GigaScience

Open and re-usable annotated mass spectrometry dataset of a chemodiverse 1,600 plant extracts collection.

--Manuscript Draft--

Manuscript Number:	GIGA-D-22-00126
Full Title:	Open and re-usable annotated mass spectrometry dataset of a chemodiverse 1,600 plant extracts collection.
Article Type:	Data Note
Funding Information:	Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung (CRSII5_189921/1) — Pr. Jean-Luc Wolfender
Abstract:	As privileged structures, natural products often display potent biological activities. However, the discovery of novel bioactive scaffolds is often hampered by the chemical complexity of the biological matrices they are found in. Large natural extracts collections are thus extremely valuable for their chemical novelty potential but also complicated to exploit in the frame of drug-discovery projects. In the end, it is the pure chemical structures that are desired for structural determination and evaluation of biological potential. Researchers interested in the exploitation of large and chemodiverse extracts collections should thus elaborate strategies aiming to efficiently tackle such chemical complexity and access these structures. The establishment of carefully crafted digital layers documenting the spectral and chemical complexity as well as bioactivity results of natural products extracts collections can help to efficiently prioritize time-consuming but mandatory isolation efforts. In this note, we report the results of our initial exploration of a collection of 1,600 plant extracts in the frame of a drug discovery effort. After describing the taxonomic coverage of this collection, we present the results of its high-resolution mass spectrometric profiling and the exploitation of these profiles using computational solutions. The resulting annotated mass spectral dataset and associated chemical and taxonomic metadata are made available to the community. We are currently continuing our exploration of this extract collection for drug-discovery purposes (notably looking for novel anti-trypanosomatids, anti-infective and prometabolic compounds) and eco-metabolomics insights. We hope that such a dataset can be exploited and reused by researchers interested in computational natural products exploration.
Corresponding Author:	Pierre-Marie Allard University of Fribourg Fribourg, Freiburg SWITZERLAND
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Fribourg
Corresponding Author's Secondary Institution:	
First Author:	Pierre-Marie Allard
First Author Secondary Information:	
Order of Authors:	Pierre-Marie Allard
	Arnaud Gaudry
	Luis-Manuel Quirós-Guerrero
	Adriano Rutz
	Miwa Dounoue-Kubo
	Emmanuel Defossez
	Christophe Long

	Antonio Grondin
	Bruno David
	Jean-Luc Wolfender
Order of Authors Secondary Information:	
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	No
If not, please give reasons for any omissions below. as follow-up to "**Experimental design and statistics** Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript? "	Not applicable. Not statistical analysis in this datanote.
Resources A description of all resources used, including antibodies, cell lines, animals	Yes

and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

Open and re-usable annotated mass spectrometry dataset of a chemodiverse 1,600 plant extracts collection.

Pierre-Marie Allard[1,2,3], pierre-marie.allard@unifr.ch, 0000-0003-3389-2191
Arnaud Gaudry[1,2], arnaud.gaudry@unige.ch, 0000-0002-3648-7362
Luis-Manuel Quirós-Guerrero[1,2], luis.guerrero@unige.ch, 0000-0002-1630-8697
Adriano Rutz[1,2], adriano.rutz@unige.ch, 0000-0003-0443-9902
Miwa Dounoue-Kubo[4], miwa-k@ph.bunri-u.ac.jp, 0000-0002-7271-3128
Emmanuel Defossez[3,5], emmanuel.defossez@unifr.ch, 0000-0002-3279-9190
Christophe Long[6], christophe.long@pierre-fabre.com, 0000-0002-7677-7700
Antonio Grondin[7], antonio.grondin@pierre-fabre.com, 0000-0002-2952-6271
Bruno David[7], brunoxdavid@gmail.com, 0000-0002-6222-9228
Jean-Luc Wolfender[1,2], jean-luc.wolfender@unige.ch, 0000-0002-0125-952X

[1] Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, 1211 Geneva 4, Switzerland
[2] School of Pharmaceutical Sciences, University of Geneva, 1211 Geneva 4, Switzerland
[3] Department of Biology, University of Fribourg, 1700 Fribourg, Switzerland
[4] Faculty of Pharmaceutical Sciences, Tokushima Bunri University, 180 Yamashiro-cho, Tokushima, 770-8514, Japan.
[5] Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland
[6] Direction Scientifique Naturactive, Pierre Fabre Medicament, 17 Avenue Jean Moulin, 81100 Castres, France.
[7] Green Mission Pierre Fabre, Institut de Recherche Pierre Fabre, 3 Avenue Hubert Curien, BP 13562, 31562 Toulouse, France.

Abstract

As privileged structures, natural products often display potent biological activities. However, the discovery of novel bioactive scaffolds is often hampered by the chemical complexity of the biological matrices they are found in. Large natural extracts collections are thus extremely valuable for their chemical novelty potential but also complicated to exploit in the frame of drug-discovery projects. In the end, it is the pure chemical structures that are desired for structural determination and evaluation of biological potential. Researchers interested in the exploitation of large and chemodiverse extracts collections should thus elaborate strategies aiming to efficiently tackle such chemical complexity and access these structures. The establishment of carefully crafted digital layers documenting the spectral and chemical complexity as well as bioactivity results of natural products extracts collections can help to efficiently prioritize time-consuming but mandatory isolation efforts. In this note, we report the results of our initial exploration of a collection of 1,600 plant extracts in the frame of a drug discovery effort. After describing the taxonomic coverage of this collection, we present the results of its high-resolution mass spectrometric profiling and the exploitation of these profiles using computational solutions. The resulting annotated mass spectral dataset and associated chemical and taxonomic metadata are made available to the community. We are currently continuing our exploration of this extract collection for drug-discovery purposes (notably looking for novel anti-trypanosomatids, anti-infective and prometabolic compounds) and eco-metabolomics insights. We hope that such a dataset can be exploited and reused by researchers interested in computational natural products exploration.

