



TRAINING LARGE LANGUAGE MODELS FOR NATURAL PRODUCTS WITH MASKED LEARNING USING DEEPMOL: A BENCHMARK FOR SIMILARITY OVERLAP RESOLUTION

UC Projeto
Mestrado em Bioinformática
Escola de Engenharia da Universidade do Minho

Discente: Pedro Pereira PG55703
Orientador: João Capela e Miguel Rocha

[HTTPS://GITHUB.COM/PHCPO110/PROJECT](https://github.com/PHCPO110/PROJECT)



INTRODUÇÃO

Os produtos naturais apresentam uma diversidade estrutural e biossintética significativa, o que levanta desafios na sua representação computacional e comparação.

Neste trabalho, desenvolvemos um pipeline modular para aplicar e comparar diferentes métodos de *fingerprint* sobre o mesmo conjunto de *natural products*, com o objetivo de explorar a similaridade entre moléculas a partir de múltiplas perspectivas. Embora a avaliação de similaridade biossintética e o treino do modelo NPBERT estejam ainda por concluir, esta fase inicial permitiu estruturar uma abordagem coerente e escalável, adequada à integração dessas componentes.



Fingerprints utilizadas

Fingerprint	Tipo de vetor	Descrição da representação
NPClassifierFP	Inteiros	Contagens de sobreposição entre várias Morgan FPs com diferentes raios
Biosynfoni	Binário	Codifica presença de subestruturas biossintéticas
NeuralNPFP	Contínuo	Vetores aprendidos por redes neuronais supervisionadas
MHFP	Inteiros	Vetores de MinHash com índices de subestruturas
MorganFingerprint	Binário	Padrão estrutural circular
NPBERT	Contínuo	Vetor de 512 dimensões com <i>embeddings</i> químicos aprendidos

Como as *fingerprints* geram vetores de naturezas distintas (binários, contínuos e inteiros), não era possível compará-las diretamente.

Para resolver isso, utilizámos a distância de Manhattan como métrica universal.


Em seguida, normalizámos os vetores de similaridade para o intervalo [0, 1], permitindo comparação entre métodos.

Por fim, calculámos correlações entre esses vetores, o que possibilitou gerar representações como MSTs e violin plots.




RESEARCH GAP

Falta de comparação de *fingerprints* para *Natural Products* (NP)




Embora existam várias *fingerprints* ainda não há estudos abrangentes que comparem o desempenho destas representações especificamente em tarefas envolvendo NPs [1, 2].

Falta de *benchmarks* de BERT *models* para NP



Os modelos de linguagem, como o BERT, começaram a ser adaptados ao domínio químico com resultados promissores, mas ainda carecem de *benchmarks* dedicados a NPs que permitam avaliar de forma justa e reprodutível o seu desempenho [2, 3].

Falta de métricas de similaridade bio sintéticas para *fingerprints* - Biosynfoni



Biosynfoni é uma *fingerprint* construída com base na lógica de biossíntese dos NPs, mas o seu uso prático ainda é limitado pela ausência de métricas quantitativas que avaliem biossimilaridade de forma robusta [1, 4].

OBJETIVOS

01

Comparar FPs para NPs

02

Criar modelos BERT para NPs

MÉTODOS

01

Dataset

02

Geração de *fingerprints*

03

Comparação

05



CRIAÇÃO DO DATASET

DATASET PARA TREINO COM MASKED LEARNING

Fontes: COCONUT (~695k moléculas) e LotusDB (~276k), integradas.

Normalização: baseada na rotina ChEMBL (via DeepMol).

InChIKey *cleaning*: eliminar duplicados com o mesmo InChIKey (até ao 26.º caractere).

