Training Large Language Models for Natural Products with Masked Learning Using DeepMol: A Benchmark for Similarity Overlap Resolution

Pedro Pereira[1][0000−1111−2222−3333], Second Author[2,3][1111−2222−3333−4444], and Third Author[3][2222−−3333−4444−5555]

[1] Escola de Engenharia da Universidade do Minho , Portugal
pg55703@alunos.uminho.pt
http://www.springer.com/gp/computer-science/lncs
[2] ABC Institute, Rupert-Karls-University Heidelberg, Heidelberg, Germany
{abc,lncs}@uni-heidelberg.de

Abstract. The abstract should briefly summarize the contents of the paper in 150–250 words.

Keywords: First keyword · Second keyword · Another keyword.

1 Context and Motivation

The rapid advancement of deep learning (DL) and natural language processing (NLP) has significantly impacted various fields, including computational chemistry and bioinformatics. One of the emerging challenges in these domains is the representation and analysis of molecular structures of natural products, which play a crucial role in drug discovery and biotechnology. Traditional methods for molecular similarity assessment, such as 2D and 3D fingerprinting techniques, have been extensively used. However, these methods often struggle to capture complex molecular relationships, especially in the case of natural products, which frequently display high stereochemical complexity, rich functional group diversity, and non-linear scaffold architectures that arise from their biosynthetic origins (Johnson and Maggiora, 2006) [14].

A central challenge in computational chemistry is the accurate assessment of molecular similarity. Determining how similar two molecules are is fundamental for tasks such as virtual screening, clustering, and property prediction. However, molecular similarity is a multifaceted concept that depends heavily on the type of representation and the metric used. For instance, two molecules may appear dissimilar using 2D fingerprints but may share similar 3D conformations or biosynthetic origins. This ambiguity, often referred to as the "molecular similarity problem" (Johnson and Maggiora, 2006) [14], complicates the design of universal similarity methods and highlights the need for more expressive, data-driven molecular representations.

To explore alternative molecular representations, transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) have

been adapted to the chemical domain, particularly using SMILES strings as input. Examples of such adaptations include ChemBERTa, a self-supervised model pretrained on large chemical datasets for property prediction (Chithrananda et al., 2020) [9], and NPBERT, a transformer model tailored for natural products using masked substructure prediction (Nguyen-Vo et al., 2021) [7]. These models use a masked language modeling (MLM) approach to learn context-aware molecular embeddings by predicting masked tokens within SMILES strings. Preliminary studies have shown that such embeddings may capture structural patterns relevant for molecular property prediction and classification tasks (Zheng and Tomiura, 2024) [3]. However, their application in tasks such as molecular similarity assessment or virtual screening is still limited and lacks systematic benchmarking against more established methods such as graph neural networks (GNNs), which tend to outperform them in most scenarios (Menke et al., 2021; Bolcato et al., 2022)[6, **?**].

This project focuses on the training of large language models (LLMs) tailored for natural products, leveraging masked learning techniques within the DeepMol framework. By benchmarking various similarity overlap resolution strategies, we aim to improve the classification and retrieval of natural product molecules. The study explores different molecular similarity metrics, including 2D, 3D, and biosynthetic similarities, and evaluates the role of deep learning models in enhancing molecular feature extraction.

2 State-of-the-art

2.1 Similarity in computational chemistry

Molecular similarity refers to the degree to which two molecular structures resemble each other based on defined criteria such as structural, physicochemical, or electronic properties (Johnson Maggiora, 2006) [14]. It plays a central role in cheminformatics and computational chemistry, particularly in tasks such as virtual screening, clustering, and property prediction, under the assumption that structurally similar molecules tend to exhibit similar biological or physicochemical behaviors (Willett, 2003)[**?**]. In drug design, molecular similarity enables virtual screening by identifying compounds structurally similar to known actives. In chemical space exploration, it helps cluster or map molecules based on shared features, facilitating diversity analysis. For property prediction, similarity metrics allow inference of properties for unknown compounds based on their proximity to annotated molecules (Johnson and Maggiora, 2006)[14].