Keywords

Plant extracts collection, metabolomics, drug discovery, LC-MS, natural products, mass spectrometry, chemodiversity, open-science, biodiversity digitization

Data Description

In the frame of a partnership between academia and industry, a collaboration was established in 2017 between the Laboratory of Phytochemistry and Bioactive Natural Products at the University of Geneva (LPBNP), Switzerland, and the Pierre Fabre Laboratories (PFL) in Toulouse, France. This collaboration had the objective to explore and exploit the chemodiversity of a large collection of plant extracts (furnished by the industrial partner) using state-of-the-art mass spectrometric profiling methods and associated data mining solutions (performed and developed by the academic partner). The main goal of this research project is to establish a precise and exploitable description of the wide chemical diversity displayed in the plant extracts collection in order to orient further isolation efforts aiming to discover, in a first move, novel anti-trypanosomatids scaffolds and deepen chemotaxonomic knowledge of such a large set.

Context

Pierre Fabre Laboratories were created in 1962 by Pierre Fabre, a French pharmacist from Castres (Tarn, France). They have specialized since the very beginning in the exploration and valorization of plants as medicines, health, and beauty products. Innovation based on the vegetal world in PFL encompasses chemotherapy treatments with vinca alkaloids extracted from the leaves of *Catharanthus roseus* [1], dermo-cosmetic products such as hair dyes from *Lawsonia inermis* extracts [2], or Celastrol-enriched extracts for psoriatic skin obtained from in vitro plant cell culture of *Tripterygium wilfordii* [3]. In 1998, Pierre Fabre decided to launch a High Throughput Screening (HTS) program based on plant extracts with the objective to find novel anti-cancer drugs. The first samples of plant parts were collected in 1998 and when the HTS program ended in December 2015, the collection contained just over 17,000 unique samples. Multiple scientific articles originating from the exploitation of this unique collection have been published over those years [4–7]. Hereafter is a selection of natural products-based scaffolds that entered antitumor medicinal chemistry programs: dimeric derivatives of artemisinin [8]; flavagline derivatives [9]; griseofulvin [10]; narciclasine and pancratistatin [11]; neoboutomellerone [12]; triptolide [13].

The PFL collection, which is among the largest collection plant samples worldwide with over 17,000 unique samples, was registered on April 1, 2020 at the European Commission under the accession number 03-FR-2020. This official registration recognizes the legality of the access and management process. To date three european collections are recognized

(https://ec.europa.eu/environment/nature/biodiversity/international/abs/pdf/Register%20of%20Collections.pdf).

In 2015 the Nature Open Library program was launched in order to share the unique PFL expertise with industrial or academic partners and to foster the research, development and industrialization of plant assets. In this context, PFL provided access to their private plants' collection, one of the most important in the industry, numbering over 17,000 classified samples, including some rare species. (https://www.pierre-fabre.com/en/press_release/pierre-fabre-laboratories-unveils-open-nature-library-a-unique-open-innovation). Upon restructuration, this program was discontinued as of 2018. Some academic collaborations established at this time were kept active.

The partnership established between LPBNP and PFL in 2017 had for principal objective the chemical characterization of the full PFL plant extracts collection. In order to evaluate the feasibility of such an ambitious program, a pilot project was defined. It focussed on a selection of 1,600 samples (corresponding to approximately 10% of the extracts prepared in the full PFL collection). The results of this pilot study, namely the generated data and outcomes are shared in this data note.

Note : from now on, in this manuscript, the full collection (17,000 extracts) is referred to as the "*PFL collection*". The selected set is simply referred to as "*the 1,600 extracts collection*".

Sample collection

All samples from the PFL collection were collected with respect to the different regulations around genetic resources at the time of collection. After drying and grounding, they were stored at room temperature, in the dark, in high-density polyethylene (HDPE) bar-coded 0.5L or 2L-pots. The storage room is access secured and protected by an automatic fire protection system with inert gas (50% nitrogen, 50% argon). Precise localization of the geographical sampling sites, unique IDs, barcodes and quantities, are stored in the PFL internal data management system. Furthermore, to perform later sample inspections if wrong identification of the original plants was suspected, all dry plant parts were sampled and stored in their intact form and not in the ground form.

Several constraints were followed while building this collection. First, to have the best chances of finding new chemical entities, a large chemical diversity was desirable. Taking the assumption that taxonomical position and chemical production were deeply related, samples were collected in order to maximize diversity within superior plant taxa (classes, orders, family and genera, see Figure 1). Second, only plants providing reasonable biomass quantities (i.e. > 50 g dry ground samples) were collected, hence allowing isolation of potentially active compounds in sufficient quantities to be characterized and to perform preliminary bioassays. Third, for preserving samples on the long term but also for allowing only stable compounds to be left in the samples, they were dried for three days at 55°C. This point foresaw issues of potential HTS hits due to unstable compounds which would be difficult to isolate and manipulate.

