recentes (43; 42). Este método representa o espectro de massa como um vetor de comprimento fixo através da discretização: o eixo m/z é dividido em compartimentos estreitos e regularmente espaçados, e a cada compartimento é atribuída a soma das intensidades de todos os picos que caem dentro de seus limites. A previsão de espectro se transforma, então, em um problema de regressão vetorial, que é adequado para implementação em GPU e escala melhor do que o método de quebra de ligações. No entanto, devido a um espaço alvo com milhões de compartimentos de massa ser demasiadamente grande, contagens realistas de compartimentos perdem informações essenciais de alta resolução sobre as fórmulas químicas dos picos, descartando uma vantagem chave da análise MS/MS em favor de um problema de aprendizado tratável. Tais modelos também são suscetíveis a efeitos de

borda, onde erros de medição de m/z do instrumento podem causar a transposição dos picos dos mesmos íons produto entre os limites dos compartimentos de um espectro para outro.

2.5.3 Métodos Avançados em Predição de Espectros de Massa

No campo de predição de espectros de massa, diversos preditores de espectro baseados em aprendizado profundo têm sido desenvolvidos, cada um explorando estratégias distintas. O método 3DMolMS (3D Molecular Network for Mass Spectra Prediction), por exemplo, gera previsões de espectros em baixa resolução utilizando informações tridimensionais da molécula, o que melhora a performance, especialmente em regimes de poucos dados (18). GrAFF (Graph neural network for Approximation via Fixed Formulas of Mass Spectra), por outro lado, modela espectros como uma distribuição sobre um vocabulário fixo de fórmulas, aproveitando a observação empírica de que a maioria dos picos em espectros de moléculas pequenas pode ser explicada por uma fórmula pequena comum ou uma fórmula maior menos comum (25). Finalmente, SCARF (Subformulae Classification for Autoregressively Reconstructing Fragmentations), também realiza a classificação de fórmulas, mas, em vez de usar um vocabulário fixo, parametriza uma distribuição sobre todas as subfórmulas possíveis da fórmula precursora (14). Esses métodos, embora promissores em resolução de predição e anotação de picos, dependem de heurísticas projetadas manualmente, o que pode limitar sua generalização para novos tipos de compostos, especialmente aqueles maiores onde as heurísticas podem não se aplicar.

Além disso, a dependência de procedimentos autorregressivos em modelos como o SCARF pode tornar esses métodos menos eficientes em termos de tempo de predição. Em contraste, o MassFormer, que é construído sobre a arquitetura robusta e flexível do Graphormer, não apenas evita a necessidade de intervenções manuais complexas, mas também promove uma implementação que é potencialmente mais rápida e adaptável a diferentes conjuntos de dados e condições experimentais. Essa combinação de eficiência computacional, generalização superior e integração flexível com diversas condições experimentais torna o MassFormer uma escolha promissora para a análise de espectros em larga escala, superando desafios comuns enfrentados por outras abordagens mais tradicionais.

3

Metodologia

3.1 Formulação do problema

A previsão de espectros pode ser conceituada como um problema de aprendizado supervisionado. Neste contexto, utiliza-se com um conjunto de dados $\{x_i, z_i, y_i\}_{i=1}^{n}$, onde x_i representa uma molécula e y_i o seu espectro sob condições experimentais z_i. O índice i identifica cada amostra individual no conjunto de dados, enquanto n representa o número total de amostras disponíveis. O objetivo principal é aprender os parâmetros θ da função de previsão $f_\theta : X \times Z \to Y$, na qual X é o espaço de compostos químicos e Y é o espaço dos espectros.

Os espectros de massa são caracterizados por um conjunto de picos, cada um com uma localização m/z e uma intensidade associada. Ao discretizar as localizações dos picos em m *bins* de largura fixa, um espectro de massa pode ser representado como um vetor esparso m-dimensional. Neste vetor, cada pico na localização j possui uma intensidade $y_j \geq 0$. Dessa forma, o problema de previsão de espectro é formulado como uma regressão vetorial, com $Y = \mathbb{R}_{\geq 0}^{m}$.

Adicionalmente, os metadados espectrais $z \in Z$, tais como a energia de colisão e o aduto precursor, são fornecidos como informações complementares à molécula de entrada x. Estes metadados ajudam a informar o modelo sobre as condições específicas sob as quais os espectros foram obtidos, permitindo que a previsão seja ajustada para refletir essas condições variáveis e melhorar a precisão e relevância dos resultados preditos.

3.2 Featurização Química

A featurização da molécula de entrada x é um aspecto crítico, pois influencia diretamente a estrutura da função de predição f_θ e pode impactar significativamente o desempenho do modelo. As representações moleculares, comumente chamadas de impressões digitais moleculares

ou descritores moleculares, são utilizadas para representar moléculas usando características químicas projetadas manualmente. Exemplos comuns dessas características incluem a presença de subestruturas pré-definidas, como as encontradas em MACCS (6), e contagens hash de subestruturas locais como em ECFP (29).

Além disso, as representações em grafos moleculares capturam a estrutura de uma molécula ao representar explicitamente átomos como nós e ligações como arestas. Os atributos dos nós podem codificar diversas propriedades químicas associadas ao átomo, como elemento, carga formal e número de hidrogênios ligados, enquanto os atributos das arestas podem incluir informações sobre o tipo de ligação e aromaticidade. Estas representações são particularmente adequadas para serem processadas por GNNs e graph transformers, como discutido na Seção 4.3, e tendem a ser mais expressivas do que as impressões digitais moleculares. Importante destacar que as representações gráficas não contêm informações estereoquímicas, de conformadores ou coordenadas 3D, uma vez que tais dados não estão disponíveis em nossos conjuntos de dados MS.

As características dos nós, arestas e metadados espectrais utilizados no MassFormer são detalhadas na Tabela 1.

Tipo de Característica	Nome da Característica	Valores
Nó	Elemento	C, N, O, P, S, F, Cl
Nó	Grau	0, ..., 10, 11+
Nó	Número de Hidrogênios	0, ..., 8, 9+
Nó	Hibridização Orbital	SP, SP2, SP3, SP3D, SP3D2, Outro
Nó	Aromaticidade	Verdadeiro, Falso
Nó	Membro do Anel	Verdadeiro, Falso
Aresta	Tipo de Ligação	Simples, Dupla, Tripla, Aromática, Outro
Aresta	Conjugação da Ligação	Verdadeiro, Falso
Metadados	Energia de Colisão Normalizada	(0, 200)
Metadados	Aduto Precursor	$[M+H]^+$, $[M+H\text{-}H_2O]^+$, $[M+H\text{-}2H_2O]^+$, $[M+2H]^{2+}$, $[M+H\text{-}NH_3]^+$, $[M+Na]^+$
Metadados	Massa do Precursor (Da)	(0, 1000)

Tabela 1: Featurização. Características de entrada fornecidas ao modelo MassFormer. As características dos nós capturam informações dos átomos, as características das arestas capturam informações das ligações, e as características de metadados capturam informações específicas do espectro.

3.3 Graph Transformer Químico

Os transformers (34) são uma família de redes neurais que se destacam pelo uso de mecanismos de atenção para modelar sequências. Desenvolvidos inicialmente para tradução automática

de máquinas, esses modelos demonstraram eficácia em uma variedade de campos, como visão computacional e aprendizado por reforço, alcançando desempenhos notáveis mesmo em problemas que não são típicos de modelagem sequencial.

Os grafos transformers, uma adaptação dos transformers tradicionais, foram propostos para modelar interações globais entre todos os nós de um grafo. Isso é motivado pela capacidade desses modelos de capturar interações parietais globais, em contraste com muitas redes neurais gráficas que tendem a modelar apenas relações locais em uma única camada e exigem profundidade significativa para capturar interações em distâncias maiores.

O Graphormer, uma implementação específica de grafo transformer, mostrou-se particularmente eficaz em tarefas de predição de propriedades de moléculas em química, destacando-se em contextos de dados limitados. Este modelo distingue-se pelo uso inovador de codificações posicionais que incorporam informações do caminho mais curto entre nós e os dados de aresta correspondentes, processados através do algoritmo de Floyd-Warshall (8).

O mecanismo de atenção no Graphormer calcula a importância relativa entre pares de nós baseando-se em matrizes de projeção aprendidas e distâncias codificadas entre os nós, complementadas por incorporações de aresta ao longo do caminho mais curto entre eles. Esse processo é detalhado pelas seguintes equações matemáticas:

$$a_{ij} = \text{softmax}\left(\frac{(W_Q h_i)^T (W_K h_j)}{\sqrt{d}} + b_{ij} + c_{ij}\right) \tag{3.1}$$

$$c_{ij} = \frac{1}{N} \sum_p w_p^T e_p \tag{3.2}$$

Onde h_i e h_j representam as representações vetoriais dos nós i e j, transformadas pelas matrizes de projeção W_Q e W_K para queries e keys, respectivamente. O escalar b_{ij}, indexado pela distância do caminho mais curto entre i e j, ajusta a atenção baseando-se na proximidade estrutural. O termo de incorporação de aresta c_{ij} é calculado como uma média ponderada das incorporações das arestas e_p ao longo do caminho mais curto, ponderadas por w_p, que são pesos aprendíveis. Essas incorporações de arestas capturam características específicas das conexões. O resultado da atenção é normalizado pela raiz quadrada da dimensão d antes de aplicar a função softmax, que converte o resultado em uma distribuição de probabilidade, facilitando a aprendizagem de relações complexas entre os nós.

Além disso, o modelo inclui um nó de leitura, similar ao token CLS usado em transformers de processamento de linguagem natural (5), que agrega e resume informações de todo o grafo para facilitar predições de propriedades em nível de grafo na camada final do transformer.

