

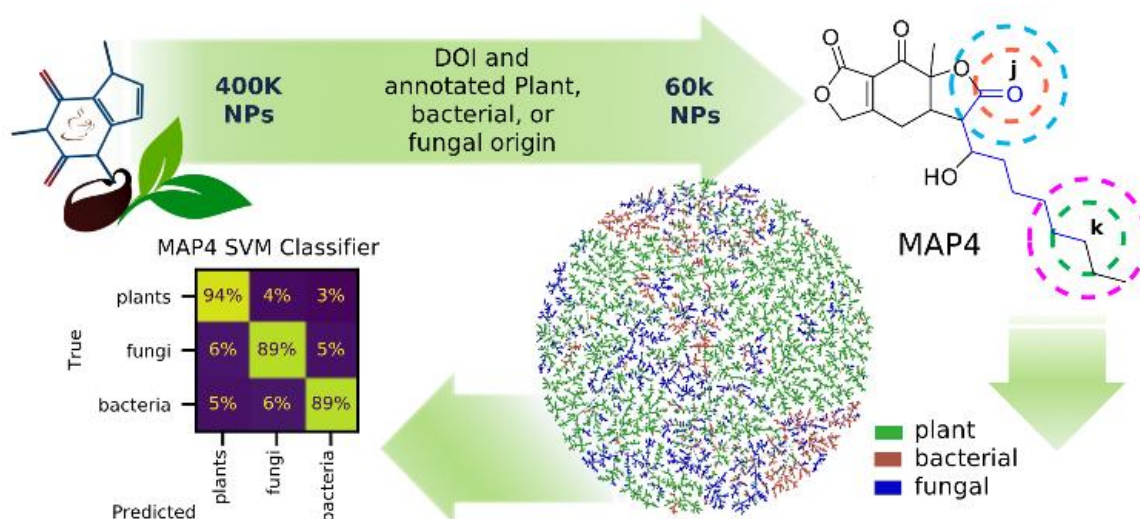
Classifying Natural Products from Plants, Fungi or Bacteria in the COCONUT Database

Alice Capecchi¹, and Jean-Louis Reymond*¹

¹ *Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern*

* Email: jean-louis.reymond@dcb.unibe.ch

Graphical Abstract



Abstract

Natural products (NPs) represent one of the most important resources for discovering new drugs. Here we asked whether NP origin can be assigned from their molecular structure in a subset of 60,171 NPs in the recently reported Collection of Open Natural Products (COCONUT) database assigned to plants, fungi, or bacteria. Visualizing this subset in an interactive tree-map (TMAP) calculated using MAP4 (MinHashed atom pair fingerprint) clustered NPs according to their assigned origin (https://tm.gdb.tools/map4/coconut_tmap/), and a support vector machine (SVM) trained with MAP4 correctly assigned the origin for 94% of plant, 89% of fungal, and 89% of bacterial NPs in this subset. An online tool based on an SVM trained with the entire subset correctly assigned the origin of further NPs with similar performance (<https://np-svm-map4.gdb.tools/>). Origin information might be useful when searching for biosynthetic genes of NPs isolated from plants but produced by endophytic microorganisms.

Keywords: natural products, databases, cheminformatics, chemical space, visualization, molecular fingerprints, machine learning, support vector machine, origin classification.

Introduction

Due to the importance of natural products (NPs) in drug discovery,[1, 2] there is a considerable interest in describing and understanding their structural diversity, particularly by exploiting NP databases[3] using *in silico* methods such as machine learning (ML).[4] Computational approaches have been reported to distinguish between NPs and non-Nps,[5–9] between terrestrial and marine NPs,[10] and to classify NP structural types[11, 12] and visualize their chemical space.[13]

In our own approach to this problem,[14] we recently analyzed NPAtlas, an open-access database listing 25,523 NPs from bacterial or fungal origin,[15] by computing the MAP4 fingerprint (MinHashed Atom-Pair fingerprint up to four bonds)[16] of each NP and creating a TMAP (tree-map)[17] of the resulting high-dimensional dataset. In this analysis, NPs from bacterial or fungal origin formed separated clusters. This separation effect was confirmed by showing that a support vector machine (SVM) trained with the MAP4 of NPAtlas was able to distinguish bacterial or fungal origin, including a recently reported NP isolated from the marine sponge *Phakellia fusca* assigned by our classifier to be of bacterial origin, in line with the fact that many NPs from this sponge originate from endosymbiotic actinobacteria.[18, 19]

The possibility to assign the origin of NPs from their structure was intriguing because most NPs are secondary metabolites produced by biosynthetic gene clusters[20] which are sometimes transferred between different organisms.[21] Such horizontal gene transfer may reflect adaptative relationships between host organisms such as plants and sponges and endosymbiotic bacteria or fungi.[22] Among the many endophytic NPs,[23, 24] striking examples include the cancer drug paclitaxel, a plant NP also produced by endophytic fungi of the yew tree,[25, 26] and maytansine, used in antibody-drug conjugates for cancer therapy

and produced by endophytic bacteria in plants.[27] Due to the very widespread occurrence of endophytic bacteria and fungi in plants, we asked whether our MAP4 analysis might be able to distinguish plant NPs from bacterial and fungal NPs. To test this hypothesis, we considered the recently reported COCONUT database, a recently reported open-access database currently offering the most extensive coverage and including plant NPs.[28]

Results and discussion

Chemical space analysis of plant and microbial NPs from the COCONUT database

COCONUT collects over 400 thousand NPs from 52 different databases, 135 thousand of which are annotated with a taxonomical origin. For our analysis, we considered the 68 thousand entries annotated with a source organism that were also associated with a publication. We focused on those annotated as originating from plants (50 %), fungi (23 %), or bacteria (16 %), leaving out a smaller subset of NPs originating from animals (2 %), homo sapiens (2.5 %), of marine origin (1.5 %), or lacking a superclass annotation (5 %). The selected subset of 60,171 NPs contained 33,772 plant NPs, 15,648 fungal NPs and 10,751 bacterial NPs.

The subset spanned from molecular weight $MW = 81$ Da for 1,2-dihydropyridine, a plant NP,[29] to $MW = 2,901$ Da for lacticin 481, a bacterial peptide.[30] Plant NPs dominated the intermediate molecular weight range ($200 < MW < 800$), while fungal NPs were most abundant in the low molecular weight range ($MW \leq 200$) and bacterial NPs in the high MW range ($MW \geq 800$). The three series had rather similar distributions of the fraction of sp^3 carbon atoms (F_{sp3}), which measures the degree of saturation. However, the estimated octanol:water partition coefficient $AlogP$ indicated that highly polar NPs were almost absent from fungal NPs. Furthermore, plant NPs had overall higher percentages of glycosides, while peptides were almost absent from plant NPs and most abundant in bacterial NPs (Table 1).

Table 1. Property distribution and origin of the 60,171 COCONUT entries with a DOI and annotated as plants, fungal, or bacterial.