The phylogenetic coverage of the 1,600 extracts collection is depicted in Figure 1. Within the Streptophyta phylum the collection represented ca. 64 % of the known orders and ca. 30 % of all known botanical families. The numbers rapidly decreased to ca. 4 % of the known genus and less than 0.5 % of all known species. See horizontal bar plot on Figure 1. The botanical families of the collection are relatively well distributed across the global phylogeny, despite some orders (e.g. Sapindales, Rosales) being better represented than others (e.g. Boraginales, Cyatheales). The scripts for taxonomical names resolving and figure generation are available at https://github.com/mandelbrot-project/pf_1600_datanote/blob/main/src/phylogenetic_coverage.R

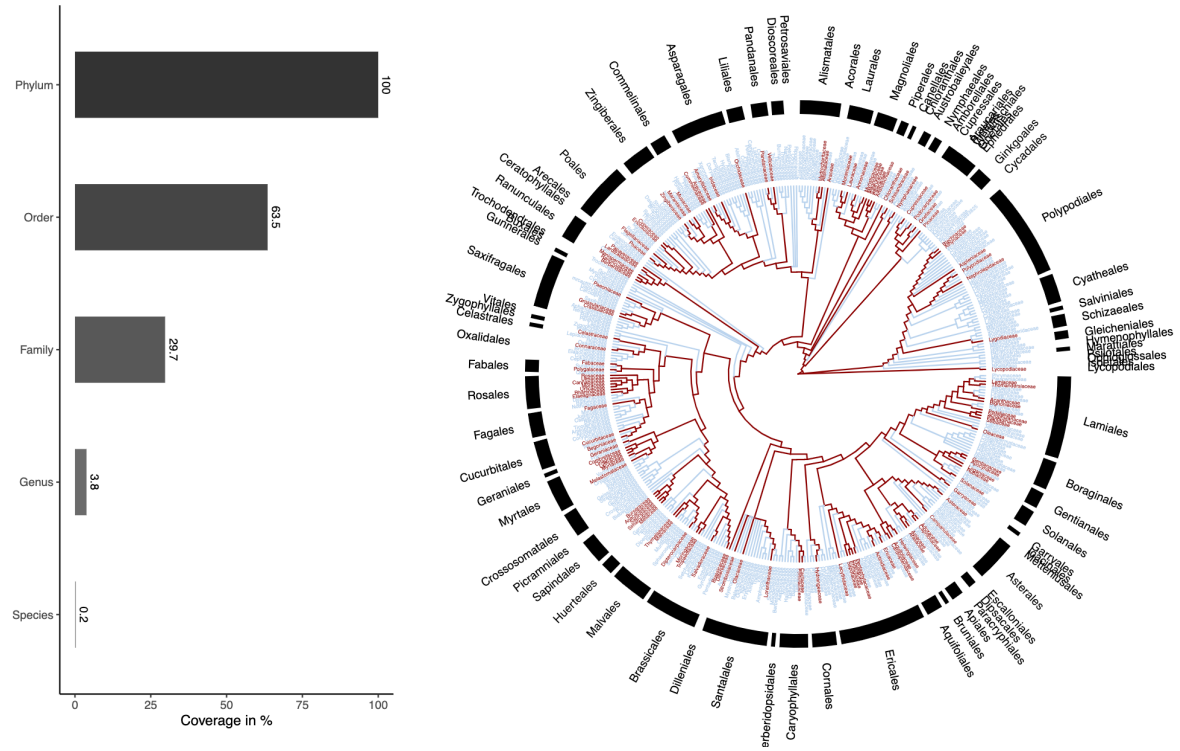

Figure 1: Taxonomical coverage of the profiled collection (1,600). On the left, the bar plot represents the overall coverage at main taxa level and up to the phylum Streptophyta. On the right, the taxonomical coverage is represented using a taxonomical tree of all families within the Streptophyta phylum. The families present within the current collection are highlighted in red. The Open Tree of Life (ott3.3) was used for taxonomy resolving.

Extraction and sample preparation

Since the PFL collection was designed to provide extracts for HTS experiments, only medium polarity compounds were targeted. Using ethyl acetate (EtOAc) extracts purified over silica gel allowed to reduce the amount of known classes of pan-assay interference compounds (PAINS) such as condensed tannins [14].

Starting from the dried plant material, the extracts were prepared as follows. A sample of 8 g (leaves, whole plant or aerial parts) or 10 g (subterranean parts, roots, bark) was extracted over a period of 8-10 h with 80 mL or 100 mL of EtOAc respectively. The extracts were filtered through paper and dried under vacuum in a rotary evaporator to a reduced volume (ca. 5 mL). The residue was transferred to a pre-weighted recipient and dried under vacuum until constant weight. The extracts were dissolved in EtOAc at a concentration of 30 mg/mL. An aliquot of 1 mL was mixed with 200 mg of silica gel (Merk 60, 40-63 μm) and dried under vacuum. The resulting powder mixture was transferred to a SPE cartridge (6 mL, 1 g SiO2). The cartridge was eluted with 10 mL

of dichloromethane to remove non polar compounds, then the cartridge was eluted with 8 mL of a mixture dichloromethane : methanol (85:15), and subsequently washed with 2 mL of the same mixture to fully recover the compounds of interest. The filtrate was recovered (ca 10 mL final volume) and dried under vacuum. Extracts were dissolved in DMSO to a concentration of 5 mg/mL and transferred to 1 mL 96 deep-well plates for further analysis.

The extracts were dissolved in DMSO to reach a concentration of 2.5 mg/mL. The plates originally contained 80 extracts per plate, the columns 1 and 12 being empty. These columns were filled, respectively, by a QC (quality control) sample and a DMSO blank. The QC samples was constituted by a mixture of *Cinchona pubescens* Vahl, *Panax ginseng* C.A.Mey., *Ginkgo biloba* L., *Arnica montana* L. and *Salvia officinalis* L. methanolic extracts dissolved in a 1:1:1:1:1 ratio at 5 mg/mL in DMSO.