3.4 Base de dados

Neste projeto foi utilizado o MassBank of North America (MoNA) como base de dados dos espectros de massa e metadados (19). Segundo Galgonek and Vondrášek (10) MoNA é um repositório centrado em metadados e de autocuração de espectros de massa de metabólitos, metadados e compostos associados. MoNA integra dados de muitos outros conjuntos de dados, como LipidBlast, MassBank e GNPS, e atualmente contém aproximadamente 2.000.000 de espectros de cerca de 600.000 compostos químicos.

Neste estudo, para simplificar a análise, o foco foi exclusivamente em espectros gerados por instrumentos Orbitrap (15) que utilizam a técnica de colisão Higher Energy Collision Dissociation (HCD), uma escolha justificada pelo fato de representarem a maior parcela dos dados disponíveis no banco MoNA. Especificamente, a análise foi restrita a espectros em modo positivo, selecionando apenas aqueles que apresentam seis dos adutos precursores mais comuns.

Esses adutos foram selecionados com base em uma análise preliminar dos dados disponíveis, que revelou que [M+H]+, [M+H-H2O]+, [M+H-2H2O]+, [M+2H]2+, [M+H-NH3]+ e [M+Na]+ são os adutos mais frequentemente observados. Essa seleção focada permite uma investigação mais aprofundada das características e comportamentos mais frequentemente observados em amostras, facilitando a identificação e caracterização dos compostos. Após a filtragem dos dados, foram utilizados no total 13.225 espectros e 1.376 compostos químicos

3.5 Tecnologias Utilizadas

Na elaboração deste projeto, a escolha das tecnologias foi pautada pela necessidade de eficiência, flexibilidade e compatibilidade com os requisitos avançados de processamento de dados. Utilizamos Python 3.8 [1] devido à sua vasta biblioteca padrão e suporte robusto para ciência de dados e aprendizado de máquina. Esta versão do Python oferece melhorias significativas de desempenho e compatibilidade com as bibliotecas mais recentes.

O PyTorch [2] foi escolhido como a biblioteca principal de aprendizado profundo devido à sua interface intuitiva, flexibilidade e capacidade de execução dinâmica, que são ideais para pesquisa experimental e prototipagem rápida. A integração do PyTorch com CUDA 11.3 [3] permitiu a utilização eficiente de GPUs para acelerar os cálculos, essencial para o treinamento de modelos complexos como o MassFormer, que requer um grande volume de operações matemáticas e manipulação de dados.

Docker [4] foi utilizado para encapsular o ambiente de desenvolvimento, garantindo a reprodutibilidade e a consistência do projeto em diferentes plataformas e sistemas. Isso facilita

[1] https://www.python.org
[2] https://pytorch.org
[3] https://developer.nvidia.com/cuda-toolkit
[4] https://www.docker.com

a configuração e distribuição do ambiente necessário para executar o MassFormer, proporcionando uma maneira eficaz de gerenciar dependências e configurações.

A biblioteca MassFormer [5] foi escolhida especificamente para a tarefa de previsão de espectros de fragmentos a partir de moléculas, utilizando técnicas de grafo transformer para capturar as complexas relações estruturais em dados moleculares. Além disso, outras bibliotecas de ciência de dados, como scikit-learn[6], pandas[7], numpy[8], matplotlib[9] e seaborn[10], foram utilizadas para o pré-processamento de dados, análises estatísticas e visualizações. Estas ferramentas são essenciais para manipular grandes conjuntos de dados, realizar cálculos numéricos e apresentar os resultados de maneira clara e eficaz.

3.6 Arquitetura do Modelo

O MassFormer utiliza uma arquitetura de grafo transformer para prever espectros de fragmentos a partir de uma molécula de entrada, representada como um grafo molecular (42). Esta abordagem é visualmente resumida na figura 7. Inicialmente, o grafo de entrada é pré-processado em embeddings de nós e de arestas. Os embeddings de nós codificam informações químicas sobre os átomos, como o elemento e propriedades de centralidade, como o grau. Os embeddings de aresta capturam as relações espaciais entre os átomos na molécula, combinando informações sobre o comprimento do caminho mais curto e as arestas visitadas ao longo dele.

Após o pré-processamento, o processo de predição do espectro utilizando o grafo transformer pode ser descrito nas seguintes etapas:

1. **Extração de Embeddings:** Os embeddings de nós e arestas são extraídos a partir do grafo molecular de entrada. Esses embeddings codificam informações químicas sobre os átomos (como elemento e propriedades de centralidade) e relações espaciais entre os átomos (como comprimento do caminho mais curto e arestas visitadas).

2. **Aplicação do Grafo Transformer:** O grafo transformer aplica iterativamente autoatenção multi-cabeças (Multi-Head Self-Attention, MHA) e perceptrons multicamadas (Multi-Layer Perceptrons, MLP) para manipular os dados. Durante este processo, os pesos de autoatenção capturam associações globais entre todos os pares de nós, influenciados pelos embeddings de arestas em cada iteração.

3. **Resumo dos Embeddings:** Após várias iterações, os embeddings finais são resumidos em um único embedding que representa toda a molécula. Este embedding sintetiza a informação estrutural e química necessária para a predição do espectro.

[5] https://github.com/Roestlab/massformer
[6] https://scikit-learn.org/stable/
[7] https://pandas.pydata.org
[8] https://numpy.org
[9] https://matplotlib.org
[10] https://seaborn.pydata.org

4. **Adição de Metadados:** A representação química é combinada com metadados do espectro, que incluem informações sobre o precursor (como o aduto formado durante a ionização) e o instrumento (como a energia de colisão). Estes metadados são cruciais para influenciar o processo de fragmentação e o espectro resultante.

5. **Predição do Espectro:** O embedding combinado é passado para um MLP que realiza a predição do espectro na forma de um vetor positivo esparso. Cada dimensão do vetor representa uma localização de pico em caixas (*bins*), com a magnitude correspondendo à intensidade do pico.

6. **Inicialização e Ajuste Fino:** Os parâmetros dos embeddings de entrada e das camadas do grafo transformer são inicializados a partir de um modelo Graphormer pré-treinado em um grande conjunto de dados químicos. Em seguida, o modelo é ajustado conjuntamente com o preditor de espectro MLP utilizando o conjunto de dados MS/MS.

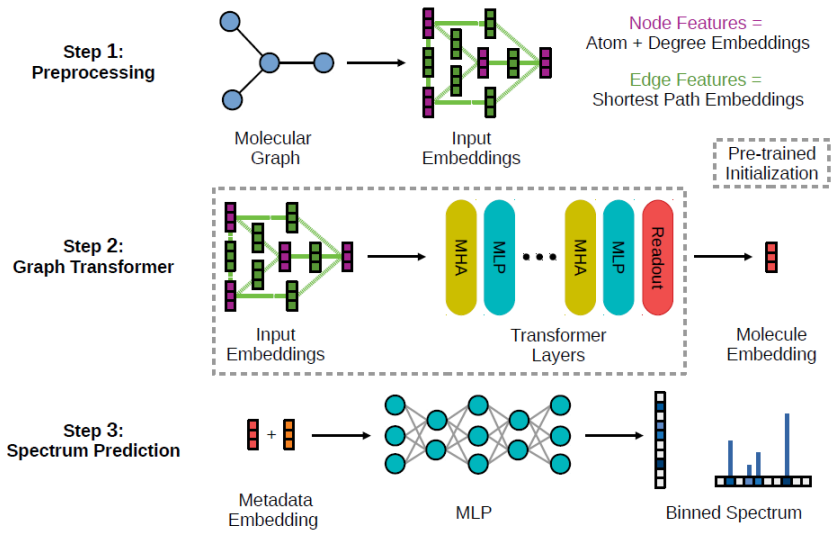

Figura 7: Arquitetura geral do massformer. Fonte: Young et al. (42)

3.7 Treinamento do Modelo

Inicialmente, o modelo base utilizado, o Graphormer, é treinado em uma ampla gama de dados de moléculas para aprender representações químicas genéricas utilizando o conjunto de dados PCQM4Mv2 (22; 26). Após esse treinamento preliminar, o MassFormer é refinado, especificamente para a tarefa de predição de espectros de massa. Este processo de refinamento ajusta os parâmetros do modelo às características particulares dos dados de espectrometria de massa.

Para a divisão dos dados em conjuntos de treinamento e teste, foram consideradas duas estratégias distintas: uma divisão aleatória simples por composto, utilizando a chave InChI (InChIKey) (16), e uma divisão mais desafiadora baseada em Scaffold, que estratifica os compostos por seu esqueleto Murcko antes de dividir (2). A estratégia de divisão por InChIKey [11] baseia-se em strings hash de representações químicas, funcionando essencialmente como uma divisão aleatória simples baseada na identidade do composto. Esta abordagem tem a vantagem de ser direta e preserva uma distribuição homogênea dos compostos através dos conjuntos de dados.

Por outro lado, a divisão por Scaffold utiliza os esqueletos Murcko, que agrupam compostos de maneira mais detalhada antes da divisão, assegurando que todos os compostos (e espectros associados) de um mesmo grupo sejam alocados na mesma partição. Este método introduz um desvio distribucional entre os dados de treino e teste, sendo comumente utilizado para avaliar modelos de aprendizado profundo em aplicações com pequenas moléculas (39).

Ambos os métodos de divisão foram projetados para evitar o vazamento de informações entre os conjuntos de treinamento e teste, particularmente em casos onde os espectros diferem apenas em metadados, como energia de colisão ou tipo de precursor, e não em estrutura. Esta precaução é vital para garantir que o modelo aprenda a generalizar a partir de características estruturais das moléculas e não de peculiaridades dos dados de treinamento.

3.8 Hiperparâmetros utilizados

Os parâmetros utilizados no MassFormer são detalhadas na Tabela 2.

Parâmetro	**Valor**
loss	"cos"
sim	"cos"
lr	0.001
batch_size	100
weight_decay	0.001
num_epochs	20
train_seed	5585
split_seed	420
test_frac	20%

Tabela 2: Parâmetros e valores utilizados para o treinamento do modelo.

