Natural Products meet Masked Learning

João Capela[1†] and Miguel Rocha[1,2*]

[1]CEB - Centre of Biological Engineering, University of Minho, Braga, Portugal.
[2]LABBELS - Associate Laboratory, Braga/Guimarães, Portugal.

*Corresponding author(s). E-mail(s): mrocha@di.uminho.pt;
Contributing authors: joao.capela@ceb.uminho.pt;
†These authors contributed equally to this work.

Abstract

1 Introduction

Natural Products (NP) constitute an important class of known molecules with high structural diversity, tending to be more rigid, have more fused ring systems, and have more chiral centres compared to synthetic molecules [1]. Such diversity is the result of the involvement of multiple biosynthetic pathways and tailoring reactions [2]. Their diverse and specialised nature, crafted by evolution, provides a source of bioactive properties that humans have extensively exploited for agriculture and pharmaceuticals.

Efforts have been made to represent their unequivocal diversity and difference from synthetic molecules computationally. Molecular structural keys are 1D representations [3] that have been explored for representing NPs [4, 5]. They are vectors of integers where each index of the vector represents one specific substructure. Then, an algorithm that normally resorts to SMARTS notations, captures or counts the presence or absence of these substructures in a given molecule.

Regardless of being referred to as The Natural Compound Molecular Fingerprint (NC-MFP) [4], this was the first structural key to be developed for NPs. It was designed to better capture the structural characteristics of natural compounds by incorporating scaffolds, scaffold-fragment connection points, and fragments derived from the Dictionary of Natural Products (DNP) classification system. Using a hierarchical scaffold approach to identify the scaffolds, multiple structural levels are generated using the Bemis and Murko (BM) method, where larger molecular frameworks are progressively

broken down into simpler core structures. In NC-MFP, this approach creates a scaffold tree with different levels and prioritizes the ones belonging to classes in the Dictionary of Natural Products (DNP). However, the development and reimplementation of NC-MFP are hampered by disconnected processing stages, software dependencies, and the risk of data leakage during structural key generation. This occurs because connection points and fragments are not dataset-agnostic, being tailored-generated. In the original code, the structural key is generated with the whole dataset for classification tasks, disregarding data splitting, so there is the risk of leaking information of the test set to ML models.

Another fragment key designed specifically for NPs is Biosynfoni. It consists of 39 substructure keys that capture information of the building blocks and patterns described in Dewick's 2009 book on NP biosynthetic logic. This structural key outperforms some of the most used fingerprints (i.e. Morgan and RDKit fingerprints, and MACCS keys) in capturing biosynthetic distance and performs on par for predicting NP class.

Another type of NP representation is neural-network-based, where a Neural Network (NN) is trained either in a supervised or self-supervised manner to extract an embedded chemical/molecular representation. The former refers to a way of training NNs, where the model is trained to predict specific categorical and/or discrete endpoints (e.g. whether a compound is an NP with binary/multi-class classification and/or physico-chemical properties with regression). The latter concept refers to training an NN to learn the inputted representation and does not need for curated labels assigned to each molecule.

NP_AUX and NP_AE, described in [1], are an NN-generated molecular representation designed for NPs. It is trained on a curated dataset of natural products from COCONUT and synthetic decoys from ZINC, selected based on Tanimoto similarity and Natural Product Likeness (NPL) score. The fingerprint is extracted from the activations of the last hidden layer of a multi-layer perceptron (MLP) or an autoencoder-like (AE) model, with a 64-dimensional representation encoding natural product-relevant features. Three architectures were tested: a baseline MLP, an auxiliary-task MLP (NP_AUX) that predicts molecular descriptors, and an autoencoder (NP_AE) reconstructing ECFP4 fingerprints. The neural fingerprint outperforms traditional fingerprints like ECFP4 and NC-MFP in natural product classification, similarity searches, and virtual screening, as shown by improved AUC and Enrichment Factor (EF1%). This approach provides a data-driven alternative to handcrafted fingerprint.

In the era where language models (LMs) are gaining traction, NPBERT was developed [6], a transformer-based molecular embedding trained on data from the ChEMBL and ZINC databases to encode natural products in a 512-dimensional latent space. These embeddings were used to train machine learning (ML) models to predict antimalarial activity. This embedding was generated based on a masked learning task, where a BERT model was trained to predict tokens in a sentence. In this case, tokens are substructure keys generated by Morgan fingerprints for each atom in a molecule, and a sentence is the concatenation of all the molecular tokens. By learning how to predict masked substructures based on the context, this model learns the interdependence of the molecule substructures.

Although several representations of NP were explored before, they lack comparative evaluations for ML

2 Methods

3 Results

4 Discussion: comparison with other tools

5 Conclusion

6 Declarations

6.1 Availability of data and materials

6.2 Competing interests

6.3 Funding

6.4 Authors' contributions

6.5 Acknowledgements

References

[1] Menke J, Massa J, Koch O. Natural product scores and fingerprints extracted from artificial neural networks. Computational and Structural Biotechnology Journal. 2021;19:4593–4602.

[2] Kim HW, Wang M, Leber CA, Nothias LF, Reher R, Kang KB, et al. NPClassifier: a deep neural network-based structural classification tool for natural products. Journal of Natural Products. 2021;84(11):2795–2807.

[3] David L, Thakkar A, Mercado R, Engkvist O. Molecular representations in AI-driven drug discovery: a review and practical guide. Journal of Cheminformatics. 2020;12(1):56.

[4] Seo M, Shin HK, Myung Y, Hwang S, No KT. Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP) for natural product-based drug development. Journal of Cheminformatics. 2020;12(1):6.

[5] Nollen LM, Meijer D, Sorokina M, van der Hooft J. Biosynfoni: A Biosynthesis-informed and Interpretable Lightweight Molecular Fingerprint. 2025;.

[6] Nguyen-Vo TH, Trinh QH, Nguyen L, Do TT, Chua MCH, Nguyen BP. Predicting antimalarial activity in natural products using pretrained bidirectional encoder

representations from transformers. Journal of Chemical Information and Modeling. 2021;62(21):5050–5058.