LC-MS/MS analyses

Chromatographic separation was performed on a Waters Acquity UPLC system interfaced to a Q-Exactive Focus mass spectrometer (Thermo Scientific, Bremen, Germany), using a heated electrospray ionization (HESI-II) source. Thermo Scientific Xcalibur 3.1 software was used for instrument control. The LC conditions were as follows: column, Waters BEH C18 50 × 2.1 mm, 1.7 μm; mobile phase, (A) water with 0.1% formic acid; (B) acetonitrile with 0.1% formic acid; flow rate, 600 $\mu L \cdot min^{-1}$; injection volume, 2 μL; gradient, linear gradient of 5−100% B over 7 min and isocratic at 100% B for 1 min. The optimized HESI-II parameters were as follows: source voltage, 3.5 kV (pos); sheath gas flow rate (N2), 55 units; auxiliary gas flow rate, 15 units; spare gas flow rate, 3.0; capillary temperature, 350.00°C, S-Lens RF Level, 45. The mass analyzer was calibrated using a mixture of caffeine, methionine–arginine–phenylalanine–alanine–acetate (MRFA), sodium dodecyl sulfate, sodium taurocholate, and Ultramark 1621 in an acetonitrile/methanol/water solution containing 1% formic acid by direct injection. The data-dependent MS/MS events were performed on the three most intense ions detected in full scan MS (Top3 experiment). The MS/MS isolation window width was 1 Da, and the stepped normalized collision energy (NCE) was set to 15, 30 and 45 units. In data-dependent MS/MS experiments, full scans were acquired at a resolution of 35,000 FWHM (at *m/z* 200) and MS/MS scans at 17,500 FWHM both with an automatically determined maximum injection time. After being acquired in a MS/MS scan, parent ions were placed in a dynamic exclusion list for 2.0 s.

Data-processing and Molecular Networking

The MS data were converted from .RAW (Thermo) standard data format to .mzXML format using the MSConvert software, part of the ProteoWizard package [15]. The converted files were treated using the MZMine software suite v. 2.53 [16]. The parameters were adjusted as follows: the centroid mass detector was used for mass detection with the noise level set to 1.0E4 for MS level set to 1, and to 0 for MS level set to 2. The ADAP chromatogram builder was used and set to a minimum group size of scans of 5, minimum group intensity threshold of 1.0E4, minimum highest intensity of 5.0E5 and *m/z* tolerance of 12 ppm [17]. For chromatogram deconvolution, the algorithm used was the wavelets (ADAP). The intensity window S/N was used as S/N estimator with a signal to noise ratio set at 10, a minimum feature height at 5.0E5, a coefficient area threshold at 130, a peak duration ranges from 0.0 to 0.5 min and the RT wavelet range from 0.01 to 0.03 min. Isotopes were detected using the isotopes peaks grouper with a *m/z* tolerance of 12 ppm, a RT tolerance of 0.01 min (absolute), the maximum charge set at 2 and the representative isotope used was the most intense. Each feature list was filtered before alignment to keep only features with an associated MS2 scan and a RT between 0.5 and 8.0 min using the feature filtering. Peak alignment was performed using the join aligner method (*m/z* tolerance at 40 ppm), absolute RT tolerance 0.2 min, weight for *m/z* at 2 and weight for RT at 1 and a weighted dot-product cosine similarity of 0.3. The aligned feature list (119,182 features) was exported using the export to GNPS module. Features occurring in QC samples only or in blanks were removed before molecular networking, resulting in a final feature list of 117,005 features. The MZMine parameters used for the data treatment are available at https://massive.ucsd.edu/ProteoSAFe/DownloadResultFile?file=f.MSV000087728/updates/2022-05-02_pmallard_e88304cd/other/210302_VGF_pos_parameters.xml

The full spectral dataset corresponding to the 1,600 plant extracts collection was uploaded on the MassIVE repository https://doi.org/doi:10.25345/C59J97. Via the GNPS Explorer interface it is possible to efficiently navigate through the uploaded spectral file and their associated metadata. See this link https://gnps-explorer.ucsd.edu/MSV000087728?dataset_accession=MSV000087728&metadata_source=MASSIVE&metadata_option=f.MSV000087728%2Fupdates%2F2021-11-15_pmallard_334e9199%2Fmetadata%2Fgnps_metadata.tsv. For example, the name of a plant of interest can be typed under the NCBI Taxonomy header resulting in the direct filtering of the dataset for this specific plant. Individual spectral files can then be selected (and eventually compared if multiple are selected) and viewed using the GNPS dashboard. See for example this interactive view of the total ion chromatogram of *Desmodium heterophyllum* Hook. & Arn aerial parts.

The initial study of this aligned peak list with classical unsupervised statistics (e.g. PCA, PCoA) indicated a strong batch effect which could be tracked down to a specific date and attributed to a change of column during the course of the mass spectrometry analysis of the full collection. Such batch effects are almost inevitable when samples are profiled on long periods (here spanning over several months) and are particularly complicated to mitigate especially when dealing with chemodiverse datasets where poor overlap among the samples is expected. This incentivized us to develop a novel computational mass spectrometry solution for MS2 BasEd SaMple VectOrization (MEMO). MEMO allows to organize samples in large and chemodiverse collections in a retention-time agonist fashion thus strongly mitigating batch effect and allowing the comparison of samples acquired over heterogeneous chromatographic conditions. Applied to the current plant extract collection, MEMO allows efficient reduction of the observable batch-effect and clustering samples according to their content [18]. See the following link for a view of the batch effect on the classically aligned dataset and its mitigation using MEMO fingerprints (https://mandelbrot-project.github.io/memo_publication_examples/plant_extract_dataset/pcoa_vgf_color_before_after.html)