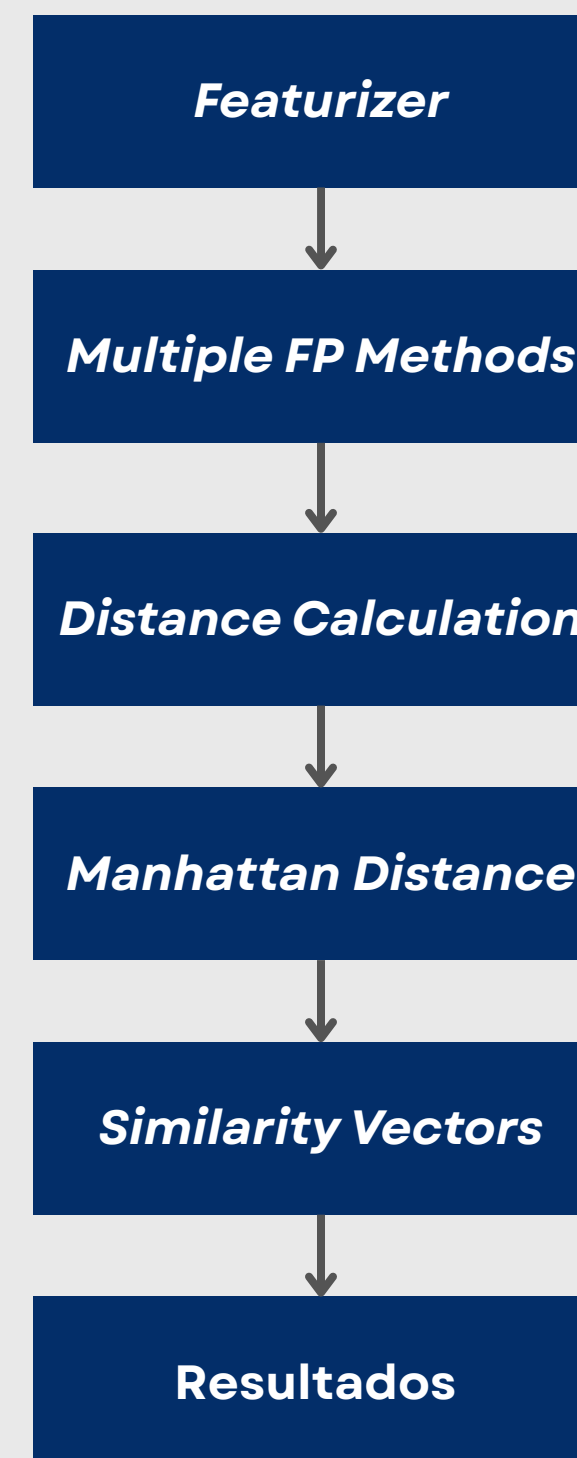
Filtro *anti-leakage*: remoção de moléculas com similaridade >60% (Tanimoto) com o *test set*.

Partição: divisão usando *scaffold split*.

GERAÇÃO DE FINGERPRINTS

Compara *fingerprints* aplicadas ao mesmo conjunto de produtos naturais, avaliando a similaridade entre moléculas e entre métodos.

1. **Featurizer:** Gera as representações moleculares.
2. **Múltiplos métodos de *fingerprint*:** Aplica os vários métodos ao dataset.
3. **Cálculo de Distâncias:** Comparação dos *FPs*.
4. **Distância de *Manhattan*:** Métrica usada para medir a dissimilaridade
5. **Normalização:** Todos os vetores são normalizados para garantir comparabilidade.
6. **Análise de Correlação:** Escala os vetores para garantir comparabilidade
7. **Matriz de Correlação:** Mede a relação entre os métodos.
8. **MST (*Minimum Spanning Tree*):** Visualizar as relações entre métodos.
9. **Violin Plots:** Mostra a distribuição das similaridades.



COMPUTAR A DISTÂNCIA

Manhattan Distance

A distância de *Manhattan* mede a dissimilaridade entre dois vetores de forma simples e eficiente:

$$d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n |x_i - y_i|$$

É apropriada para vetores de com diferentes domínios como binários, inteiros ou *floats* porque trata igualmente cada dimensão, sem amplificar desvios individuais.