[11] "InChIKey"é um identificador único gerado a partir da estrutura química de uma molécula, derivado do "International Chemical Identifier"(InChI). O InChIKey é uma versão compacta e padronizada do InChI, projetada para ser mais fácil de usar em pesquisas, indexações e compartilhamento de informações químicas.

Enquanto o InChI é uma string que descreve a estrutura química detalhada de uma molécula, incluindo informações sobre a conectividade dos átomos, estereoquímica e outras características, o InChIKey é um resumo criptográfico (hash) de comprimento fixo dessa string. Ele é composto por 27 caracteres e facilita a busca de compostos químicos em bancos de dados, permitindo uma maneira rápida e eficiente de identificar moléculas específicas.

3.8.1 Descrição dos Hiperparâmetros

Nesta subseção, apresentamos uma explicação detalhada de cada hiperparâmetro utilizado no treinamento do modelo MassFormer, conforme listado na Tabela 2.

- **loss**: Define a função de perda utilizada para otimizar o modelo. Neste caso, "cos" refere-se à distância cosseno, que é adequada para comparar vetores em um espaço onde a magnitude é importante. A escolha dessa função é discutida com mais detalhes na subseção seguinte.
- **sim**: Representa a métrica de similaridade utilizada para avaliar o desempenho do modelo durante o treinamento e validação. Assim como o parâmetro de perda, "cos" indica a utilização da similaridade cosseno.
- **lr** (Learning Rate): A taxa de aprendizado inicial para o otimizador. Um valor de 0.001 é escolhido para permitir uma convergência suave sem ajustes bruscos nos pesos do modelo.
- **batch_size**: O tamanho do lote utilizado durante o treinamento, definido como 100. Este tamanho é escolhido para equilibrar a eficiência computacional e a estabilidade do treinamento.
- **weight_decay**: Parâmetro que ajuda na regularização do modelo, adicionando uma penalidade nos pesos durante o treinamento, com o objetivo de prevenir o sobreajuste.
- **num_epochs**: Número total de épocas de treinamento, definido como 20, indicando quantas vezes o conjunto de dados é passado integralmente pelo modelo durante o treinamento.
- **train_seed** e **split_seed**: Sementes para a inicialização do gerador de números aleatórios que influencia a inicialização dos pesos do modelo e a divisão dos dados, respectivamente. Isso garante a reprodutibilidade dos resultados.
- **test_frac**: Fração dos dados reservada para teste, indicada aqui como 20%. Este valor define quanto do conjunto de dados será usado exclusivamente para avaliar o desempenho do modelo após o treinamento.

Esses parâmetros são essenciais para configurar adequadamente o treinamento do modelo, garantindo que ele aprenda a partir dos dados disponíveis da maneira mais eficaz possível.

3.8.2 Cálculo da Perda (loss) e Similaridade

O desempenho do modelo de predição de espectros é avaliado principalmente através da distância cosseno, que é particularmente adequada devido à natureza relativa das intensidades

dos picos no espectro de massas. A distância cosseno é usada como função de perda e é definida como segue:

$$\mathrm{CD}(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{m} y_i \hat{y}_i}{\sqrt{\sum_{j=1}^{m} y_j^2 \sum_{k=1}^{m} \hat{y}_k^2}} \tag{3.3}$$

onde y é o espectro real e $\hat{y} = f_\theta(x,z)$ é o espectro predito pelo modelo. Minimizar a distância cosseno maximiza a similaridade cosseno, que é uma métrica comum para comparação de espectros. A similaridade cosseno é particularmente útil para avaliar quão similares são os espectros, independentemente da escala das intensidades dos picos.

Para calcular a similaridade espectral nos experimentos, os espectros, tanto preditos quanto alvos, são primeiro medidos em relação à energia de colisão antes de comparar a similaridade. Este método ajuda a prevenir scores de similaridade inflados que podem ocorrer quando a energia de colisão é muito alta ou muito baixa. Os scores de similaridade são calculados para cada molécula considerando todos os adutos de precursor, e então são feitas médias adicionais entre as moléculas para fornecer uma medida robusta e representativa da precisão do modelo em todo o conjunto de dados.

4

Resultados e Discussões

4.1 Métricas de avaliação

Nesta seção, são apresentados os resultados obtidos a partir do refinamento e da avaliação do modelo MassFormer. As métricas de avaliação utilizadas para analisar a eficácia e a precisão do modelo incluem uma variedade de indicadores relacionados à similaridade espectral e ao erro de predição. As métricas primárias focam na similaridade entre os espectros previstos e reais, tanto para moléculas (`mol_sim_obj_mean` e `mol_sim_cos_std_mean`) quanto para espectros (`spec_sim_obj_mean` e `spec_sim_cos_std_mean`). Além disso, as métricas de perda (`mol_loss_obj_mean` e `spec_loss_obj_mean`) quantificam o desvio dos valores preditos em relação aos reais, proporcionando uma avaliação direta da performance preditiva do modelo. A métrica `best_val_loss_mean` é utilizada para avaliar a perda mínima obtida no conjunto de validação, refletindo a eficiência do modelo sob condições ideais de ajuste. A discussão a seguir explora como cada uma dessas métricas se comporta sob diferentes configurações de treinamento e o que isso revela sobre a capacidade do modelo de generalizar a partir de dados espectrais complexos.

4.1.1 Similaridade de Espectro (`spec_sim_obj_mean`)

Um valor de 0.539 para a divisão por InChIKey sugere uma maior eficácia nessa configuração, comparada a 0.5362 para a divisão por Scaffold, refletindo uma ligeira superioridade na conservação da integridade estrutural dos dados no método de divisão por InChIKey.

4.1.2 Similaridade de Moléculas (`mol_sim_obj_mean`)

Um valor de 0.5013 na divisão por InChIKey versus 0.5071 na divisão por Scaffold indica que a abordagem de Scaffold, apesar de introduzir um desvio distribucional, captura com mais

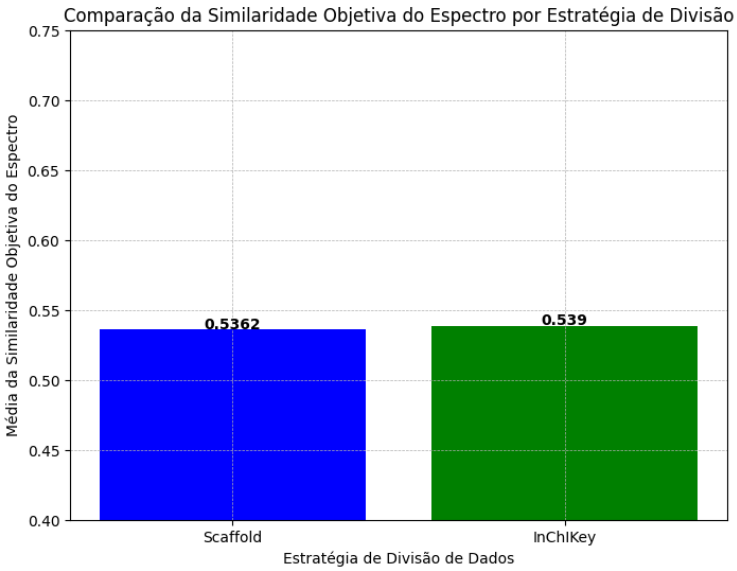

Figura 8: Comparação da Similaridade de Espectro entre as divisões Scaffold e InChIKey.

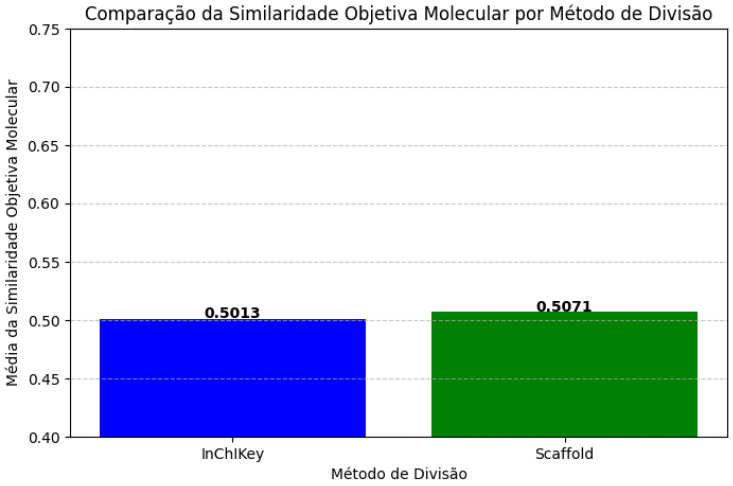

Figura 9: Comparação da Similaridade de moléculas entre as divisões Scaffold e InChIKey.

eficiência as características moleculares essenciais para a tarefa de predição de espectros. Isso sugere que a estratificação detalhada dos compostos pode ajudar o modelo a aprender representações moleculares mais robustas e generalizáveis.

4.1.3 Desvio Padrão da Similaridade de Moléculas (`mol_sim_cos_std_mean`)

InChIKey (0.3946): O valor mais alto indica que o modelo, quando treinado e testado com a estratégia de divisão InChIKey, apresenta uma variação ligeiramente maior nas previsões de similaridade de cosseno entre as moléculas, sugerindo uma distribuição mais ampla de qualidade de previsão.

Scaffold (0.3614): Um valor ligeiramente mais baixo sob a divisão Scaffold sugere que o modelo é mais consistente ao prever a similaridade de espectros entre as moléculas quando a divisão dos dados é feita de maneira mais estratificada.