To analyze the spectral diversity of the profile collection, a molecular network (MN) was created using the online workflow (https://ccms-ucsd.github.io/GNPSDocumentation/) on the GNPS website (http://gnps.ucsd.edu) using the .mgf spectra file generated at the previous step. [19] The precursor ion mass tolerance was set to 0.02 Da and a MS/MS fragment ion tolerance of 0.02 Da. A network was then created where edges were filtered to have a cosine score above 0.7 and more than 6 matched peaks. Further, edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. Finally, the maximum size of a spectral family was set to 100, and the lowest scoring edges were removed from molecular families until the molecular family size was below this threshold. The spectra in the network were then searched against GNPS spectral libraries. All matches kept between network spectra and library spectra were required to have a score above 0.7 and at least 6 matched peaks. The resulting molecular network is available at https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=3197f70bed224f9ba6f59f62906839e9.

In addition to the MN, spectra were organized using TMAP visualization [20]. Briefly, a minimum spanning tree is built based on the distance between spectra. This distance was computed based on the presence or absence of MS/MS peaks and losses to the precursor in the different spectra. The generated TMAP is presented in Figure 2 and

an interactive visualization is available at https://mandelbrot-project.github.io/pf_1600_datanote/220408_VGF_samples_spectral_tmap_pos.html. Such visualization allows to highlight taxa-specific spectral subspaces. In addition, structural annotation (see next section for details concerning the metabolite annotation process) were overlaid to their corresponding spectra. Here structural annotations were limited to compounds reported in the Eukaryote domain. For example the spectra region 1 in Figure 2 B, C, and D is mainly specific to samples belonging to the Meliaceae family and most of these spectra are annotated as limonoids derivatives. Another region, called region 2 in Figure 2, contains mainly spectra specific to the Annonaceae family and corresponds to acetogenins derivatives. Finally, the spectral region 3 is specific to the Apocynaceae family and is mainly annotated as tryptophan alkaloids derivatives. On the other hand, botanical families such as the Fabaceae appear to occupy a much wider spectral space. These examples showcase the importance of the taxonomic coverage of an extract collection to maximize the chemical diversity. Indeed, if some chemical classes, such as the flavonoid derivatives, appear to be widespread, some others are known to be extremely specific to some taxa, such as the acetogenins to the Annonaceae [21]. Of course, the interpretation of this spectral TMAP is highly dependent on the structural annotations' quality (see next section "Metabolite annotation") and as such should be taken with caution. Because the annotation process favored structures reported in closer taxa, it should also be noted that the comparison of annotation occurrences and taxonomy may be impacted by an annotation set biased towards denser taxa-specific ensembles of structures.

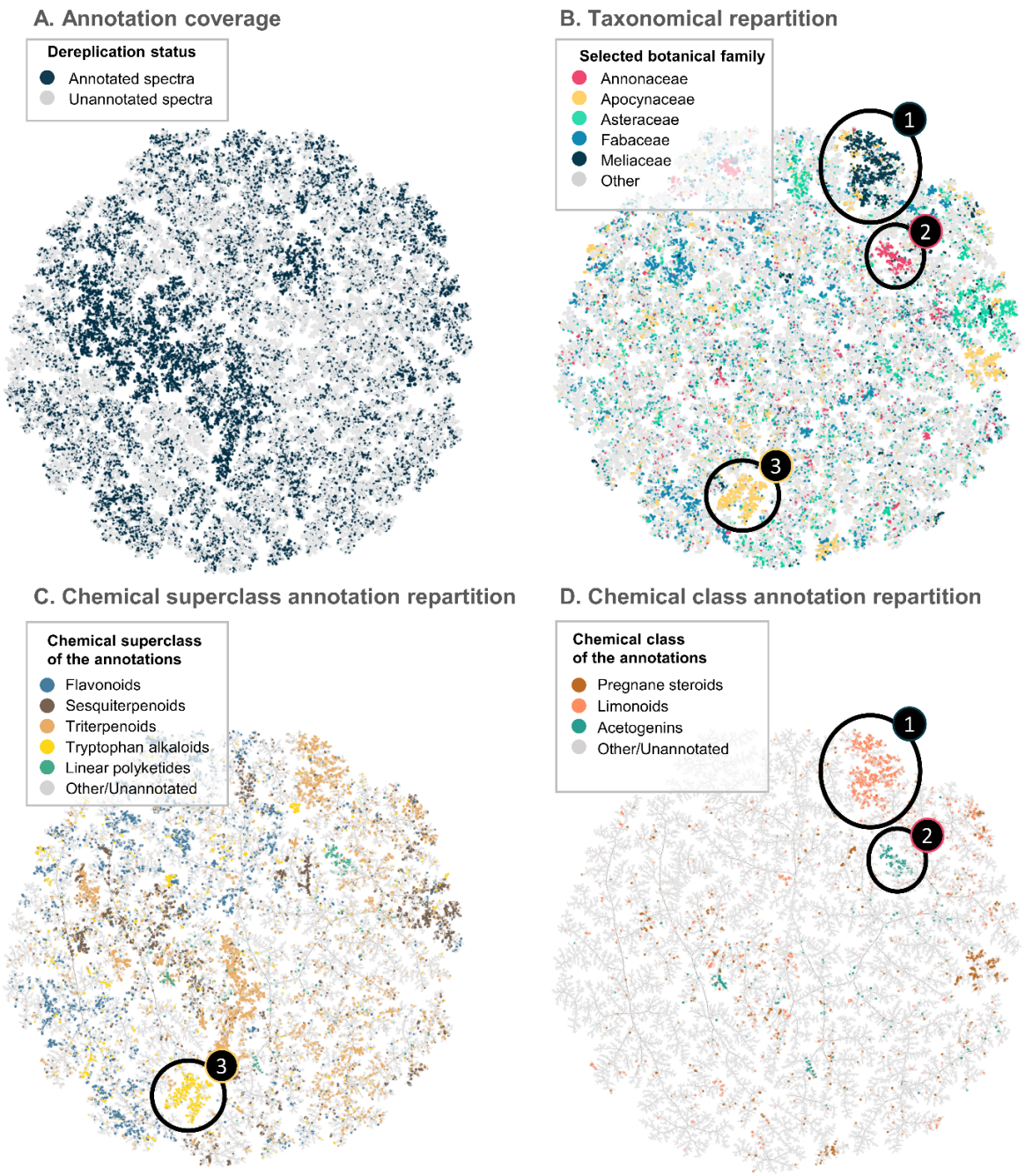
A. Annotation coverage
Dereplication status
Annotated spectra
Unannotated spectra
B. Taxonomical repartition
Selected botanical family
Annonaceae
Apocynaceae
Asteraceae
Fabaceae
Meliaceae
Other
1
2
3
C. Chemical superclass annotation repartition
Chemical superclass of the annotations
Flavonoids
Sesquiterpenoids
Triterpenoids
Tryptophan alkaloids
Linear polyketides
Other/Unannotated
3
D. Chemical class annotation repartition
Chemical class of the annotations
Pregnane steroids
Limonoids
Acetogenins
Other/Unannotated
1
2