Similarity Measures Several metrics are commonly used to assess similarity or dissimilarity between molecules, depending on the type of molecular representation (e.g., binary fingerprints or continuous descriptors). Among the most used are the Tanimoto and Dice similarity for binary vectors and Euclidean distance and cosine similarity for continuous-valued descriptors.

The **Tanimoto coefficient**, also known as the Jaccard index, is frequently employed to compare binary fingerprints. It is defined as:

$$T(A, B) = \frac{c}{a + b - c} \tag{1}$$

where a and b are the number of bits set to 1 in fingerprints A and B, respectively, and c is the number of bits set to 1 in both A and B. This coefficient measures the proportion of shared features relative to the total number of features present in both molecules. Its value ranges from 0 (no similarity) to 1 (identical).

The **Dice similarity**Dice similarity is an alternative similarity metric that places more emphasis on the shared characteristics between two fingerprints. It is defined as:

$$D(A, B) = \frac{2c}{a + b} \tag{2}$$

This metric gives twice the weight to the number of shared features (c), compared to the Tanimoto coefficient, making it more sensitive to small overlaps, particularly in sparse fingerprints.

For continuous descriptor vectors, the **Euclidean distance** is a standard dissimilarity measure that calculates the geometric distance between two points in a multidimensional space. It is computed as:

$$d_E(A, B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \tag{3}$$

where A_i and B_i are the i-th components of the descriptor vectors A and B. Lower values of d_E indicate greater similarity.

Another metric for continuous data is the **cosine similarity**, which measures the cosine of the angle between two vectors. It is defined as:

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4}$$

This metric evaluates the orientation rather than the magnitude of vectors, making it suitable for high-dimensional data. A cosine similarity of 1 implies identical orientation (maximum similarity), while 0 indicates orthogonality (no similarity).

These metrics are foundational in cheminformatics for applications such as virtual screening, clustering of chemical libraries, and similarity-based property prediction.

2.2 Global vs. Local Similarity

Global Similarity Global similarity measures the overall resemblance between entire molecular structures, considering the molecule as a whole. This type of

similarity is useful in the virtual screening and clustering of chemical libraries, where structurally related molecules can be grouped based on their global features (Willett, 2003) [**?**].

Local Similarity Local similarity focuses on specific regions or substructures within molecules, such as functional groups or active sites. This approach is particularly relevant when biological activity is determined by localized molecular interactions, enabling the identification of compounds that exhibit similar behaviour despite differing globally.

2.3 2D Similarity

2D similarity compares molecules based on their two-dimensional structural representations, focusing on atom connectivity and bond patterns. Methods such as molecular fingerprints (e.g., ECFP, MACCS keys, or topological fingerprints) encode the presence or absence of specific substructures as binary vectors. To quantify similarity between these fingerprints, several metrics can be used — the most common being the Tanimoto coefficient and the Dice similarity coefficient, which compare the overlap between feature sets. Additionally, other metrics such as cosine similarity or Hamming distance may also be applied, although they are less common in cheminformatics. While 2D similarity methods are computationally efficient and scalable, they may fail to capture crucial three-dimensional aspects of molecular interactions, such as stereochemistry, conformational flexibility, and spatial arrangements (Willett, 2003)[**?**].

3D Similarity While 3D similarity assessments can provide a more refined and detailed analysis of molecular relationships compared to 2D methods—by capturing spatial arrangement, molecular shape, and electronic distributions—they also introduce notable challenges. Generating 3D structures requires conformer generation, which can be computationally intensive, especially for flexible molecules. Moreover, a single molecule can adopt multiple low-energy conformations, and it is often unclear which conformation corresponds to the biologically active form (Wawer and Bajorath, 2011; Bolcato et al., 2022) [**?**,**?**]. These limitations can affect the robustness and interpretability of 3D similarity-based analyses.