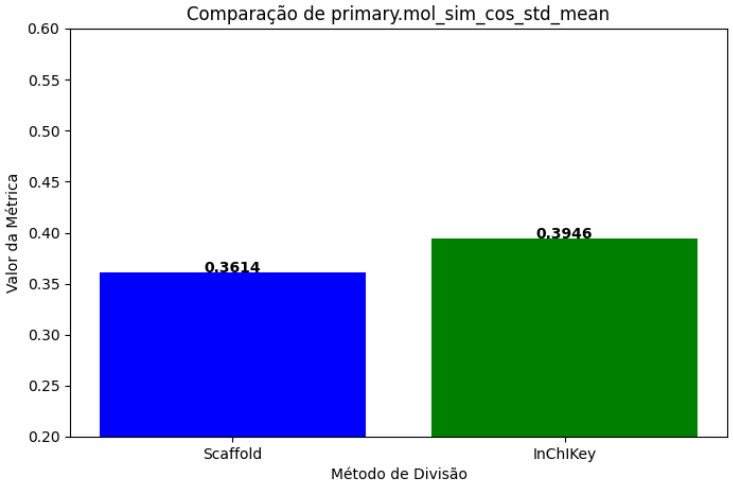

Figura 10: Comparação do Desvio Padrão da Similaridade de Moleculas entre as divisões Scaffold e InChIKey.

4.1.4 Desvio Padrão da Similaridade de Espectros (`spec_sim_cos_std_mean`)

A divisão Scaffold, com um valor de 0.4077, e a divisão InChIKey, com um valor de 0.4037, mostram desempenhos muito próximos, indicando que a abordagem de divisão do conjunto de dados não teve um impacto significativo na consistência das previsões de similaridade cosseno para esta métrica.

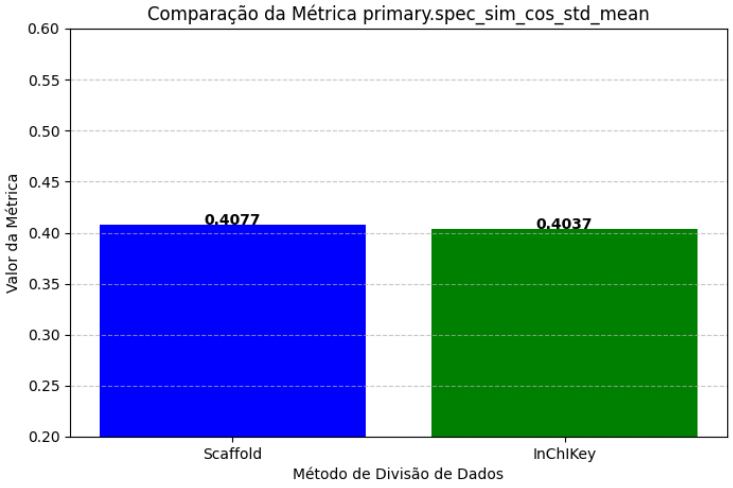

Figura 11: Comparação do Desvio Padrão da Similaridade de Espectro entre as divisões Scaffold e InChIKey.

4.1.5 Perda Média de Moléculas (`mol_loss_obj_mean`)

A estratégia de divisão por Scaffold gerou uma perda média de 0.4929, enquanto a divisão por InChIKey resultou em uma perda ligeiramente superior de 0.4987. Isso sugere que o modelo com a divisão por Scaffold se ajustou um pouco melhor aos dados, possivelmente devido à maneira como essa estratégia agrupa compostos com estruturas químicas semelhantes, proporcionando um treinamento que pode captar melhor as nuances dos grupos de moléculas.

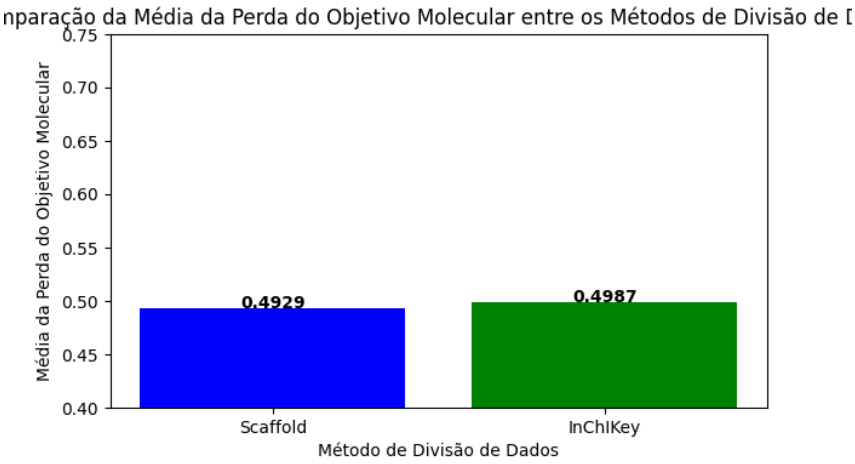

Figura 12: Comparação da Perda Média de Moléculas entre as divisões Scaffold e InChIKey.

4.1.6 Perda Média de Espectros (`spec_loss_obj_mean`)

A análise dos resultados mostra uma ligeira diferença nas médias de perda entre as duas estratégias de divisão de dados. A divisão por Scaffold apresentou uma perda média de 0.4638, enquanto a divisão por InChIKey resultou em uma perda ligeiramente menor de 0.461. Essa diferença sutil sugere que, embora a estratégia Scaffold possa enfrentar desafios adicionais devido à variação estrutural dentro dos grupos, ela ainda assim mantém uma performance comparável à abordagem mais aleatória e direta da divisão por InChIKey.

4.1.7 Melhor Perda de Validação (`best_val_loss_mean`)

Os resultados indicam que a estratégia de divisão por Scaffold obteve uma melhor perda de validação (0.4929) em comparação com a estratégia por InChIKey (0.4987), mostrando que, apesar dos desafios de sua abordagem estruturada, Scaffold ajudou a moldar um treinamento que capacita o modelo a generalizar de forma mais eficaz em cenários complexos.

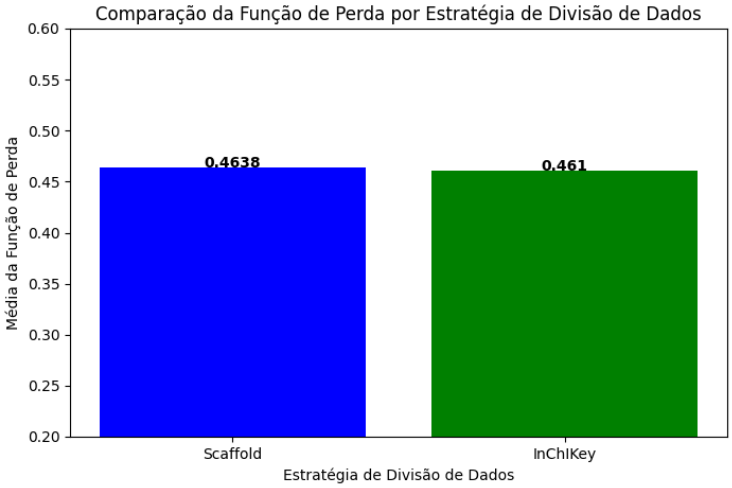

Figura 13: Comparação da Perda Média de Espectro entre as divisões Scaffold e InChIKey.

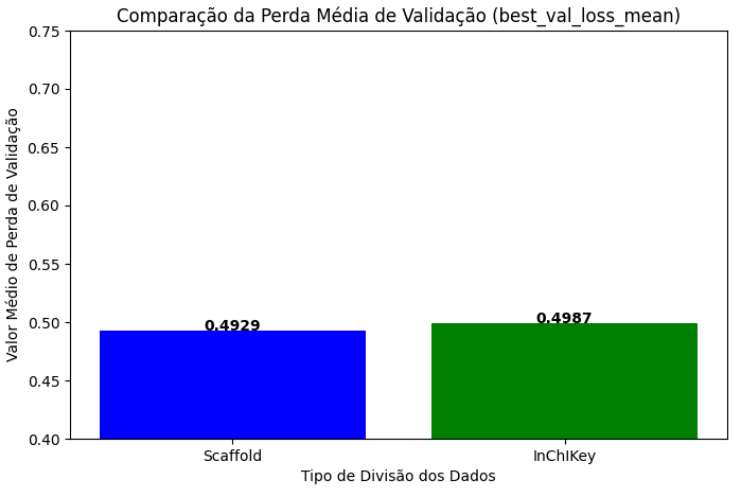

Figura 14: Comparação da Melhor Perda de Validação entre divisões Scaffold e InChIKey.

4.2 Desempenho do MassFormer

4.2.1 Maiores similaridades de cosseno

A análise das melhores predições geradas pelo modelo MassFormer revela padrões interessantes que podem ser correlacionados com a estrutura molecular e a complexidade do espectro. Observou-se que moléculas com estruturas mais simples ou com pequenas cadeias aromáticas tendem a resultar em predições de alta precisão. Isso pode ser atribuído à menor complexidade dos espectros gerados por tais moléculas, facilitando assim a tarefa de aprendizado e predição do modelo.

Especificamente, moléculas curtas e simples frequentemente apresentam espectros com um número reduzido de picos dominantes. Essa característica é particularmente vantajosa para o modelo, pois simplifica o espaço de saída que precisa ser aprendido. Além disso, a presença de

pequenas cadeias aromáticas pode levar a padrões de fragmentação mais previsíveis e menos variáveis, contribuindo ainda mais para a precisão das predições.

Um aspecto notável nas melhores predições é a ocorrência de espectros que exibem predominantemente uma única linha intensa. Esse fenômeno indica que há pouca fragmentação ocorrendo ou que os principais fragmentos são muito mais abundantes do que outros. Esse padrão simplificado no espectro permite que o modelo concentre seus recursos de aprendizado na captura e reprodução dessas características principais, o que é evidenciado pela alta similaridade coseno alcançada nessas predições.

Essas observações sugerem que a estrutura molecular e a simplicidade do espectro são fatores cruciais que influenciam a eficácia das predições do modelo. Portanto, entender e explorar essas relações pode ser fundamental para futuros aprimoramentos e aplicações práticas do MassFormer em contextos de análise espectrométrica.