Figure 2: Spectral diversity of the profiled plant collection (1,600 extracts). The TMAP approach is employed to display the >100,000 of spectra resulting from the alignment of the 1,600 untargeted MS/MS experiments. In this TMAP each dot represents a feature's spectrum and they are linked together according to their similarity. In **A**, blue dots (36% of the total amount of spectra) correspond to annotated spectra while gray dots (64%) correspond to unannotated spectra. In **B**, dots are colored according to the botanical family of the sample where the highest MS1 peak area for the corresponding feature was recorded. In **C** and **D**, dots are colored according to the NP Classifier superclass and class, respectively, of their annotation (for annotated dots). In **B**, it is possible to spot spectral regions (1, 2 and 3) specific to given botanical families. The region 1 is specific to the Meliaceae family and these spectra are mainly annotated as limonoids derivatives, the region 2 is specific to the Annonaceae family and spectra are mainly annotated as acetogenins derivatives and the region 3 is specific to the Apocynaceae family and spectra are annotated as tryptophan alkaloids derivatives. Note that the structural annotation results are reweighed according to the taxonomical proximity of the biological source of the candidate structure and the biological source of the annotated spectra. A bias toward taxa-specific structures can thus be observed. This interactive structural TMAP can be browsed at https://mandelbrot-project.github.io/pf_1600_datanote/220408_VGF_samples_spectral_tmap_pos.html

Metabolite annotation

Experimental spectral libraries search

The full spectral dataset corresponding to the 1,600 plant extracts collection being uploaded on the MASSIVE repository https://doi.org/doi:10.25345/C59J97, a continuous identification workflow is automatically carried against GNPS experimental spectral libraries. The latest results of this continuous identification workflow can be observed at https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=b753bf1e39cb4875bdf3b786e747bc15&view=advanced_view. The latest iteration at the time of publication (2022-04-18) indicated that a total of 2665 unique compounds were spectrally matched (see https://gnps.ucsd.edu/ProteoSAFe/result.jsp?task=ee2e8e9cf8214f48ab6ee01df652a3f2&view=all_unique_compounds)

Theoretical spectral libraries search

In addition to experimental spectral libraries search we have shown that spectral matching against theoretical spectral libraries of natural products was an efficient way to cover a much wider spectral space [22]. Furthermore we showed that taking into account the taxonomical distance between the biological source of candidate structure and the biological source of the annotated extracts greatly improved the overall quality of the annotation results [23]. Thus, in addition to the spectral search performed at the molecular networking step against publicly available spectral libraries (see previous section) a taxonomically informed metabolite annotation was performed. For this we

first established a large theoretical spectral database of natural products following a previously established metabolite annotation workflow [24]. This spectral database and associated biological sources metadata were constructed using chemical structure and information compiled during the LOTUS Initiative's first project aiming to establish an open and evolutive ressource compiling natural products biological occurrences [25]. The spectral database is publicly available at https://doi.org/10.5281/zenodo.5607264. The biological sources metadata are available at the following address https://doi.org/10.5281/zenodo.6378204. The taxonomically informed metabolite annotation was performed using the following scripts (https://github.com/mandelbrot-project/met_annot_enhancer/releases/tag/v0.1). The parameters used for the taxonomically informed metabolite annotation process and the resulting tables can be found at https://massive.ucsd.edu/ProteoSAFe/dataset_files.jsp?task=b753bf1e39cb4875bdf3b786e747bc15#%7B%22table_sort_history%22%3A%22main.collection_asc%22%2C%22main.attachment_input%22%3A%22updates%2F2022-04-28_pmallard_b0e0f70c%22%7D