Biosynthetic Similarity Biosynthetic similarity evaluates molecules based on their biosynthetic origins and the enzymatic pathways through which they are produced. Compounds synthesized via similar biosynthetic routes often share core structural motifs and may exhibit related biological activities. To computationally assess this form of similarity, the Biosynfoni method introduces a manually curated fingerprint comprising 39 biosynthetically relevant substructures, inspired by classical biosynthetic logic as outlined by Dewick (Nollen et al., 2025)[5]. These substructures were selected to reflect common building blocks and intermediates found in the biosynthesis of natural products, such as polyketides, terpenoids, and alkaloids.

Each molecule is represented as a binary vector indicating the presence or absence of these biosynthetic fragments. Biosynthetic proximity is then quantified by comparing these fingerprints using conventional similarity metrics — most commonly the Tanimoto coefficient — to determine the degree of overlap in biosynthetic building blocks between two molecules. A high biosynthetic proximity score suggests that two molecules may derive from related biosynthetic pathways, even if their overall structures differ significantly. This approach provides a more mechanistically grounded similarity measure, which is particularly valuable in natural product research for scaffold hopping, compound prioritization, and the discovery of novel bioactive entities.

2.4 BERT for masked learning applied to chemistry

The application of deep learning models based on transformer architectures, such as BERT , has gained attention in computational chemistry as a strategy to learn molecular representations from SMILES strings. These models are trained through masked language modeling tasks, enabling them to capture syntactic and structural patterns in molecular sequences. Although early studies have shown promise in property prediction and virtual screening scenarios (Zheng and Tomiura, 2024) [3], the overall success of BERT-based models in chemistry remains limited. In comparative benchmarks, they have often underperformed when compared to graph neural networks, which more naturally encode molecular topology and connectivity (Menke et al., 2021; Bolcato et al., 2022) [6, **?**].

SMILES The Simplified Molecular Input Line Entry System (SMILES) is a compact and widely adopted textual notation used to represent molecular structures as linear sequences of characters. This format enables efficient storage, manipulation, and processing of chemical compounds, making it particularly well-suited for cheminformatics and machine learning applications (Weininger, 1988) [11].

In SMILES strings, atoms are represented by their chemical symbols, e.g. C for carbon, O for oxygen, and H for hydrogen. Importantly, uppercase letters denote aliphatic atoms, while lowercase letters are used for aromatic atoms (e.g. c for aromatic carbon), which allows SMILES to directly encode aromaticity (O'Boyle, 2012)[12].

Chemical bonds are represented using specific characters: a single bond is either implicit or indicated by -, a double bond by =, and a triple bond by #. Branching is expressed using parentheses — for instance, CC(O)C represents a carbon chain with a hydroxyl side group. Ring closures are denoted by numeric labels, where the same number indicates the start and end of a ring, as in C1CCCC1.

Stereochemistry is encoded using symbols such as / and for cis-trans isomerism around double bonds, and @ or @@ to indicate tetrahedral chirality at stereocenters. This encoding allows SMILES to preserve three-dimensional structural information in a linear format.

Overall, SMILES provides a balance between human readability and machine interpretability, and its sequence-based structure makes it especially compatible with natural language processing models.

2.5 SMILES Tokenization in Chemical Language Models

Unlike traditional NLP tasks that use WordPiece or Byte-Pair Encoding, most chemical language models do not rely on general-purpose subword tokenization. Instead, they adopt chemically meaningful tokenization strategies adapted to the specific structure and syntax of SMILES .