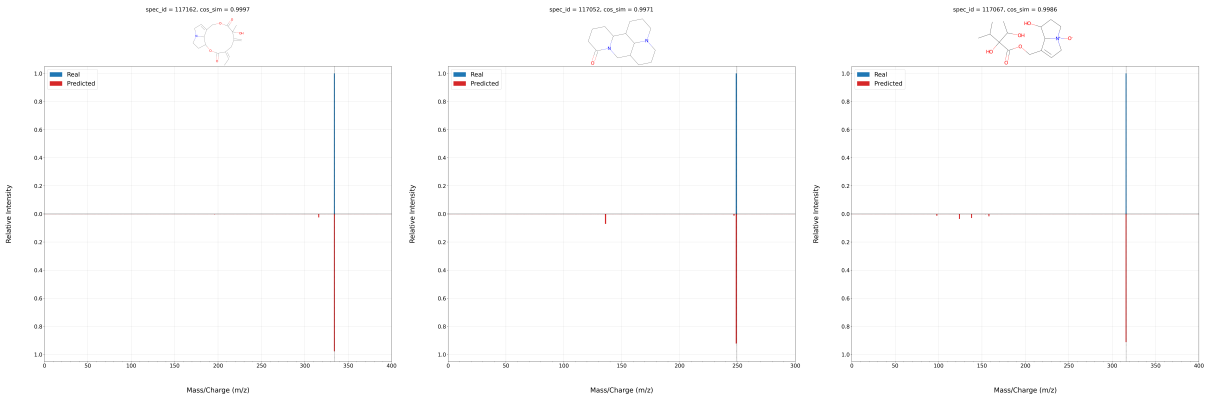

Figura 15: TopK: Amostras A, B e C melhores predições

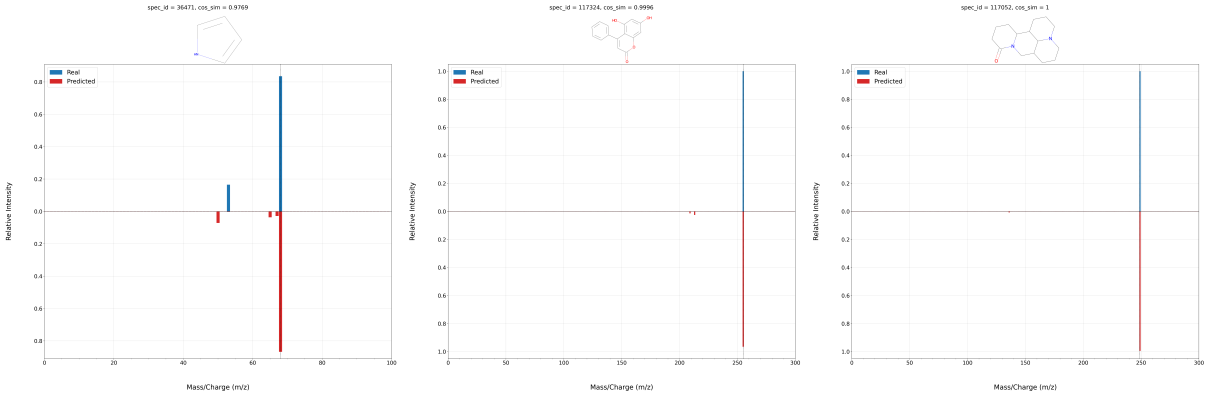

Figura 16: TopK: Amostras D, E e F melhores predições

4.2.2 Piores similaridades

Os resultados das piores predições, destacam desafios específicos enfrentados pelo modelo MassFormer ao lidar com espectros de moléculas de estrutura complexa. Estas moléculas fre-

quentemente apresentam anéis poli-aromáticos, múltiplas substituições heteroatômicas, ou extensas cadeias alifáticas, que podem gerar uma grande variedade de fragmentos durante a ionização.

Estes exemplos, que exibem uma similaridade cosseno de zero, revelam limitações na capacidade do modelo de generalizar a partir de dados espectrais complexos. Isso sugere que abordagens de treinamento aprimoradas, aumento na diversidade dos dados de treinamento, ou ajustes na representação dos dados podem ser necessários para melhorar as predições do modelo em casos desafiadores.

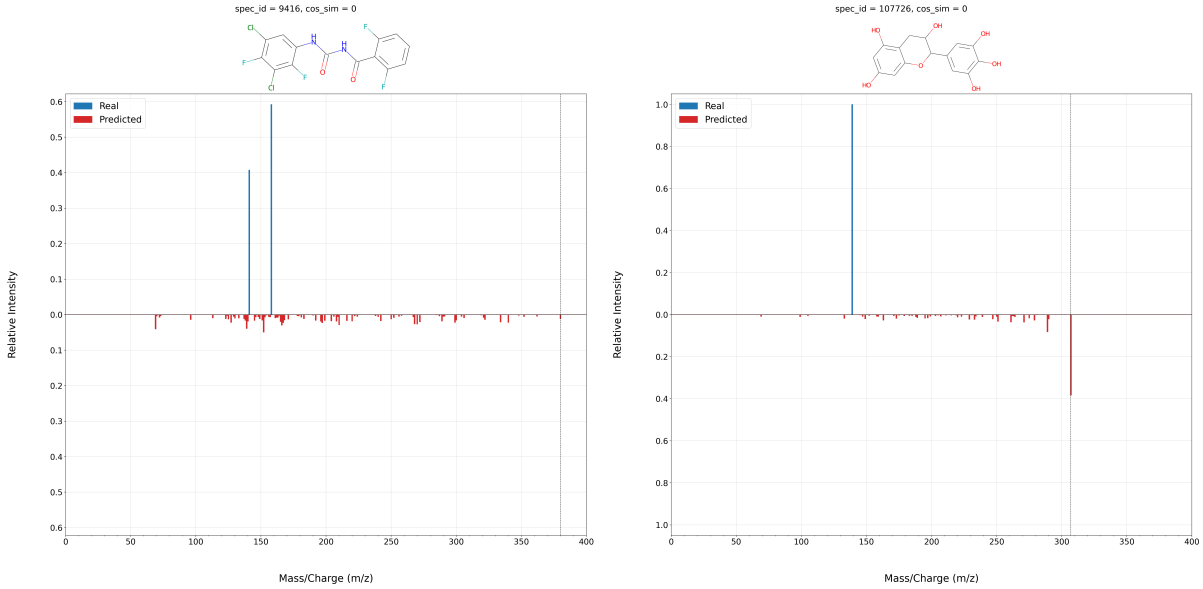

Figura 17: BotomK: Amostras A e B piores predições

4.2.3 Análise de Amostras com Valores Aleatórios de Similaridade de Cosseno

As amostras apresentadas com valores aleatórios de similaridade de cosseno revelam variados graus de eficácia na previsão dos espectros pelo modelo. Essas variações podem ser atribuídas à complexidade das estruturas moleculares envolvidas e à capacidade do modelo de interpretar e reproduzir corretamente as características espectrais.

Essas observações indicam que a capacidade do modelo de prever com precisão os espectros varia significativamente com a natureza química das amostras e destaca a necessidade de melhorias contínuas no treinamento e refinamento do modelo para lidar com a diversidade de desafios apresentados por diferentes tipos de moléculas.