Visualization of the metabolite annotations

In order to obtain an overview of the metabolite annotation results on the profiled collection, these were compared to the ensemble of molecules described in LOTUS, See Figure 3. For this, a TMAP [20,26,27] was built to connect similar chemical structures using the MAP4 fingerprint [28]. To resume, in Figure 3, each dot represents a chemical structure and it is linked to its neighbor according to its structural proximity. In Figure 3 A, the color code indicates the repartition of chemical structures mostly found to be present in plants (green color). In Figure 1 B. each of the annotated structures within the 1,600 plants extracts collection is displayed (strong blue) within the rest of the reported structures in LOTUS (light gray) indicating a relatively heterogeneous structural coverage of the annotations with some denser areas. Finally, a qualitative color coding in Figure 1 C. allows to distinguish chemical classes in the overall TMAP. The NPClassifier classification is employed [29]. The framed bar plot displays the most frequently annotated chemical classes in the present dataset (strong color) versus the total count for each class in LOTUS (light color). Multiple factors can explain discrepancies between the repartition of chemical classes annotated in the 1,600 extracts collection versus the overall repetition observed in LOTUS. For example cyclic peptides (green bar in the framed bar plot) are poorly covered. This can be explained by the fact that such structures are most often found in microbial organisms (as displayed when comparing plot A. and B.) On the other hand the high coverage of limonoids in the collection can be explained by the high proportion of species from the Sapindales order (see Figure 1.) as this order is known to be the main responsible of the biosynthesis of such scaffolds [30]. An interactive

version of the structural TMAP presented in Figure 3 is available at https://mandelbrot-project.github.io/pf_1600_datanote/220408_annotation_vs_lotusdnp_tmap.html
A coverage of annotations of the dataset versus the full LOTUS annotations is available as interactive plots (coverage at the pathway level, at the superclass level and at the class level.)