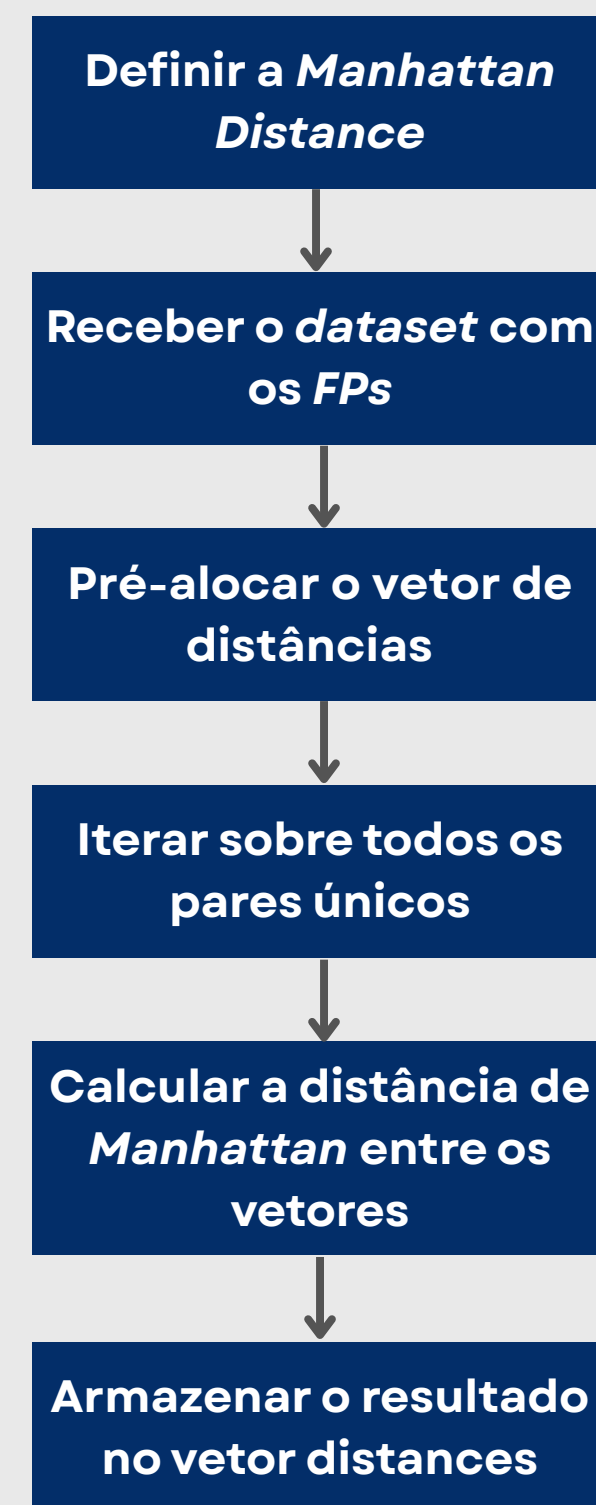
Procedimento

Para cada método de *fingerprint*, foram calculadas todas as distâncias de *Manhattan* entre pares únicos de moléculas. O vetor de distâncias obtido é depois normalizado para o intervalo [0,1]:

$$\text{sim}(x, y) = 1 - \frac{d(x, y) - \min(d)}{\max(d) - \min(d)}$$

Isto garante que:

- *Fingerprints* com diferentes escalas de valor possam ser comparadas de forma justa;
- Valores altos de distancia resultem em baixa similaridade, e vice-versa;
- Todos os métodos passem a gerar vetores de similaridade na mesma escala comparável.





PRÓXIMOS PASSOS

SIMILARIDADE BIOSSINTÉTICA

Pode ser entendida como a proximidade entre duas moléculas com base na sua origem bio sintética comum, por exemplo, por pertencerem à mesma via metabólica ou partilharem classes bio sintéticas e é avaliada de forma indireta, através da consistência entre os agrupamentos gerados por uma *fingerprint* e a partilha de vias bio sintéticas entre os compostos.

Distância Bio sintética

É definida como o número de passos enzimáticos que separam dois compostos numa via bio sintética. Esta medida reflete a proximidade funcional e bio sintética entre moléculas, considerando o percurso enzimático que leva de um precursor a um produto.[4]