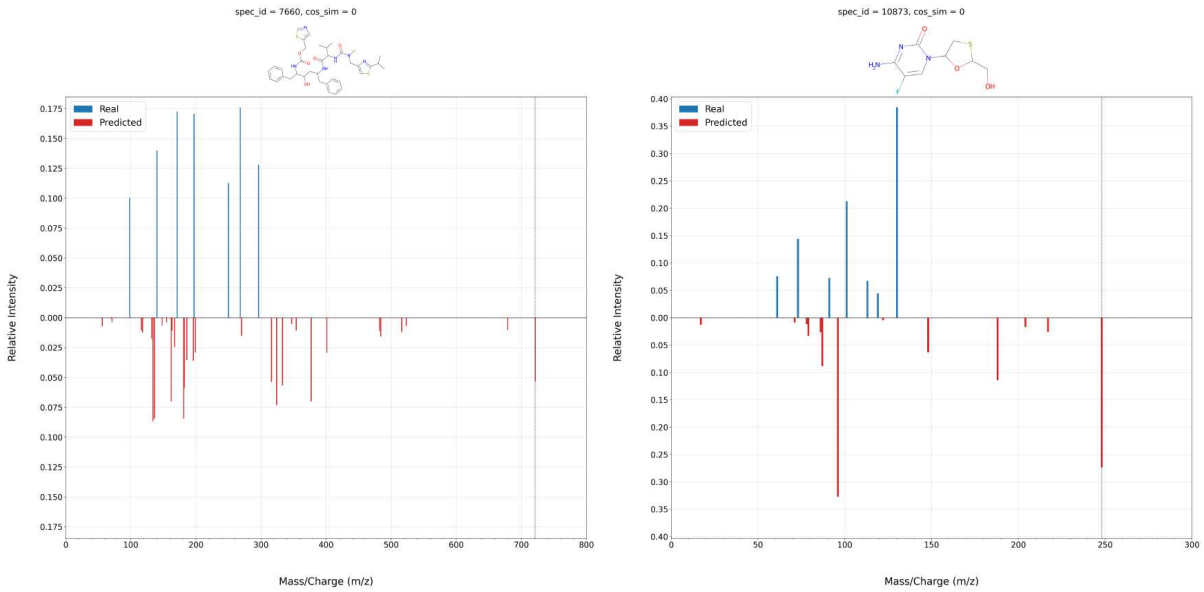

Figura 18: BotomK: Amostras C e D piores predições

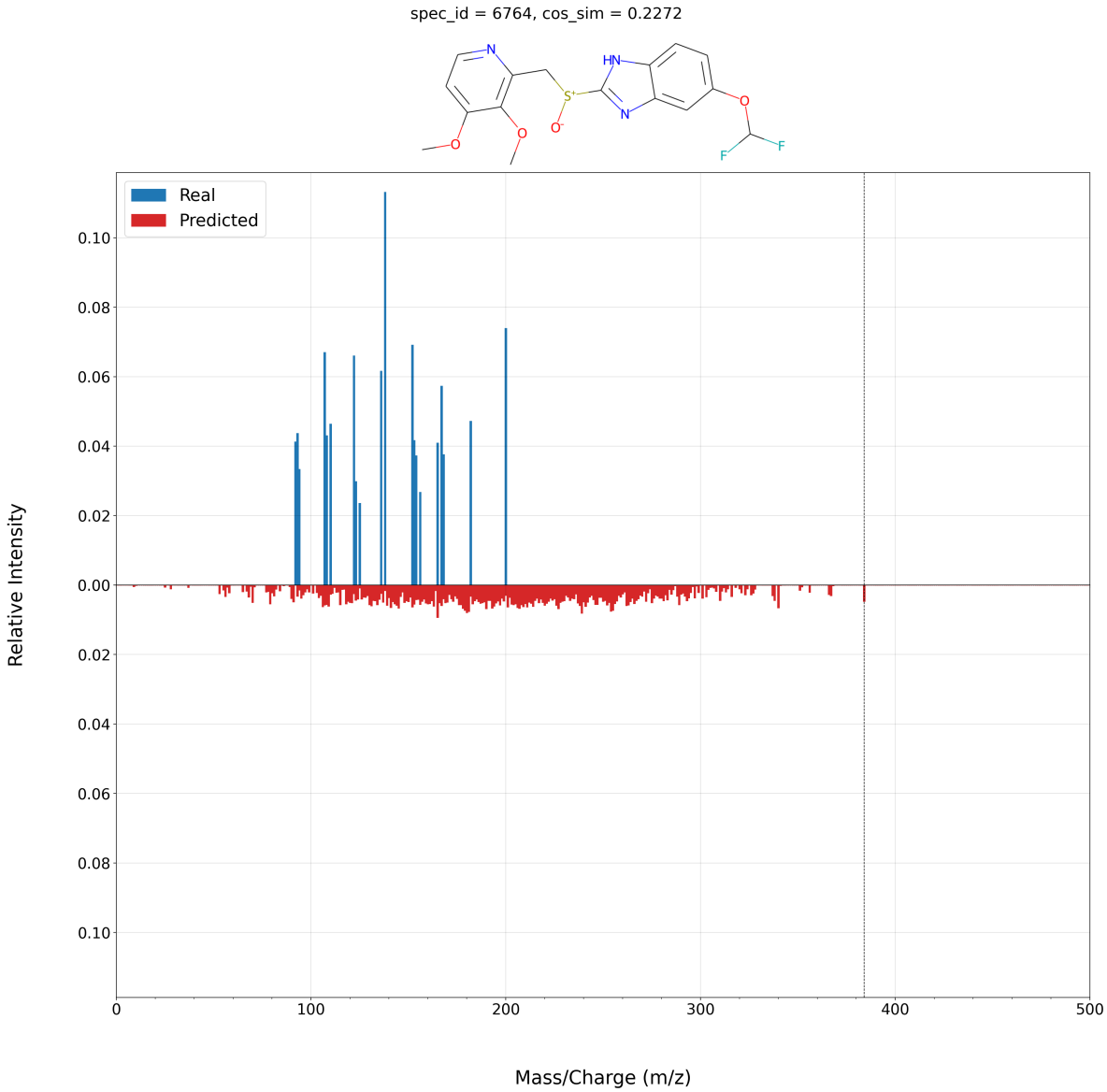

Figura 19: Randk: A

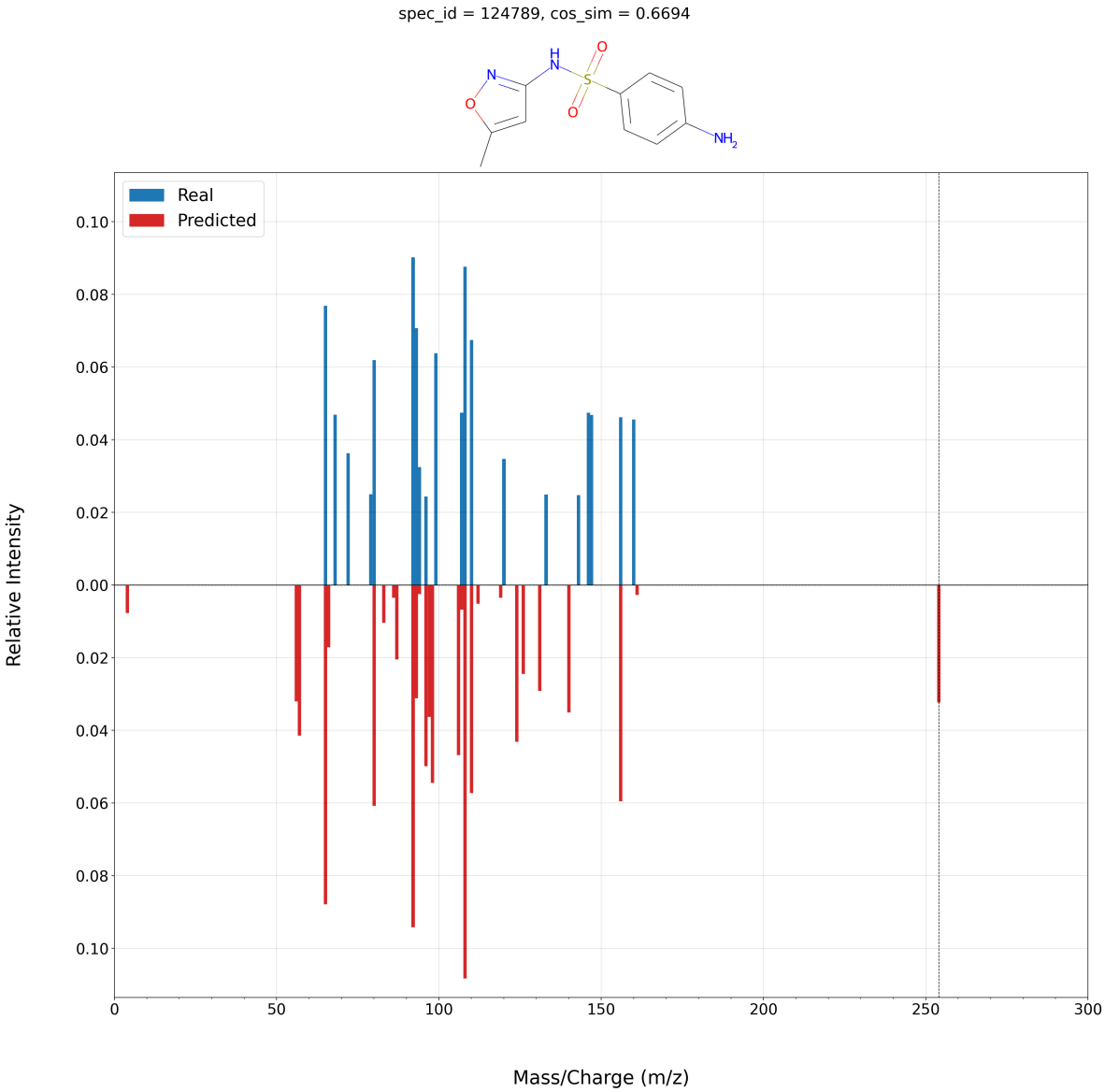

Figura 20: Randk: B

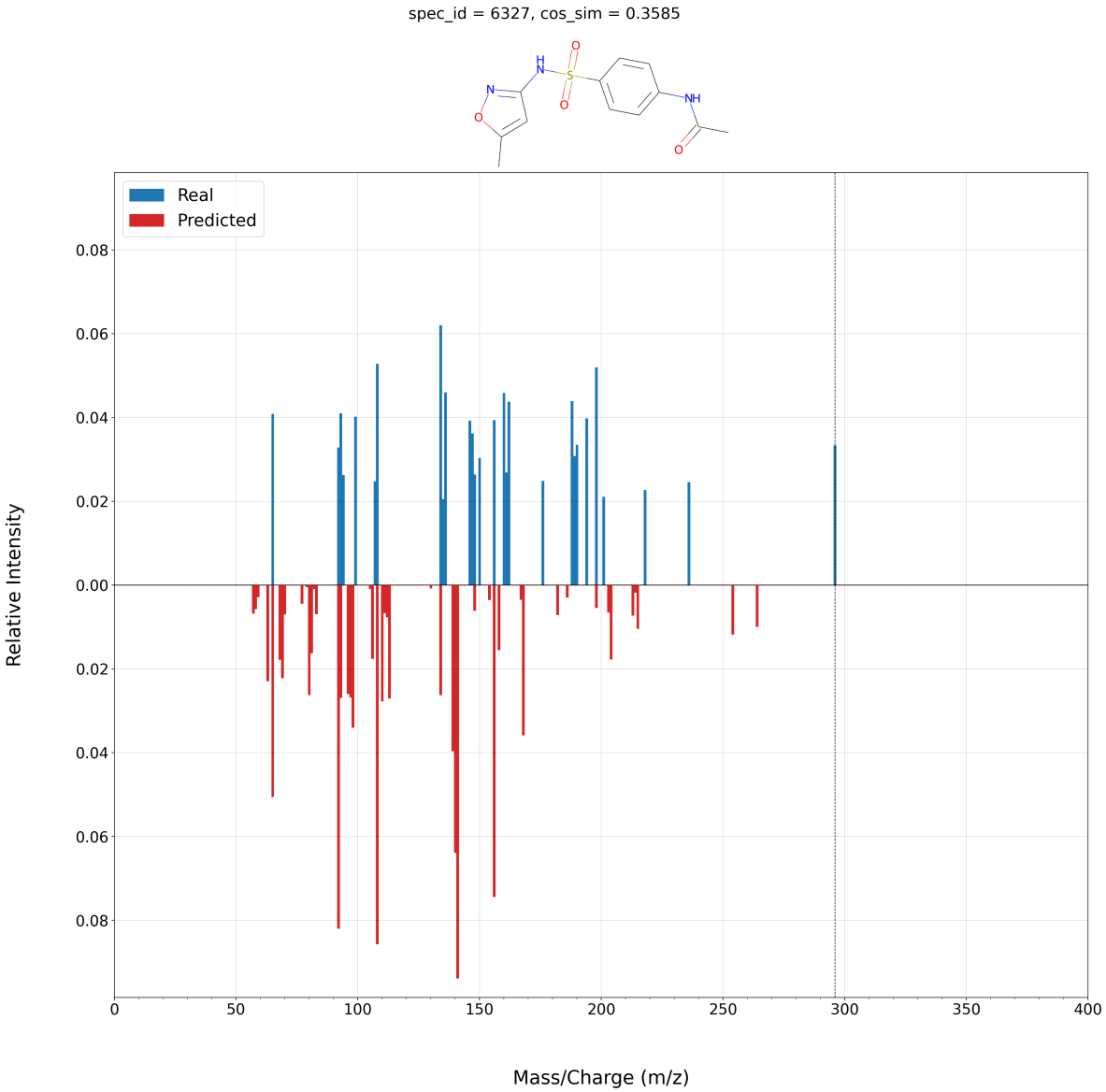

Figura 21: Randk: C

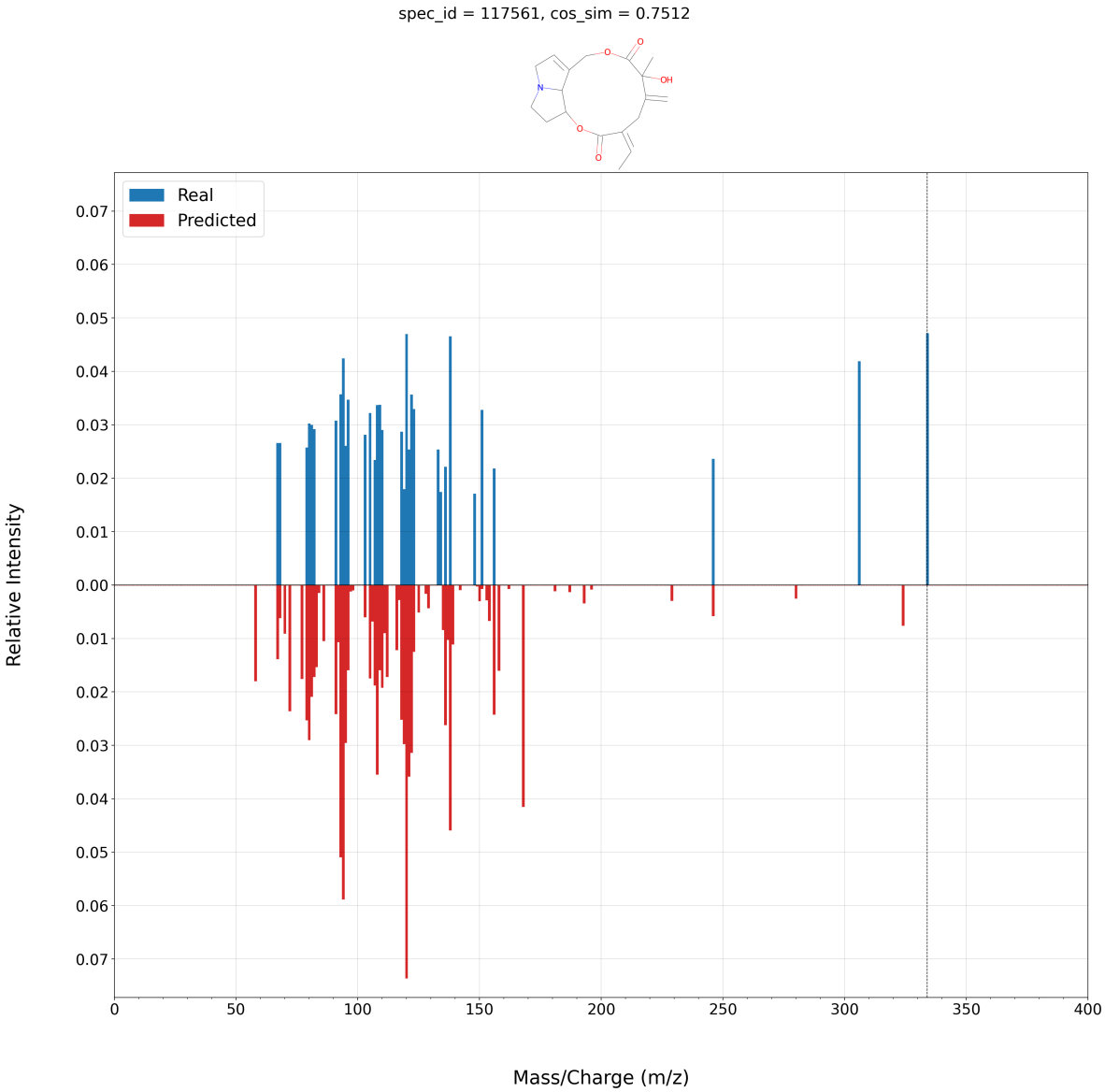

Figura 22: Randk: D

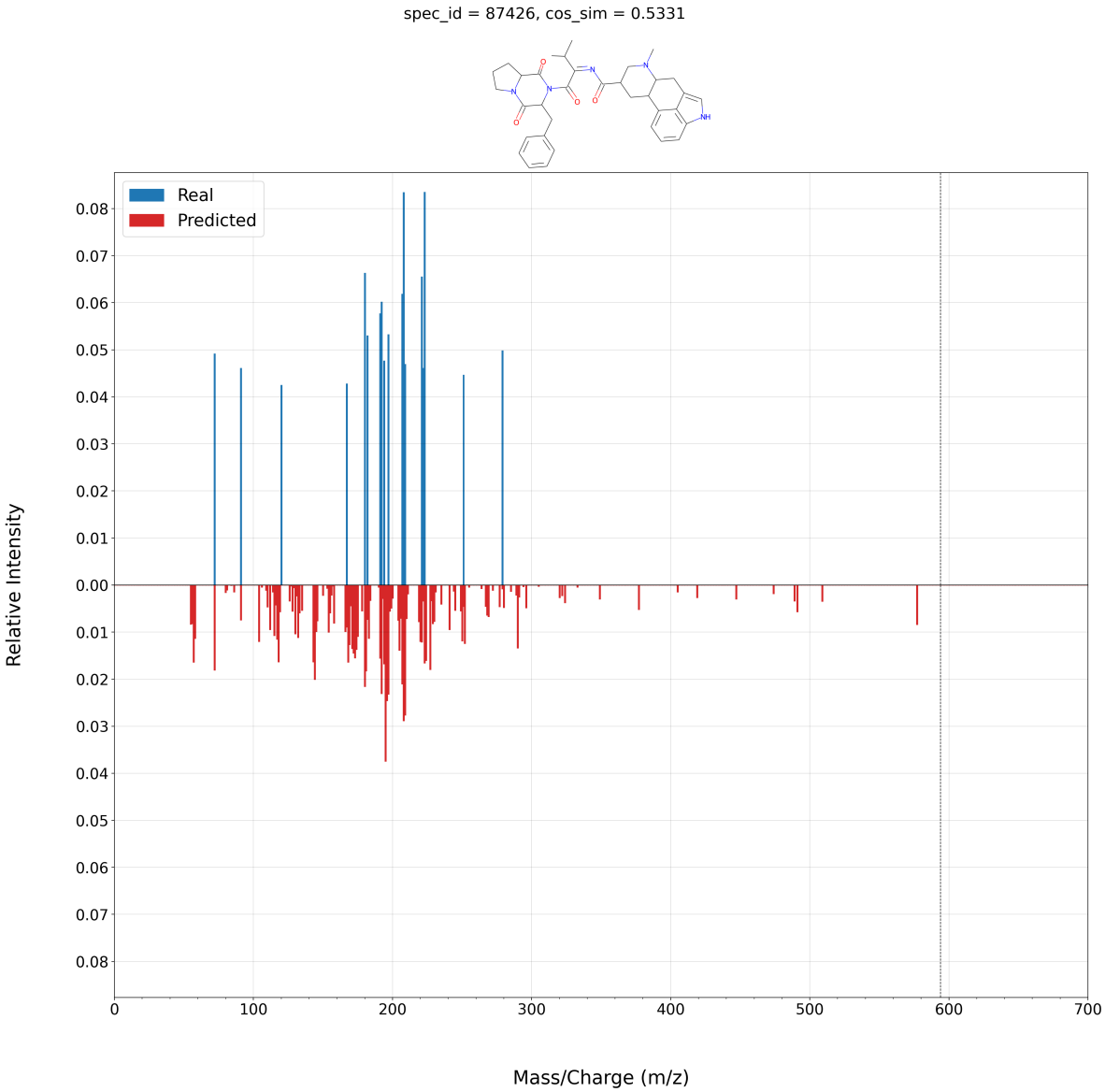

Figura 23: Randk: E

5

Conclusão

5.1 Conclusão

Este trabalho apresentou a aplicabilidade do *MassFormer*, um modelo avançado baseado em grafo transformer, no contexto de um conjunto de dados público, o MassBank of North America (MoNA). Foi possível demonstrar que o modelo é capaz de prever espectros de fragmentos de massa com uma precisão considerável, apesar das limitações inerentes aos dados disponíveis publicamente.

Ambos os métodos de divisão de dados testados, divisão por InChIKey e divisão por Scaffold, demonstraram resultados comparativamente similares, indicando que o modelo possui uma robustez adequada para generalizar a partir de diferentes abordagens de estratificação dos dados. Esta característica é crucial para o desenvolvimento de ferramentas analíticas em espectrometria de massa, onde a diversidade dos tipos de compostos químicos e as condições experimentais podem variar significativamente.