	Plants NPs ^{a)}	Fungal NPs ^{a)}	Bacterial NPs ^{a)}
MW \leq 200 ^{b)}	7,072 (21%)	4,919 (31%)	2,237 (21%)
200 \leq MW $<$ 800 ^{b)}	24,078 (71%)	10,111 (65%)	6,066 (56%)
MW \geq 800 ^{b)}	2,622 (8%)	618 (4%)	2,448 (23%)
Fsp3 \leq 0.2 ^{c)}	4,213 (13%)	1,580 (10%)	1,073 (10%)
0.2 \leq Fsp3 $<$ 0.8 ^{c)}	22,032 (65%)	11,334 (72%)	7,986 (74%)
Fsp3 \geq 0.8 ^{c)}	7,527 (22%)	2,734 (18%)	1,692 (16%)
AlogP \leq -2 ^{d)}	4,855 (14%)	373 (2%)	1,446 (13%)
-2 \leq AlogP $<$ 8 ^{d)}	28,315 (84%)	15,000 (96%)	8,906 (83%)
AlogP \geq 8 ^{d)}	602 (2%)	275 (2%)	399 (4%)
Glycosides ^{e)}	8,260 (24%)	797 (5%)	1,793 (17%)
Peptides ^{f)}	194 ($<$ 1%)	676 (4%)	2,053 (19%)

a) COCONUT entries with a DOI and the specified taxonomical origin annotated; percentages refer to the total number of the selected entries within the specified class: 33,772 plants NPs, 15,648 fungal NPs, and 10,751 bacterial NPs. b) Molecular weight (MW) calculated with RDKit. c) Fraction of sp3 (Fsp3) calculated with RDKit. d) Octanol: water partition coefficient calculated with RDKit following the Crippen method (AlogP). e) Containing a cyclic N- or O-acetal substructure defined through SMARTS language. f) Containing a dipeptide substructure defined through SMARTS language.

To get a closer insight into structural features, we calculated the MAP4 fingerprint for each of the 60,171 selected NPs. MAP4 encoding combines the characteristics of substructure fingerprints, which are well suitable for small molecules, and of atom pair fingerprints, which are instead preferable for larger structures, and it has been proven suitable for both.[16] It consists of listing all pairs of circular substructures of radius 1 and 2 as SMILES, separated by their topological distance in bonds, and MinHashing the resulting set of SMILES pairs to a defined dimensionality (1024 in the present analysis). We then represented the MAP4 annotated NP dataset using the dimensionality reduction method TMAP. This method is suitable for very large high-dimensional datasets and performs better than t-SNE or UMAP in preserving local and global relationships in the data.[17] To create a TMAP, the algorithm computes an approximate nearest neighbor graph by locality sensitive hashing (LSH), cuts edges to obtain the minimum spanning tree of this graph, and creates an optimized 2D representation of the minimum spanning tree, in which each node represent a molecule connected to its approximate nearest neighbors. This tree is then displayed with interactive

visualization tool Faerun. [31] Faerun shows each node as a sphere that can be color-coded according to various properties and uses Smilesdrawer[32] to depict molecular structures.

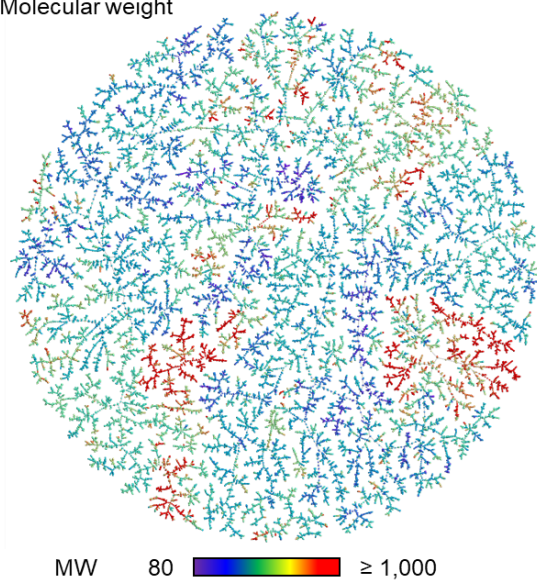
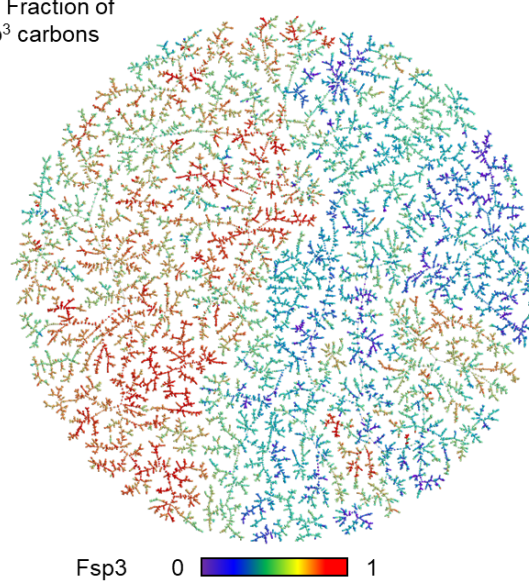
The TMAP of our NP subset is available interactively at

https://tm.gdb.tools/map4/coconut_tmap/.

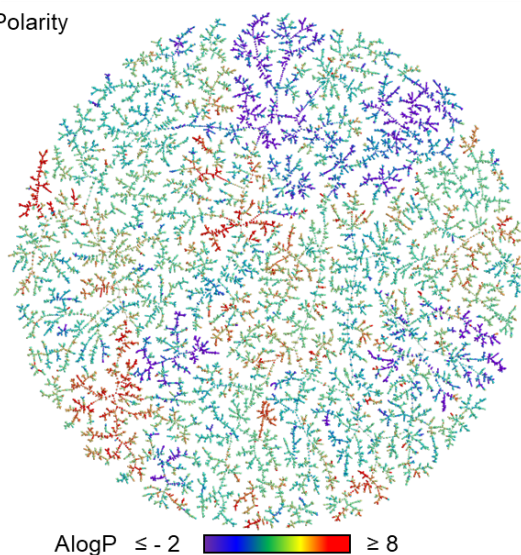
The TMAP of our NP subset color-coded by MW showed that most high MW compounds appeared in two groups, the first one (at right on the TMAP), contained peptides and related macrocycles, and the second one (at middle/lower left on the TMAP) corresponded to glycosylated triterpenoids (Figure 1a). Color-coding by Fsp3 showed that the TMAP separated high Fsp3 molecules (left half of the TMAP), comprising many terpenes, steroids, and glycosides, from low Fsp3 molecules (right half of the TMAP) featuring many polyphenols and related polyaromatic molecules (Figure 1b). Furthermore, the color-code by the calculated octanol:water partition coefficient AlogP, estimating polarity, showed several islands of highly polar NPs (low AlogP, magenta) corresponding mostly to nucleosides and glycosylated polyphenols (upper part of the TMAP), glycosylated triterpenoids (lower left on the TMAP) and peptides (middle right on the TMAP), as well as a few groups of apolar NPs (high AlogP, red), corresponding primarily to lipids, terpenes, and steroids (Figure 1c).