For instance, in **ChemBERTa** (Chithrananda et al., 2020), tokenization is performed using a custom vocabulary built from characters or chemically relevant SMILES substrings. The tokenizer splits SMILES into atom-level or group-level tokens, including atoms (e.g., `C`, `Cl`), rings (`1`, `2`), branches (`(`, `)`), and stereochemical indicators (`@`, `/`, `\`). This tokenization preserves the syntactic and chemical structure of the molecule while allowing the model to learn meaningful patterns during masked language modeling.

In contrast, **NPBERT** (Nguyen-Vo et al., 2021) [7] does not operate directly on SMILES strings. Instead, it uses atom-level feature vectors derived from **Morgan fingerprints** with radius 2. These features describe the chemical environment of each atom and are used as input tokens, enabling the model to leverage established cheminformatics descriptors while applying transformer-based architectures.

These tokenization strategies ensure that chemical language models operate on inputs that reflect molecular structure more accurately than standard NLP tokenizers.

How BERT Works BERT is powered by a deep neural network architecture known as Transformers. This architecture incorporates a mechanism called "self-attention", which allows BERT to measure the importance of each word based on its surrounding context, both preceding and following it. This bidirectional training is BERT's key innovation, enabling it to generate contextualized word embeddings—representations of words that capture their meanings within a given sentence.

BERT Training BERT models are generally trained using a technique called *masked language modeling* (MLM), in which some words or characters within the input sequence are randomly masked and replaced with a special `[MASK]` token. The model is trained to predict these masked elements based solely on their surrounding context. This learning strategy allows BERT to capture bidirectional dependencies across the entire sequence, enabling a deeper understanding of contextual relationships between tokens.

The training objective is optimized using the **cross-entropy loss function**, which quantifies the difference between the predicted probability distribution and

the true distribution of the masked tokens (typically one-hot encoded). Cross-entropy penalizes incorrect predictions more heavily when the model assigns high confidence to the wrong class, encouraging more accurate and calibrated outputs (Murphy, 2012) [10]. The loss for a single sample is given by:

$$Cross - Entropy = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \tag{5}$$

where C is the number of classes (vocabulary size), y_i is the ground truth label (1 for the correct token, 0 otherwise), and $\hat{y}_i$ is the predicted probability for token i.

3 Natural Products and Their Representations

Natural Products (NP) constitute a diverse class of molecules with high structural complexity, often containing fused ring systems, multiple chiral centres, and rigid frameworks. Their diversity arises from multiple biosynthetic pathways, leading to specialized bioactive compounds that have been widely exploited in pharmaceuticals, agriculture, and biotechnology. Several computational representations have been developed to capture the distinct structural, functional, and biosynthetic characteristics of natural products (NPs). Unlike traditional fingerprints, these methods often encode biologically relevant features, such as biosynthetic building blocks or learned embeddings based on large datasets. Table 1 provides a comparative overview of the most prominent NP-specific representations, including structural fingerprints, biosynthetic fragment keys, and deep learning-based embeddings.

These representations provide alternative ways to encode natural products depending on the application. NC-MFP and Biosynfoni focus on capturing interpretable structural or biosynthetic features, while NP AUX, NP AE and NPBERT leverage data-driven deep learning techniques to learn embeddings useful for tasks such as classification or virtual screening.

References

1. @articlecorreia2024deepmol, title=Deepmol: an automated machine and deep learning framework for computational chemistry, author=Correia, João and Capela, João and Rocha, Miguel, journal=Journal of Cheminformatics, volume=16, number=1, pages=1–17, year=2024, publisher=Springer
2. @articlebolcato2022value, title=On the value of using 3D shape and electrostatic similarities in deep generative methods, author=Bolcato, Giulia and Heid, Esther and Boström, Jonas, journal=Journal of Chemical Information and Modeling, volume=62, number=6, pages=1388–1398, year=2022, publisher=ACS Publications
3. @articlezheng2024bert, title=A BERT-based pretraining model for extracting molecular structural information from a SMILES sequence, author=Zheng, X and Tomiura, Y, journal=Journal of Cheminformatics, volume=16, number=1, pages=1–12, year=2024, publisher=Springer