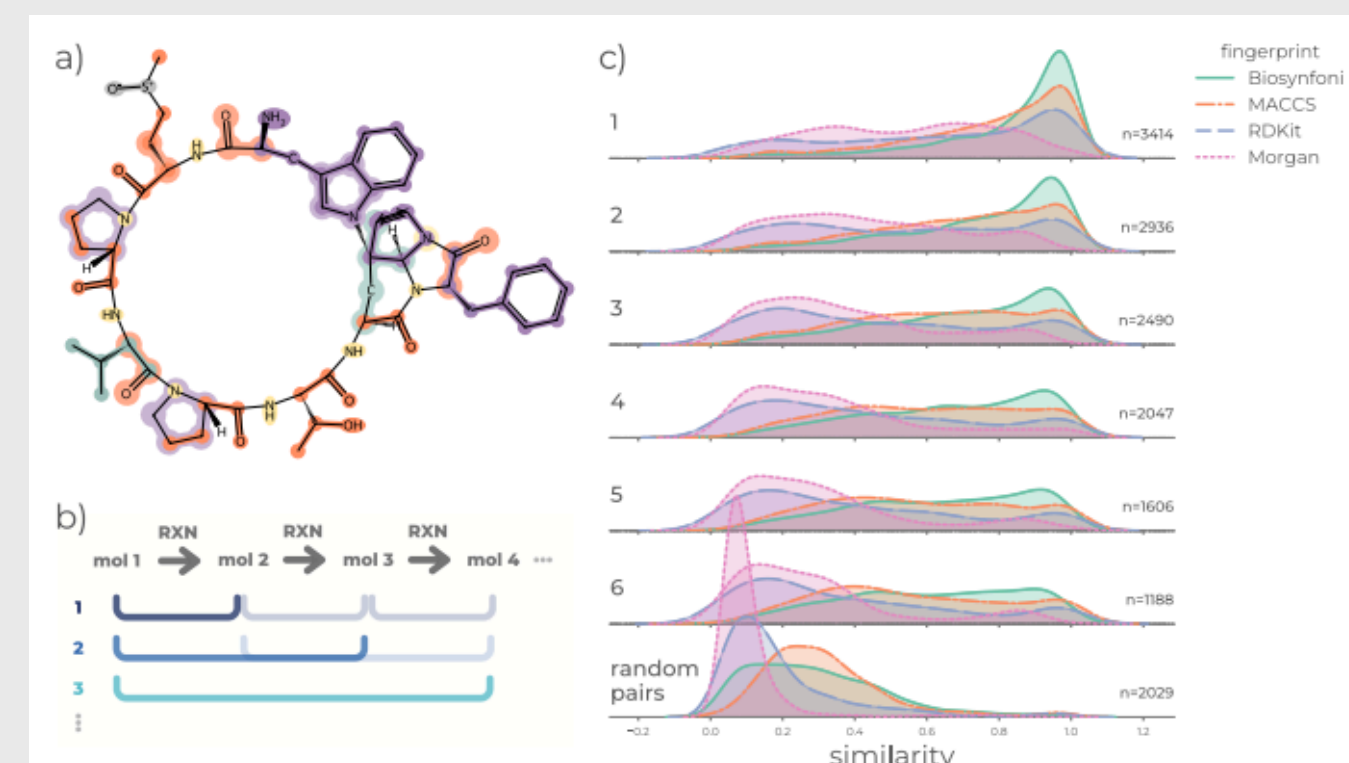


Fig1- c) Fingerprint similarity as a proxy for biosynthetic similarity per fingerprint. Nollen, M. A., Voß, K., Boecker, S., & Wessjohann, L. A. (2023). Biosynfoni: A biosynthesis-informed and interpretable lightweight molecular fingerprint. Journal of Cheminformatics, 15(1), 78.

CRIAR *BERT* MODELS PARA NPS



O NPBERT é uma representação molecular baseada em BERT treinada especificamente para produtos naturais. O modelo vai ser pré-treinado com 2 milhões de moléculas dos bancos ChEMBL e ZINC, utilizando a técnica de *Masked Language Modeling*.^[5]

Principais etapas do treino

- Tokenização: Moléculas são convertidas em strings de subestruturas químicas extraídas com RDKit.
- Embedding + Posição: Cada subestrutura é representada por um vetor + um embedding posicional, tal como no BERT original.
- Treino com MLM: Em 15% das posições, os tokens são mascarados “?” aleatoriamente, e o modelo aprende a prever os tokens corretos com base no contexto (semelhante ao treino de linguagem natural).

Neste caso o NPBERT já estava incorporado no Deepmol e foi treinado com o dataset criado.



CONSIDERAÇÕES FINAIS

- Foi desenvolvido um pipeline completo e flexível para comparação de *fingerprints* moleculares, suportando métodos binários, contínuos e inteiros.
- A utilização da distância de Manhattan e a normalização dos vetores permitiram comparações justas entre métodos com domínios numéricos distintos.
- A comparação focou-se até agora na dimensão estrutural e algorítmica, preparando o terreno para a futura análise de similaridade biossintética, baseada em anotações reais de vias metabólicas (KEGG e PlantCyc).
- O treino e integração do NPBERT será incorporado numa fase posterior, com o objetivo de avaliar o contributo de representações baseadas em modelos de linguagem.



REFERÊNCIAS BIBLIOGRÁFICAS

Capela, J., Rocha, M. (2024). Natural Products meet Masked Learning. Springer Lecture Notes in Computer Science.

Menke, J., Massa, J., Koch, O. (2021). Natural product scores and fingerprints extracted from artificial neural networks. CSBJ.

Nguyen-Vo, T.H., Trinh, Q.H., Nguyen, L., Do, T.T.T., Chua, M.C.H., Nguyen, B.P. (2021). Predicting Antimalarial Activity in Natural Products Using Pretrained Bidirectional Encoder Representations from Transformers. Journal of Chemical Information and Modeling, 62(21), 5050–5058.

Nollen, L.M., Meijer, D., Sorokina, M., van der Hooft, J. (2025). Biosynfoni: A Biosynthesis-informed and Interpretable Lightweight Molecular Fingerprint.

Zheng, X., Tomiura, Y. (2024). A BERT-based pretraining model for extracting molecular structural information from SMILES. Journal of Cheminformatics.



OBRIGADO