Color-coding by the annotated origin showed that NPs from plants, fungi, or bacteria formed many well-defined clusters spread across the entire TMAP (Figure 1d). On the one hand, this separation illustrated how NP origin corresponded to differences in molecular structure that were well perceived by the MAP4 fingerprint used to generate the map. On the other hand, the taxonomical origin color code also showed that each subset contained diverse structural types. While there was no correlation of origin with properties such as MW, Fsp3, or AlogP, most glycosides were associated with plants, and most peptides were of bacterial or fungal origin, in line with Table 1 (Figure 1e). These relationships were also well visible by color-coding the TMAP by prioritized categories (Figure 1f).

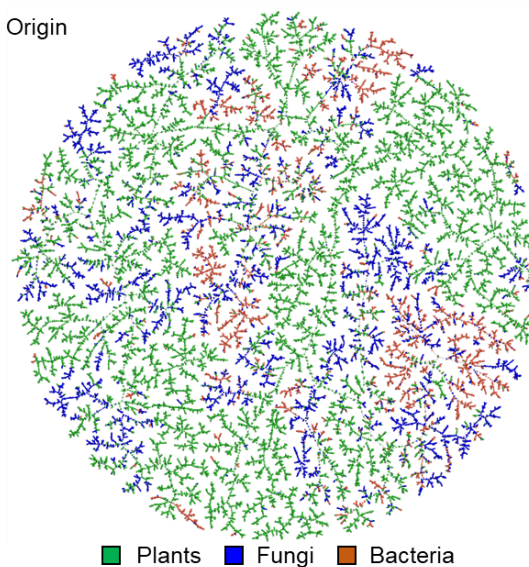
a) Molecular weight

b) Fraction of sp^3 carbons

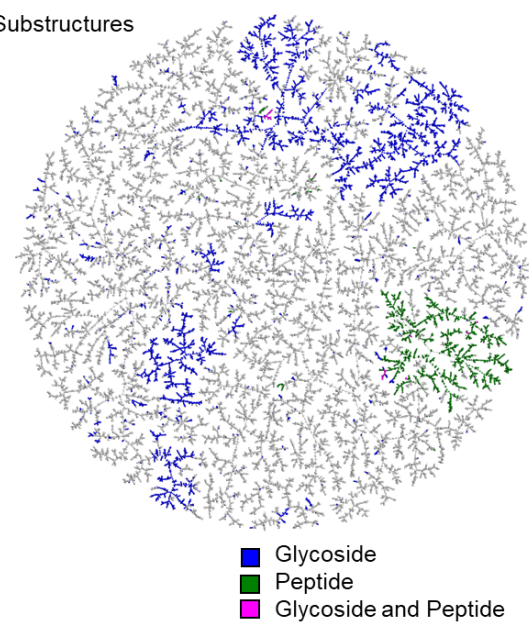
c) Polarity



d) Origin



e) Substructures



f) Prioritized categories

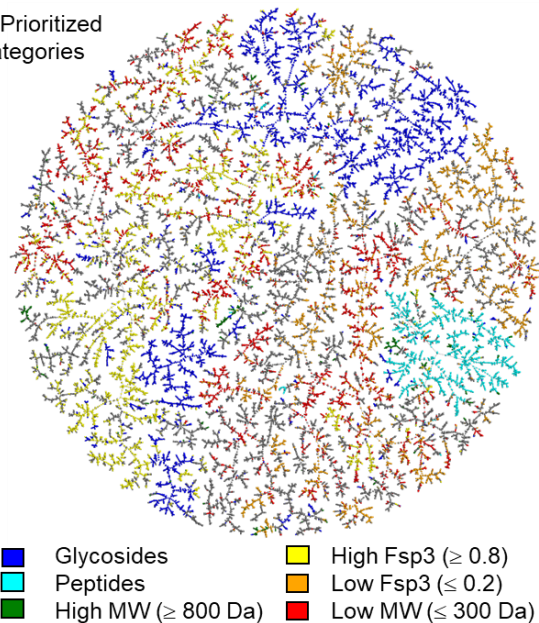


Figure 1. MAP4 TMAP of the 60 thousand selected COCONUT entries. The maps are colored according to (a) molecular weight MW in Da, (b) fraction of sp³ carbon atoms Fsp³, (c) calculated octanol:water partition coefficient AlogP, (d) COCONUT annotated origin, (e) presence of a glycoside (blue) or peptide (green) substructure, or both (magenta), (f) prioritized categories: glycosides (blue) > peptides (cyan) > high MW (green) > high Fsp³ (yellow) > low Fsp³ (orange) > low MW (red). Entries not belonging to any category are reported in gray. All maps are accessible in an interactive format at https://tm.gdb.tools/map4/coconut_tmap/.

Statistical modeling of NP origin using support vector machines (SVM)

The clear separation of NPs from plants, fungi, or bacteria in the TMAP above clearly showed that our MAP4 fingerprint distinguished between NPs of plant, bacterial or fungal origin. To further investigate this separation, we trained an SVM classifier using the MAP4 similarity matrix of half of the COCONUT subset and used the other half to evaluate it. Indeed, the obtained MAP4 SVM correctly predicted the origin of 94% of plant NPs, 89% of fungal NPs, and 89% of the bacterial NPs (MAP4 SVM), resulting in a balanced accuracy of 0.897, an MCC (Matthews correlation coefficient) of 0.890, and an F1 score of 0.920 (for a detailed explanation of the used metrics, please refer to the method section 4.5).

To better identify the role of the MAP4 molecular encoding in the reported successful prediction, we compared the performances of a MAP4 SVM with the performances of an SVM trained using ECFP4 (Extended Connectivity Fingerprint ECFP of radius 2, ECFP4 SVM) and the RDKit atom pair fingerprint (AP SVM). We chose ECFP4 and the RDKit AP as widely used and available examples of substructures fingerprints and atom pair fingerprints. As a baseline model, we also included an SVM trained with a set of 11 calculated physico-chemical properties, namely MW, Fsp³, HBD (hydrogen bond donor) count, HBA (hydrogen bond acceptor) count, AlogP, the number of carbons, oxygens, and nitrogens, the total number of atoms, number of bonds, and TPSA (topological polar surface area) (properties SVM). The selected 60 thousand COCONUT entries were divided into five subsets, and each model was trained and evaluated five times using the five different 80-20

training test splits combinations of one subset as test set and the other four as training set. Then the mean balanced accuracy, MCC, and F1 score of the five evaluations were calculated.

The results of this evaluation are presented in Table 2 and Figure 2. Remarkably, all four SVM performed reasonably well. The good performance of the property based SVM reflected the fact that relatively large NP families with characteristic properties are essentially all from the same origin. For example, almost all large peptides or cyclic peptides are assigned to bacteria, while most glycosylated triterpenoids and polyphenols are assigned to plants. Nevertheless, there was a significant performance increase with the ECFP4 SVM and MAP4 SVM, which performed best, showing that correct origin assignment works better if specific substructures are considered. Among the four SVM evaluated, our MAP4 SVM performed best with significantly higher values compared to the ECFP4 SVM, probably because the MAP4 fingerprint encodes a more precise representation of the molecular structures than ECFP4. Indeed, MAP4 considers pairs of local substructures and the topological distance between them, while ECFP4 only encodes the presence of local substructures.