Table 1. Summary of computational representations of natural products

Name	Type	Size / Dim.	Generation Method	Datasets Used
NC-MFP	Structural fingerprint	∼10,016 bits	Based on hierarchical scaffold decomposition (Bemis-Murcko), fragment SMARTS, and DNP classification [4]	Dictionary of Natural Products (DNP)
Biosynfoni	Biosynthetic fragment key	39 bits	Manually curated biosynthetic fragments based on Dewick's logic [5]	NA (not applicable)
NP AUX	Neural network (MLP)	64-dim vector	Multi-layer perceptron trained with auxiliary tasks (e.g., descriptor prediction) [6]	COCONUT + ZINC decoys
NP AE	Neural network (Autoencoder)	64-dim vector	Autoencoder trained to reconstruct ECFP4 fingerprints [6]	COCONUT + ZINC decoys
NPBERT	Transformer-based model	512-dim vector	BERT trained on SMILES using masked learning with substructure tokens [7]	ChEMBL + ZINC

4. @articleseo2020development, title=Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP) for natural product-based drug development, author=Seo, Myungwon and Shin, Hye Kyung and Myung, Yoochan and Hwang, Seung and No, Kyoung Tai, journal=Journal of Cheminformatics, volume=12, number=1, pages=1–14, year=2020, publisher=Springer
5. @articlenollen2025biosynfoni, title=Biosynfoni: A Biosynthesis-informed and Interpretable Lightweight Molecular Fingerprint, author=Nollen, Lucina-May and Meijer, David and Sorokina, Maria and van der Hooft, Justin JJ, journal=ChemRxiv, year=2025, publisher=Cambridge Open Engage
6. @articlemenke2021natural, title=Natural product scores and fingerprints extracted from artificial neural networks, author=Menke, Janosch and Massa, Joana and Koch, Oliver, journal=Computational and Structural Biotechnology Journal, volume=19, pages=4593–4602, year=2021, publisher=Elsevier
7. @articlenguyen2021predicting, title=Predicting antimalarial activity in natural products using pretrained bidirectional encoder representations from transformers, author=Nguyen-Vo, Thanh-Hoang and Trinh, Quang H and Nguyen, Loc and Do, Trang TT and Chua, Michael CH and Nguyen, Bao P, journal=Journal of Chemical Information and Modeling, volume=62, number=21, pages=5050–5058, year=2021, publisher=ACS Publications
8. @inproceedingsdevlin2019bert, title=BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, author=Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina, booktitle=Proceedings of NAACL-HLT 2019, pages=4171–4186, year=2019

9. @articlechithrananda2020chemberta, title=ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction, author=Chithrananda, Seyone and Grand, Gabriel and Ramsundar, Bharath, journal=arXiv preprint arXiv:2010.09885, year=2020
10. @bookmurphy2012machine, title=Machine Learning: A Probabilistic Perspective, author=Murphy, Kevin P, year=2012, publisher=MIT Press
11. @articleweininger1988smiles, title=SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules, author=Weininger, David, journal=Journal of Chemical Information and Computer Sciences, volume=28, number=1, pages=31–36, year=1988, publisher=ACS Publications
12. @articleoboyle2012towards, title=Towards a universal SMILES representation—a standard method to generate canonical SMILES based on the InChI, author=O'Boyle, Noel M, journal=Journal of Cheminformatics, volume=4, number=1, pages=1–14, year=2012, publisher=Springer
13. @articlewillett2003similarity, title=Similarity-based virtual screening using 2D fingerprints, author=Willett, Peter, journal=Drug Discovery Today, volume=8, number=12, pages=626–633, year=2003, publisher=Elsevier
14. @bookjohnson2006concepts, title=Concepts and Applications of Molecular Similarity, author=Johnson, Mark A and Maggiora, Gerald M, year=2006, publisher=Wiley-Interscience