É importante destacar que, embora o *MassFormer* aplicado ao conjunto de dados MoNA tenha mostrado eficácia, os resultados ainda estão aquém dos obtidos por modelos comerciais. A principal razão para essa discrepância é a menor quantidade e qualidade dos dados no MoNA comparados aos conjuntos de dados proprietários NIST. No entanto, a natureza aberta deste modelo e do conjunto de dados oferece uma oportunidade valiosa para a comunidade científica colaborar e enriquecer continuamente a base de dados de espectros de massa. Com o tempo e o aumento do volume e da qualidade dos dados contribuídos, espera-se que o *MassFormer* possa alcançar ou até superar o desempenho dos modelos comerciais.

Em conclusão, o desenvolvimento e a aplicação do *MassFormer* dentro deste estudo não apenas provam o potencial dos modelos de aprendizado profundo em espectrometria de massa, mas também abrem caminho para futuras colaborações científicas que possam explorar e expandir ainda mais as capacidades deste e de outros modelos similares. Tal progresso continuará

a impulsionar inovações na análise de espectros de massa, com implicações diretas para a pesquisa e desenvolvimento em ciências biomédicas, químicas e ambientais.

5.2 Trabalhos Futuros

A pesquisa realizada neste trabalho abre várias avenidas para investigações futuras e desenvolvimentos no campo da espectrometria de massa utilizando aprendizado de máquina. Os resultados obtidos com o *MassFormer* utilizando o conjunto de dados do MassBank of North America (MoNA) indicam um potencial significativo para melhorias e expansões. Três áreas principais foram identificadas como focos para trabalhos futuros:

1. **Incremento de Dados no MoNA:** Uma das principais limitações identificadas no uso do MoNA foi a quantidade e diversidade de dados disponíveis. Trabalhos futuros poderão se concentrar em expandir o repositório de espectros de massa disponíveis no MoNA, incentivando a comunidade científica a contribuir com novos dados de espectros. A ampliação da base de dados não apenas melhora a precisão e a generalização dos modelos de predição, mas também enriquece as possibilidades de descoberta e identificação de novos compostos químicos.

2. **Revisão dos Metadados Anotados no MoNA:** A qualidade dos metadados associados aos espectros é crucial para o treinamento de modelos precisos e confiáveis. Uma revisão sistemática e a atualização dos metadados no MoNA ajudarão a garantir que as informações sejam precisas e detalhadas, permitindo uma melhor interpretação e utilização dos dados nos modelos de aprendizado de máquina. Esforços nesse sentido podem incluir a padronização das anotações e a verificação da consistência das informações fornecidas.

3. **Desenvolvimento de Modelos para Identificação Molecular:** Além de prever espectros de fragmentos a partir de estruturas moleculares, há uma oportunidade substancial para desenvolver modelos que realizem o processo inverso, isto é, a identificação de estruturas moleculares a partir de seus espectros de massa. Este avanço seria revolucionário, permitindo o uso prático de espectrometria de massa em aplicações como o controle de qualidade, monitoramento ambiental e diagnóstico biomédico, onde a identificação rápida e precisa de substâncias é essencial.

Esses avanços não apenas aprimorarão as capacidades analíticas dos cientistas e pesquisadores, mas também promoverão uma colaboração mais estreita entre as comunidades de química, biologia e ciência da computação, explorando plenamente o potencial da análise de dados em larga escala no campo da espectrometria de massa.

Referências bibliográficas

[1] Allen, F., Greiner, R., and Wishart, D. (2015). Competitive fragmentation modeling of esi-ms/ms spectra for putative metabolite identification. *Metabolomics*, 11:98–110.

[2] Bemis, G. W. and Murcko, M. A. (1996). The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893.

[Broad Institute] Broad Institute. What is Mass Spectrometry. https://www.broadinstitute.org/technology-areas/what-mass-spectrometry. Acessado em: 10 de junho de 2024.

[4] Cao, L., Guler, M., Tagirdzhanov, A., Lee, Y.-Y., Gurevich, A., and Mohimani, H. (2021). Moldiscovery: Learning mass spectrometry fragmentation of small molecules. *Nature communications*, 12(1):3718.

[5] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[6] Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. (2002). Reoptimization of mdl keys for use in drug discovery. *Journal of chemical information and computer sciences*, 42(6):1273–1280.

[7] Dwivedi, V. P., Joshi, C. K., Luu, A. T., Laurent, T., Bengio, Y., and Bresson, X. (2023). Benchmarking graph neural networks. *Journal of Machine Learning Research*, 24(43):1–48.

[8] Floyd, R. W. (1962). Algorithm 97: shortest path. *Communications of the ACM*, 5(6):345–345.

[9] Francois, C. (2018). Deep learning with python.

[10] Galgonek, J. and Vondrášek, J. (2024). The idsm mass spectrometry extension: searching mass spectra using sparql. *Bioinformatics*, 40(4):btae174.

[11] Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., et al. (2017). The chembl database in 2017. *Nucleic acids research*, 45(D1):D945–D954.

[12] Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. "O'Reilly Media, Inc.".

[13] Glish, G. L. and Vachet, R. W. (2003). The basics of mass spectrometry in the twenty-first century. *Nature reviews drug discovery*, 2(2):140–150.

[14] Goldman, S., Bradshaw, J., Xin, J., and Coley, C. (2023). Prefix-tree decoding for predicting mass spectra from molecules. *Advances in Neural Information Processing Systems*, 36:48548–48572.

[15] Gross, J. H. and Roepstorff, M. S. (2011). A textbook. *Springer*, 2:415–452.

[16] Heller, S. R., McNaught, A., Pletnev, I., Stein, S., and Tchekhovskoi, D. (2015). Inchi, the iupac international chemical identifier. *Journal of cheminformatics*, 7:1–34.

[17] Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

[18] Hong, Y., Li, S., Welch, C. J., Tichy, S., Ye, Y., and Tang, H. (2023). 3dmolms: prediction of tandem mass spectra from 3d molecular conformations. *Bioinformatics*, 39(6):btad354.

[19] Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., et al. (2010). Massbank: a public repository for sharing mass spectral data for life sciences. *Journal of mass spectrometry*, 45(7):703–714.

[20] Hu, W., Fey, M., Ren, H., Nakata, M., Dong, Y., and Leskovec, J. (2021). Ogb-lsc: A large-scale challenge for machine learning on graphs. *arXiv preprint arXiv:2103.09430*.

[21] Hu, W., Fey, M., Zitnik, M., Dong, Y., Ren, H., Liu, B., Catasta, M., and Leskovec, J. (2020). Open graph benchmark: Datasets for machine learning on graphs. *Advances in neural information processing systems*, 33:22118–22133.

[22] Janner, M., Li, Q., and Levine, S. (2021). Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286.

[23] Litsa, E. E., Chenthamarakshan, V., Das, P., and Kavraki, L. E. (2023). An end-to-end deep learning framework for translating mass spectra to de-novo molecules. *Communications Chemistry*, 6(1):132.

[24] Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.

[25] Murphy, M., Jegelka, S., Fraenkel, E., Kind, T., Healey, D., and Butler, T. (2023). Efficiently predicting high resolution mass spectra with graph neural networks. In *International Conference on Machine Learning*, pages 25549–25562. PMLR.

[26] Nakata, M. and Shimazaki, T. (2017). Pubchemqc project: a large-scale first-principles electronic structure database for data-driven chemistry. *Journal of chemical information and modeling*, 57(6):1300–1308.

[27] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

[28] Priyam Study Centre (2022). Mass spectrometry - instrumentation and definition. https://www.priyamstudycentre.com/2022/02/mass-spectrometry.html. Acessado em: 10 de junho de 2024.

[29] Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754.

[30] Ruttkies, C., Neumann, S., and Posch, S. (2019). Improving metfrag with statistical learning of fragment annotations. *BMC bioinformatics*, 20:1–14.

[31] Sanchez-Lengeling, B., Reif, E., Pearce, A., and Wiltschko, A. B. (2021). A gentle introduction to graph neural networks. *Distill*. https://distill.pub/2021/gnn-intro.

[32] Sarker, I. H. (2021). Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Computer Science*, 2(6):420.

[33] Thomson, J. J. (1897). Xl. cathode rays. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 44(269):293–316.

[34] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[35] Wang, F., Liigand, J., Tian, S., Arndt, D., Greiner, R., and Wishart, D. S. (2021). Cfm-id 4.0: more accurate esi-ms/ms spectral prediction and compound identification. *Analytical chemistry*, 93(34):11692–11700.

[36] Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., et al. (2019). Deep graph library: A graph-centric, highly-performant package for graph neural networks. *arXiv preprint arXiv:1909.01315*.

[37] Watson, J. T. and Sparkman, O. D. (2007). *Introduction to mass spectrometry: instrumentation, applications, and strategies for data interpretation*. John Wiley & Sons.

[38] Wei, J. N., Belanger, D., Adams, R. P., and Sculley, D. (2019). Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS central science*, 5(4):700–708.

[39] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530.

[40] Ying, C., Cai, T., Luo, S., Zheng, S., Ke, G., He, D., Shen, Y., and Liu, T.-Y. (2021). Do transformers really perform badly for graph representation? *Advances in neural information processing systems*, 34:28877–28888.

[41] Young, A., Röst, H., and Wang, B. (2024). Tandem mass spectrum prediction for small molecules using graph transformers. *Nature Machine Intelligence*, pages 1–13.

[42] Young, A., Wang, B., and Röst, H. (2021). Massformer: Tandem mass spectrum prediction for small molecules using graph transformers. *arXiv preprint arXiv:2111.04824*.

[43] Zhu, H., Liu, L., and Hassoun, S. (2020). Using graph neural networks for mass spectrometry prediction. *arXiv preprint arXiv:2010.04661*.