Table 2. SVM evaluation with balanced accuracy, MCC, and F1 score.

	Balanced acc.	MCC	F1
MAP4 SVM ^{a,b)}	0.919 \pm 0.005	0.879 \pm 0.005	0.929 \pm 0.003
ECFP4 SVM ^{a,b)}	0.890 \pm 0.005	0.827 \pm 0.006	0.897 \pm 0.003
RDKit AP SVM ^{a,b)}	0.735 \pm 0.005	0.592 \pm 0.006	0.752 \pm 0.004
Properties SVM ^{a,c)}	0.758 \pm 0.005	0.613 \pm 0.007	0.761 \pm 0.004

^{a)} Mean value and standard deviation (σ) of the five different test/training sets split of the 5-fold cross-validation. ^{b)} 1024 dimensions. ^{c)} 11 properties: MW, Fsp3, HBD) and HBA, calculated logP with the Crippen method (AlogP), number of carbons, oxygen, and nitrogen, the total number of atoms, number of bonds, and topological polar surface area (TPSA).

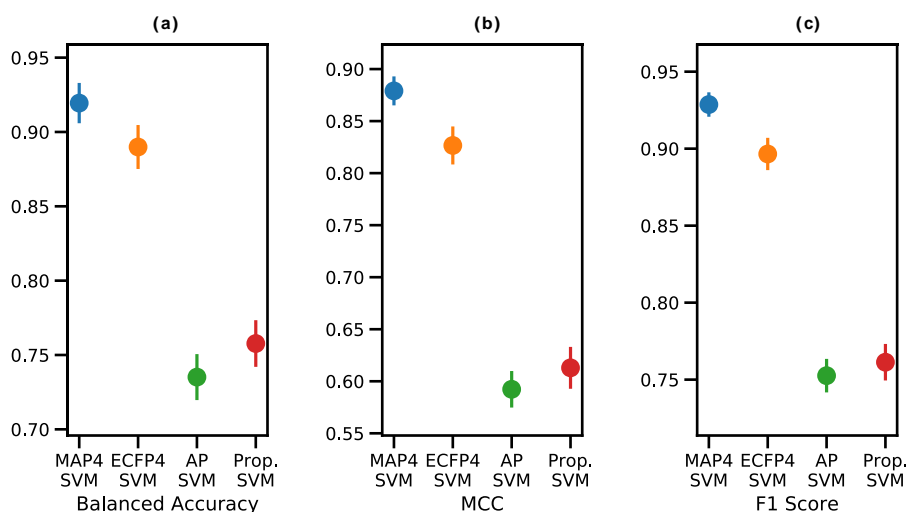


Figure 2. 5-fold cross-validation mean values and 3 σ confidence intervals of the (a) balanced accuracy, (b) MCC, and (c) F1 score for the four SVM classifiers. In all panels, the MAP4 SVM is reported in blue, the ECFP4 SVM in orange, the RDKit AP (AP) SVM in green, and the properties (Prop.) SVM in red.

Using the MAP4 SVM to assign the origin of NPs

The SVM evaluation above showed that the MAP4 analysis of NP molecular structure identified features distinguishing between NPs assigned to plants, fungi, and bacteria.

Assuming that most of the assigned origins were correct among the 60,171 NPs used for training, one may use an SVM to tentatively assign the origin of further NPs as originating from plants, fungi, or bacteria. To best exploit the information in the COCONUT database, we trained a MAP4 SVM using the entire set of 60 thousand COCONUT NPs assigned to plants, fungi, or bacteria. We used the resulting classifier to build an online tool that takes any molecular structure as input (drawn or pasted as SMILES) and returns the assigned origin

and the corresponding percentages from the SVM classifier. This tool is freely accessible online at <https://np-svm-map4.gdb.tools/>.

The online tool performed quite well in assigning the origin of newly published NPs which were not present in COCONUT. Among thirteen recently reported NPs from plants, fungi, or bacteria, eleven were correctly assigned to their origin, while only two were misassigned (Table 3, Figure 3). In details, the fungal epicospirocin 1,[33] penicimeroterpenoid A,[34] and beetleane A,[35] the bacterial bosamycin A,[36] and the plant hunzeylanine A,[37] hyperfol B,[38] pegaharmol A,[39] mucroniferal A,[40] meloyunnanine A,[41] perovsfolin A,[42], and horienoid A[43] were correctly classified. On the other hand, the fungal rhizolutin[44] and the bacterial marinoterpin A[45] were misclassified as plant NP. Note that in these cases, the percentage values to the assigned class were lower than for the correct predictions.

Table 3. MAP4 SVM origin prediction for thirteen recently published microbial and plants NPs that are not present in COCONUT.

Natural Product	Origin	MAP SVM prediction^{a)}
epicospirocin 1	fungal	fungal (97%)
penicimeroterpenoid A	fungal	fungal (82%)
beetleane A	fungal	fungal (97%)
rhizolutin	fungal	plant (55%, fungal: 29%)
bosamycin A	bacterial	bacterial (94%)
marinoterpin A	bacterial	plant (44%, bacterial: 37%)
meloyunnanine A	plant	plant (99%)
hyperfol B	plant	plant (93%)
pegaharmol A	plant	plant (77%)
mucroniferal A	plant	plant (73%)
hunzeylanine A	plant	plant (95%)
perovsfolin A	plant	plant (92%)
horienoid A	plant	plant (95%)

a) Predicted using the MAP4 SVM available online at <https://np-svm-map4.gdb.tools/>.

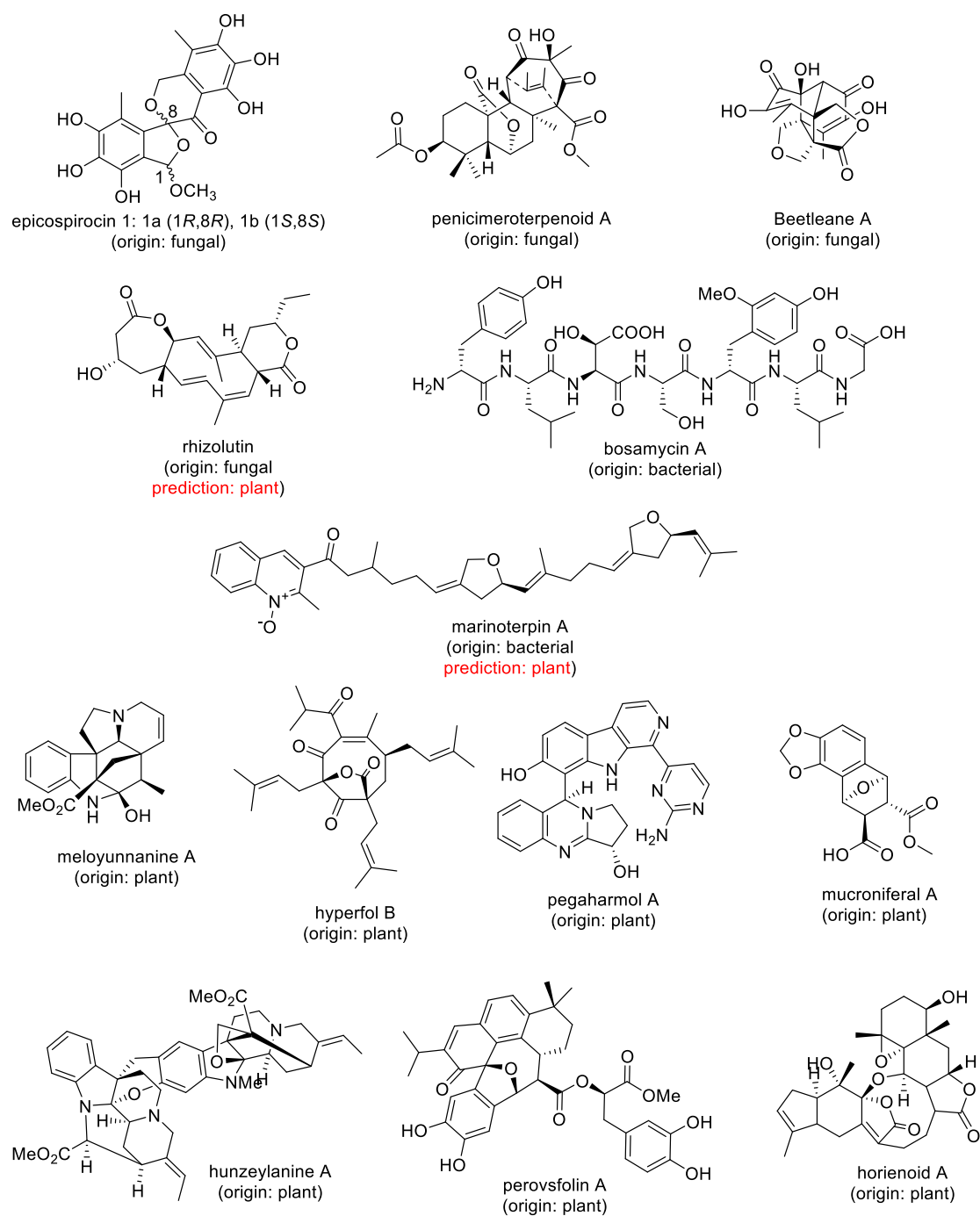


Figure 3. Chemical structure of thirteen recently published microbial and plants NPs which are not present in COCONUT. The MAP4 SVM prediction is identical with the origin unless marked otherwise.

As an additional test of our online tool, we investigated the predicted origin of the 3,364 NPs in COCONUT reported with an origin and a publication for which the organism name was reported (e.g. *Brachystemma calycinum*) but not the corresponding superclass (e.g. plant). Checking individual predictions showed that the predicted origin superclass was in many cases correct, in line with our performance evaluation. For example, the 49 NPs with *Euphorbia* as a source, many of which were peracetylated polycyclic terpene alcohols, as well as the 45 NPs with *Radula* as a source, which were polyphenols and terpenes, were all correctly assigned to a plant origin.

In several cases, the SVM prediction conflicted with the superclass of the reported source organism. For example, the indole alkaloids cephalinones A-D and cephalandoles A-C isolated from the orchid *Cephalantheropsis gracilis*[46] and whose structures were partly revised by total synthesis,[47] were all assigned to bacteria by our SVM. In fact, These NPs might stem from an endophytic bacterium considering that endophytic microorganisms produce several related indole alkaloids.[48] Our SVM also reassigned the cancer drug maytansin from an annotated plant origin in the training set to a predicted bacterial origin, in line with its endophytic origin.[27] On the other hand, our classifier also assigned a bacterial origin to two cyclic peptides (CNP0085258 and CNP0085259)[49] and a cyclotide (CNP0085363)[50] isolated from plants. Although these plants indeed contain endophytic bacteria, the plant origin of such peptides is well established,[51, 52] and the SVM assignment to bacteria reflects the fact that the majority of cyclic peptides and cyclotides in the COCONUT set used for training the SVM were assigned to bacteria, compared to only a handful of cyclotides of plant origin.

While the classifier may point to the possible endophytic origin of NPs isolated from plants, its use on NPs from other sources is problematic. For instance, among the 1,035 marine NPs from COCONUT with an annotated origin, 639 were assigned to plants by our

SVM. This prediction must be mostly wrong considering that most marine organisms such as algae, corals, and sponges are not plants. For example, the 44 NPs from the soft coral *Sinularia*, or the macrocyclic terpene lactone lobophytolide A (CNP0275045) stemming from the soft corral *lobophytum cristagalli*, [53, 54] were all incorrectly assigned to plants. However, the remaining 231 fungal and 165 bacterial predictions might be partly correct considering that many marine organisms carry endosymbionts. For example, our classifier assigned a bacterial origin for echinosulfonic acid B (CNP0318329), a brominated bis-indole NP isolated from the marine sponge *Echinodictyum gorgonoides*. [55] In this case, other authors have reported the isolation of a bacterial strain from the same sponge as a probable source of its biological activities. [56]

Conclusion

In summary, we visualized the chemical space covered by a subset of 60 thousand NPs from the COCONUT database with an assigned origin and publication using a TMAP calculated on the basis of MAP4 as molecular fingerprint, which is available at https://tm.gdb.tools/map4/coconut_tmap/. Analyzing this TMAP revealed that NPs from plant, fungal or bacterial origin form well separated groups. We then trained an SVM classifier with the MAP4 fingerprint to assign the origin of NPs and found that it performed excellently and significantly better than classifiers trained with ECFP4, RDkit AP, or physico-chemical descriptors.

To help assign NP origin, we then trained a MAP4 SVM classifier using the entire set of 60 thousand NPs. This tool is available online at <https://np-svm-map4.gdb.tools/> and returns an origin prediction for any molecular structure drawn or pasted as SMILES. We found that this classifier correctly predicts the origin of plant, bacterial or fungal NPs not included in the 60 thousand COCONUT set used for training, as exemplified with the correct

prediction of eleven out of thirteen newly published NPs. Broader testing of the classifier with further NPs from COCONUT showed limitations for NPs not from plant or microbial origin, such as marine NPs, but it also led to interesting use cases suggesting that the tool might serve as a help to assign NP origin. This concerns in particular NPs isolated from plant but which might in fact be produced by endophytic microorganisms. Such information could be essential when searching for the corresponding biosynthetic genes.

Methods

Database preprocessing

The coconut database was downloaded. Only the 135,091 (out of 400,837) entries having a taxonomical annotation were selected. The selected subset was further filtered down to the 67,730 entries having an annotation not shorter than ten characters in the DOI field. Then, the taxonomy field was split by commas and match towards the words "plant"/"plants", "fungi"/"aspergillus", "bacteria"/"bacillus"/"bacta" to select the NPs with an annotated plant, fungal, or bacterial origin, respectively. The entries common between multiple origins were assigned with the following priority: human > animal > bacteria > fungi > plant > marine. The process led to the selection of 33,772 plant NPs, 15,648 fungi NPs, and 10,751 bacterial NPs with annotated DOI, for a total of 60,171 structures. The number of carbons, oxygen, and nitrogens, the total number of atoms, number of bonds, and TPSA were extracted from the COCONUT annotations. MW, Fsp3, HBD, and HBA count, AlogP, were calculated using RDKit.[57] The presence/absence of a peptide or a glycoside moiety was evaluated using Daylight[58] SMILES arbitrary target specification (SMARTS) language. SMARTS were used with RDKit to identify COCONUT entries containing a dipeptide substructure, defined as "[NX3,NX4+][CH1,CH2][CX3](=[OX1])[NX3,NX4+][CH1,CH2][CX3](=[OX1])[O,N]",

or a containing a glycoside defined as cyclic N- or O-acetal substructure with the SMARTS “[CR][OR][CHR]([OR0,NR0])[CR]”.

Fingerprint calculation

The 1024 dimensions MinHashed atom pair fingerprint of radius 2 was calculated using the open-source code of MAP4.

TMAP layout

The indices generated by the MinHash procedure of the MAP4 calculation were used to create a locality-sensitive hashing (LSH) forest[59] of 32 trees. Then, for each structure, the 20 approximate nearest neighbors (NNs) in the MAP4 feature space were extracted from the LSH forest, and the tree layout was calculated. The LSH forest and the minimum spanning tree layout were calculated using the TMAP open-source code. Finally, Fearun[31] was used to display the obtained layout interactively.

MAP4 SVM implementation

The coconut SUBSET entries used to generate the TMAP were assigned to training or test set with a 50% random split. The SVM was trained using the MAP4 fingerprints of the training set. It utilized a custom kernel that calculates the similarity matrix between two MAP4 fingerprints, where the similarity of fingerprint *a* and fingerprint *b* is calculated (1) counting of elements with the same value and the same index across *a* and *b*, and (2) dividing the obtained value by the number of elements of fingerprint *a*. The class weights were inversely proportional to the class frequency, and the hyperparameter C was optimized using 5-fold cross-validation. During the hyperparameter optimization, 20% of the training set was left out as a validation set, and the balanced accuracy of the validation set was maximized. The hyperparameter C was optimized among the values 0.1, 1, 10, 100, and 1000, resulting in C = 1. The classifier was implemented using scikit-learn[60] with the "one versus rest" strategy,

and all not mentioned hyperparameters were used in their default values. Platt scaling,[61] was used to obtain probabilistic prediction values. After the evaluation process, a second version of the MAP4 SVM classifier was trained using both training and test to learn from all curated 60 thousand data points.

MAP4, ECFP4, RDKit AP, and properties SVMs comparison

The MAP4, ECFP4, and the RDKit AP fingerprints and a set of 11 properties (MW, Fsp3, HBD and HBA count, AlogP, number of carbons, oxygens, and nitrogens, total number of atoms, number of bonds, and TPSA) were used to train four different SVM classifiers in a 5-fold cross-validation. For all classifiers, the class weights were inversely proportional to the class frequency, and the hyperparameters were optimized using 10% of the available data (Table 4). For the properties SVM, the 11 values were scaled to zero mean and unit variance.

Table 4. Non-default and optimized hyperparameters used in the 5-fold cross-validation MAP4, ECFP4, RDKit AP, and properties SVMs comparison.

Classifier	Kernel ^{a)}	C ^{a)}	γ ^{a)}
MAP4 SVM	MAP4 ^{b)}	0.01, 0.1, 1 , 10, 100	-
ECFP4 SVM	Tanimoto ^{c)} , Dice ^{c)}	0.01, 0.1, 1 , 10, 100	-
RDKit AP SVM	Tanimoto ^{c)} , Dice ^{c)}	0.01, 0.1, 1, 10 , 100	-
Properties SVM	RBF ^{d)}	0.01, 0.1, 1 , 10, 100	0.01, 0.1, 1 , 10, 100

a) Used hyperparameters are reported in bold. b) Calculates the similarity matrix between two MAP4 fingerprints, where the similarity of fingerprint *a* and fingerprint *b* is calculated (1) counting of elements with the same value and the same index across *a* and *b*, and (2) dividing the obtained value by the number of elements of fingerprint *a*. c) Ralaivola *et al.* [62] d) Radial basis function (RBF) kernel.[63]

Classifiers evaluation metrics

The F1 score is defined as the harmonic mean of precision and recall:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\ score = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$$

Where TP stands for true positives, TN for true negatives, FP for false positives, and FN for false negatives predicted by the classifier.

The balanced accuracy is defined as:

$$\text{Balanced accuracy} = \frac{\frac{TP}{TP + FP} + \frac{TN}{TN + FN}}{2}$$

The Matthews correlation coefficient (MCC) is a correlation between the observed and the predicted class and it is defined as:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Online MAP4 SVM

The MA4 SVM classifier trained with the whole 60 thousand COCONUT subset is found at <https://np-svm-map4.gdb.tools/>. The query molecule can be provided as a drawn structure or pasted SMILES in the JSME editor[64]. The given query is canonicalized, chirality information is removed with RDKit, and the MAP4 fingerprint is calculated. To obtain probabilistic prediction values for each class, we use Platt scaling.[61]

List of abbreviations

AP – Atom pair

COCONUT – Collection of Open Natural Products

HBA – Hydrogen bond acceptor)

HBD – Hydrogen bond donor

LSH – Locality sensitive hashing

MAP4 – MinHashed atom pair fingerprint

MCC – Matthews correlation coefficient

ML – machine learning

MW – Molecular weight

NN – Nearest neighbor

NP – Natural product

SMARTS – SMILES arbitrary target specification

SMILES – Simplified molecular-input line-entry system

SVM – Supported vector machine

TMAP – Tree-map

TPSA – Topological polar surface area

Data Availability. The code used for the presented work is publicly available at

<https://github.com/reymond-group/Coconut-TMAP-SVM>.

Conflict of Interest. The authors declare no conflict of interest.

Funding. This work was supported by a grant from the Vice-Rectorate Development of the University of Bern to A. C., by the Swiss National Science Foundation Grant no. 200020_178998, and by the European Research Council Grant no. 885076.

Author contributions. AC designed and realized the project and wrote the paper. JLR designed and supervised the project and wrote the paper.

Acknowledgments. The authors thank Prof. Dr. Olivier Potterat, University of Basel, for critical reading and helpful discussions.

References

1. Dias DA, Urban S, Roessner U (2012) A Historical Overview of Natural Products in Drug Discovery. *Metabolites* 2:303–336. <https://doi.org/10.3390/metabo2020303>
2. Newman DJ, Cragg GM (2020) Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J Nat Prod* 83:770–803. <https://doi.org/10.1021/acs.jnatprod.9b01285>
3. Chen Y, de Bruyn Kops C, Kirchmair J (2017) Data Resources for the Computer-Guided Discovery of Bioactive Natural Products. *J Chem Inf Model* 57:2099–2111. <https://doi.org/10.1021/acs.jcim.7b00341>
4. Chen Y, Kirchmair J (2020) Cheminformatics in Natural Product-Based Drug Discovery. *Mol Inform.* <https://doi.org/10.1002/minf.202000171>
5. Ertl P, Roggo S, Schuffenhauer A (2008) Natural Product-likeness Score and Its Application for Prioritization of Compound Libraries. *J Chem Inf Model* 48:68–74. <https://doi.org/10.1021/ci700286x>
6. Zaid H, Raiyn J, Nasser A, et al (2010) Physicochemical Properties of Natural Based Products versus Synthetic Chemicals. *Open Nutraceuticals J* 3:
7. Yu MJ (2011) Natural Product-Like Virtual Libraries: Recursive Atom-Based Enumeration. *J Chem Inf Model* 51:541–557. <https://doi.org/10.1021/ci1002087>
8. Vanii Jayaseelan K, Moreno P, Truszkowski A, et al (2012) Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinformatics* 13:106. <https://doi.org/10.1186/1471-2105-13-106>
9. Chen Y, Stork C, Hirte S, Kirchmair J (2019) NP-Scout: Machine Learning Approach for the Quantification and Visualization of the Natural Product-Likeness of Small Molecules. *Biomolecules* 9:43. <https://doi.org/10.3390/biom9020043>
10. Pereira F (2021) Machine Learning Methods to Predict the Terrestrial and Marine Origin of Natural Products. *Mol Inform n/a*: <https://doi.org/10.1002/minf.202060034>
11. Djoumbou Feunang Y, Eisner R, Knox C, et al (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminformatics* 8:61. <https://doi.org/10.1186/s13321-016-0174-y>
12. Kim H, Wang M, Leber C, et al (2020) NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. <https://doi.org/10.26434/chemrxiv.12885494.v1>
13. Zabolotna Y, Ertl P, Horvath D, et al (2021) NP Navigator: a New Look at the Natural Product Chemical Space. <https://doi.org/10.26434/chemrxiv.14236457.v1>

14. Capecchi A, Reymond J-L (2020) Assigning the Origin of Microbial Natural Products by Chemical Space Map and Machine Learning. *Biomolecules* 10:1385. <https://doi.org/10.3390/biom10101385>
15. van Santen JA, Jacob G, Singh AL, et al (2019) The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent Sci* 5:1824–1833. <https://doi.org/10.1021/acscentsci.9b00806>
16. Capecchi A, Probst D, Reymond J-L (2020) One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J Cheminformatics* 12:43. <https://doi.org/10.1186/s13321-020-00445-4>
17. Probst D, Reymond J-L (2020) Visualization of very large high-dimensional data sets as minimum spanning trees. *J Cheminformatics* 12:12. <https://doi.org/10.1186/s13321-020-0416-x>
18. Wu Y, Liao H, Liu L-Y, et al (2020) Phakefustatins A–C: Kynurenine-Bearing Cycloheptapeptides as RXR α Modulators from the Marine Sponge *Phakellia fusca*. *Org Lett*. <https://doi.org/10.1021/acs.orglett.0c01586>
19. Han M, Liu F, Zhang F, et al (2012) Bacterial and archaeal symbionts in the South China Sea sponge *Phakellia fusca*: community structure, relative abundance, and ammonia-oxidizing populations. *Mar Biotechnol N Y N* 14:701–713. <https://doi.org/10.1007/s10126-012-9436-5>
20. Meunier L, Tocquin P, Cornet L, et al (2020) Palantir: a springboard for the analysis of secondary metabolite gene clusters in large-scale genome mining projects. *Bioinformatics* 36:4345–4347. <https://doi.org/10.1093/bioinformatics/btaa517>
21. Villa TG, Viñas M (2019) *Horizontal Gene Transfer: Breaking Borders Between Living Kingdoms*. Springer International Publishing, Cham
22. Hardoim PR, van Overbeek LS, Berg G, et al (2015) The Hidden World within Plants: Ecological and Evolutionary Considerations for Defining Functioning of Microbial Endophytes. *Microbiol Mol Biol Rev MMBR* 79:293–320. <https://doi.org/10.1128/MMBR.00050-14>
23. Strobel G, Daisy B, Castillo U, Harper J (2004) Natural Products from Endophytic Microorganisms. *J Nat Prod* 67:257–268. <https://doi.org/10.1021/np030397v>
24. Ye K, Ai H-L, Liu J-K (2021) Identification and Bioactivities of Secondary Metabolites Derived from Endophytic Fungi Isolated from Ethnomedicinal Plants of Tujia in Hubei Province: A Review. *Nat Prod Bioprospecting* 11:185–205. <https://doi.org/10.1007/s13659-020-00295-5>
25. Howat S, Park B, Oh IS, et al (2014) Paclitaxel: biosynthesis, production and future prospects. *New Biotechnol* 31:242–245. <https://doi.org/10.1016/j.nbt.2014.02.010>
26. Shankar Naik B (2019) Developments in taxol production through endophytic fungal biotechnology: a review. *Orient Pharm Exp Med* 19:1–13. <https://doi.org/10.1007/s13596-018-0352-8>
27. Kusari S, Lamshöft M, Kusari P, et al (2014) Endophytes Are Hidden Producers of Maytansine in *Putterlickia* Roots. *J Nat Prod* 77:2577–2584. <https://doi.org/10.1021/np500219a>

28. Sorokina M, Merseburger P, Rajan K, et al (2021) COCONUT online: Collection of Open Natural Products database. *J Cheminformatics* 13:2. <https://doi.org/10.1186/s13321-020-00478-9>
29. Heim WG, Sykes KA, Hildreth SB, et al (2007) Cloning and characterization of a *Nicotiana tabacum* methylputrescine oxidase transcript. *Phytochemistry* 68:454–463. <https://doi.org/10.1016/j.phytochem.2006.11.003>
30. Hooven HW van den, Lagerwerf FM, Heerma W, et al (1996) The structure of the lantibiotic lactacin 481 produced by *Lactococcus lactis*: location of the thioether bridges. *FEBS Lett* 391:317–322. [https://doi.org/10.1016/0014-5793\(96\)00771-5](https://doi.org/10.1016/0014-5793(96)00771-5)
31. Probst D, Reymond J-L, Wren J (2018) FUn: a framework for interactive visualizations of large, high-dimensional datasets on the web. *Bioinformatics* 34:1433–1435. <https://doi.org/10.1093/bioinformatics/btx760>
32. Probst D, Reymond J-L (2018) SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript. *J Chem Inf Model* 58:1–7. <https://doi.org/10.1021/acs.jcim.7b00425>
33. Zhu G, Hou C, Yuan W, et al (2020) Molecular networking assisted discovery and biosynthesis elucidation of the antimicrobial spiroketals epicospirocins. *Chem Commun*. <https://doi.org/10.1039/D0CC03990J>
34. Cheng X, Liang X, Zheng Z-H, et al (2020) Penicimeroterpenoids A–C, Meroterpenoids with Rearrangement Skeletons from the Marine-Derived Fungus *Penicillium* sp. SCSIO 41512. *Org Lett*. <https://doi.org/10.1021/acs.orglett.0c02160>
35. Cao P-R, Zheng Y-L, Zhao Y-Q, et al (2021) Beetleane A and Epicoane A: Two Carbon Skeletons Produced by *Epicoccum nigrum*. *Org Lett*. <https://doi.org/10.1021/acs.orglett.1c00731>
36. Xu ZF, Bo ST, Wang MJ, et al (2020) Discovery and biosynthesis of bosamycin from *Streptomyces* sp. 120454. *Chem Sci*. <https://doi.org/10.1039/D0SC03469J>
37. Zhang J, Yuan M-F, Li S-T, et al (2020) Hunzeylanines A–E, Five Bisindole Alkaloids Tethered with a Methylene Group from the Roots of *Hunteria zeylanica*. *J Org Chem* 85:10884–10890. <https://doi.org/10.1021/acs.joc.0c01448>
38. Lou H, Yi P, Hu Z, et al (2020) Polycyclic polyprenylated acylphloroglucinols with acetylcholinesterase inhibitory activities from *Hypericum perforatum*. *Fitoterapia* 143:104550. <https://doi.org/10.1016/j.fitote.2020.104550>
39. Li S-G, Wang Y-T, Zhang Q, et al (2020) Pegaharmols A–B, Axially Chiral β -Carboline-quinazoline Dimers from the Roots of *Peganum harmala*. *Org Lett* 22:7522–7525. <https://doi.org/10.1021/acs.orglett.0c02709>
40. Zhang J, Shi L-Y, Yin X, et al (2020) Discovery of novel potential plant growth regulators from *Corydalis mucronifera*. *Fitoterapia* 147:104776. <https://doi.org/10.1016/j.fitote.2020.104776>
41. Wu J, Zhao S-M, Shi B-B, et al (2020) Cage-Monoterpenoid Quinoline Alkaloids with Neurite Growth Promoting Effects from the Fruits of *Melodinus yunnanensis*. *Org Lett* 22:7676–7680. <https://doi.org/10.1021/acs.orglett.0c02871>

42. Tanaka N, Niwa K, Kajihara S, et al (2020) C28 Terpenoids from Lamiaceous Plant *Perovskia scrophulariifolia*: Their Structures and Anti-neuroinflammatory Activity. *Org Lett* 22:7667–7670. <https://doi.org/10.1021/acs.orglett.0c02855>
43. Fan Y-Y, Gan L-S, Chen S-X, et al (2021) Horienoids A and B, Two Heterocoupled Sesquiterpenoid Dimers from *Hedyosmum orientale*. *J Org Chem*. <https://doi.org/10.1021/acs.joc.1c00307>
44. Kwon Y, Shin J, Nam K, et al (2020) Rhizolutin, a novel 7/10/6-tricyclic dilactone, dissociates misfolded protein aggregates and reduces apoptosis/inflammation associated with Alzheimer's disease. *Angew Chem Int Ed*. <https://doi.org/10.1002/anie.202009294>
45. Kim MC, Winter JM, Asolkar RN, et al (2021) Marinoterpins A–C: Rare Linear Meroesterterpenoids from Marine-Derived Actinomycete Bacteria of the Family Streptomycetaceae. *J Org Chem*. <https://doi.org/10.1021/acs.joc.1c00262>
46. Wu P-L, Hsu Y-L, Jao C-W (2006) Indole Alkaloids from *Cephalanceropsis gracilis*. *J Nat Prod* 69:1467–1470. <https://doi.org/10.1021/np060395l>
47. Mason JJ, Bergman J, Janosik T (2008) Synthetic Studies of Cephalandole Alkaloids and the Revised Structure of Cephalandole A. *J Nat Prod* 71:1447–1450. <https://doi.org/10.1021/np800334j>
48. Ishikura M, Yamada K (2009) Simple indole alkaloids and those with a nonrearranged monoterpenoid unit. *Nat Prod Rep* 26:803–852. <https://doi.org/10.1039/B820693G>
49. Zhao J, Zhou L-L, Li X, et al (2011) Bioactive Compounds from the Aerial Parts of *Brachytemma calycinum* and Structural Revision of an Octacyclopeptide. *J Nat Prod* 74:1392–1400. <https://doi.org/10.1021/np200048u>
50. Yeshak MY, Burman R, Asres K, Göransson U (2011) Cyclotides from an Extreme Habitat: Characterization of Cyclic Peptides from *Viola abyssinica* of the Ethiopian Highlands. *J Nat Prod* 74:727–731. <https://doi.org/10.1021/np100790f>
51. Srivastava S, Dashora K, Ameta KL, et al (2021) Cysteine-rich antimicrobial peptides from plants: The future of antimicrobial therapy. *Phytother Res* 35:256–277. <https://doi.org/10.1002/ptr.6823>
52. Santos-Silva CA dos, Zupin L, Oliveira-Lima M, et al (2020) Plant Antimicrobial Peptides: State of the Art, In Silico Prediction and Perspectives in the Omics Era. *Bioinform Biol Insights* 14:1177932220952739. <https://doi.org/10.1177/1177932220952739>
53. Tursch B, Braekman JC, Daloze D, et al (1974) Chemical studies of marine invertebrates. X. Lobophytolide, a new cembranolide diterpene from the soft coral *lobophytum cristagalli* (coelenterata, octocorallia, alcyonacea). *Tetrahedron Lett* 15:3769–3772. [https://doi.org/10.1016/S0040-4039\(01\)92004-0](https://doi.org/10.1016/S0040-4039(01)92004-0)
54. Blunt JW, Copp BR, Munro MHG, et al (2010) Marine natural products. *Nat Prod Rep* 27:165–237. <https://doi.org/10.1039/B906091J>
55. Ovenden SPB, Capon RJ (1999) Echiniosulfonic Acids A–C and Echiniosulfone A: Novel Bromoindole Sulfonic Acids and a Sulfone from a Southern Australian Marine Sponge, *Echinodictyum*. *J Nat Prod* 62:1246–1249. <https://doi.org/10.1021/np9901027>

56. Dhinakaran DI, Prasad DRD, Gohila R, Lipton P (2012) Screening of marine sponge-associated bacteria from *Echinodictyum gorgonoides* and its bioactivity. *Afr J Biotechnol* 11:15469–15476. <https://doi.org/10.4314/ajb.v11i88>
57. (2020) RDKit. <https://www.rdkit.org/>
58. Daylight. <https://www.daylight.com/>. Accessed 17 Jul 2020
59. Bawa M, Condie T, Ganesan P (2005) LSH forest: self-tuning indexes for similarity search. In: *Proceedings of the 14th international conference on World Wide Web*. Association for Computing Machinery, Chiba, Japan, pp 651–660
60. Pedregosa F, Varoquaux G, Gramfort A, et al (2011) Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 12:2825–2830
61. Platt JC (1999) Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: *Advances in Large Margin Classifiers*. MIT Press, pp 61–74
62. Ralaivola L, Swamidass SJ, Saigo H, Baldi P (2005) Graph kernels for chemical informatics. *Neural Netw* 18:1093–1110. <https://doi.org/10.1016/j.neunet.2005.07.009>
63. Vert JP, Tsuda K, Schölkopf B (2004) A Primer on Kernel Methods. In: *Kernel Methods in Computational Biology*. Biologische Kybernetik, Cambridge, MA, USA, pp 35–70
64. Bienfait B, Ertl P (2013) JSME: a free molecule editor in JavaScript. *J Cheminformatics* 5:24. <https://doi.org/10.1186/1758-2946-5-24>