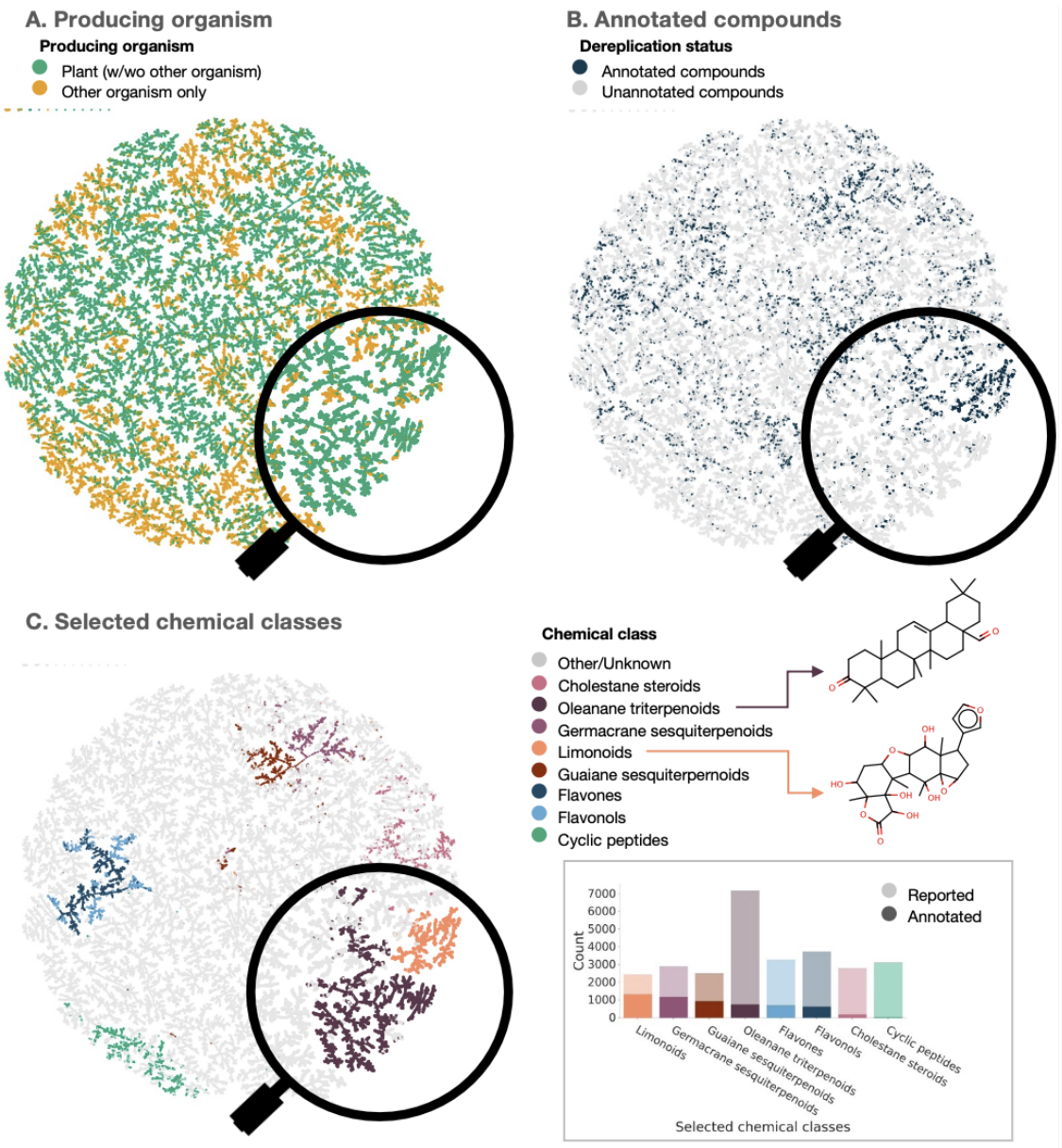

Figure 3: Chemical diversity of the profiled plant collection (1,600 extracts) and coverage against reported natural products. Visualization of reported natural products (LOTUS v1 and Dictionary of Natural Products v29.1) as a TMAP with plotting of the producing organism (A), the annotation's status, ie whether the 2D structure was annotated in the dataset (B) and selected NPClassifier classes (C). In the insert of C, the bar plot represents the number of compounds reported for each of the selected chemical classes with the opaque part of the bar representing the annotated compounds. The zoom on the limonoids and oleanane terpenoids clusters of the TMAP allows visualizing a well covered chemical class such as limonoids and a less covered one such as oleanane triterpenoids. A specific member of each class (toonaciliatin F, HRTHMOOARAGYNK, limonoid and oleanonic aldehyde, QGPIUZIWMRUUCS, oleanane triterpenoid) is represented in their planar structure form for illustration purposes. This interactive structural TMAP can be browsed at https://mandelbrot-project.github.io/pf_1600_datanote/220408_annotation_vs_lotusdnp_tmap.html

Conclusions

Here we present the results of state-of-the-art computational approaches used for both spectral organization and metabolite annotation of high-resolution mass spectrometry data acquired on a large collection of 1,600 chemically diverse plant extracts. All original mass spectrometry profiles and associated metadata are made available to the community for further analysis. The results of the metabolite annotation represent a putative chemical space that can be exploited and refined in subsequent drug discovery, chemotaxonomic or eco-metabolomics projects. Sharing these results with the community shows that, through effective partnership between academia and industry, the faith of historical and private collections of plant extracts can be changed and the richness of the associated chemical diversity can be made available to a wider public.

Availability of source code and requirements (if used in the paper)

Repository for the scripts used to generates the figure of this datanote: https://github.com/mandelbrot-project/pf_1600_datanote

License: GNU Affero General Public License v3.0

Data Availability

All .RAW (Thermo), .mzXML and .mzML datafiles in positive and negative ionization modes along with metadata and metabolite annotation tables are available on the MassIVE repository under accession number MSV000087728, available here doi:10.25345/C59J97.

Molecular networking job results are available at: https://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=3197f70bed224f9ba6f59f62906839e9.

Declarations

Competing interests

The authors declare that they have no competing interests

Funding

J-LW and P-MA are thankful to the Swiss National Science Foundation for the funding of the project (SNF N° CRSII5_189921/1).

Authors' contributions

Following the CRediT (Contributor Roles Taxonomy) roles.

Conceptualization P-MA, BD and J-LW. Data curation P-MA, AGa, LQG, AR and MDK. Funding acquisition P-MA and J-LW. Investigation P-MA and MDK. Methodology P-MA, MDK and J-LW. Project administration P-MA, BD and J-LW. Resources CL, AGr, BD and J-LW. Software P-MA, AGa, ED and AR. Supervision P-MA and J-LW. Validation P-MA, AGa and MDK. Visualization P-MA, AGa, ED and AR. Writing – original draft P-MA. Writing – review & editing P-MA, AGa, LQG, AR and AGr.

Acknowledgements

The authors are grateful to Green Mission Pierre Fabre, Pierre Fabre Research Institute, Toulouse, France for establishing and sharing this unique library of extracts.

References

1. Duflos A, Kruczynski A, Barret J-M. Novel aspects of natural and modified vinca alkaloids. *Curr Med Chem Anticancer Agents*. 2:55–702002;

2. Fabre B, Fiorini-Puybaret C. Composition cosmetique comprenant un hydrolysat de gomme de caroube. Patent.

3. Nguyen T, Cousy A, Steward N. Method for producing celastrol and pentacyclic triterpene derivatives. US Patent.

4. Vandenberghe I, Créancier L, Vispé S, Annereau J-P, Barret J-M, Pouny I, et al.. Physalin B, a novel inhibitor of the ubiquitin-proteasome pathway, triggers NOXA-associated apoptosis. *Biochem Pharmacol*. 76:453–622008;

5. Pouny I, Long C, Batut M, Aussagues Y, Jean Valère N, Achoundong G, et al.. Quinolizidine Alkaloids from Cylicomorpha solmsii. *J Nat Prod*. 84:1198–2022021;

6. Long C, Sauleau P, David B, Lavaud C, Cassabois V, Ausseil F, et al.. Bioactive flavonoids of Tanacetum parthenium revisited. *Phytochemistry*. 64:567–92003;

7. Pouny I, Batut M, Vendier L, David B, Yi S, Sautel F, et al.. Cytisine-like alkaloids from Ormosia hosiei Hemsl. & E.H. Wilson. *Phytochemistry*. 107:97–1012014;

8. Njuguna NM, Ongarora DSB, Chibale K. Artemisinin derivatives: a patent review (2006 - present). *Expert Opin Ther Pat*. Informa Healthcare; 22:1179–2032012;

9. Marion F, Kaloun EB, Lieby-Muller F, Perez M, Annereau J-P, Creancier L. Flavagline derivatives. World Patent.

10. Marion F, Lieby-Muller F, Grisoni S, Rahier N, Perez M, Sartori I. Griseofulvin derivatives. World Patent.

11. Marion F, Annereau J-P, Fahy J. Nitrogenated derivatives of pancratistatin. World Patent.

12. Beck J, Guminski Y, Long C, Marcourt L, Derguini F, Plisson F, et al.. Semisynthetic neoboutomellerone derivatives as ubiquitin-proteasome pathway inhibitors. *Bioorg Med Chem*. 20:819–312012;

13. Kaloun EB, Long C, Molinier N, Brel V, Cantagrel F, Massiot G. Partial synthesis of 14-deoxy-14-aminotriptolide. *Tetrahedron Lett*. 57:1895–82016;

14. David B, Ausseil F. High-Throughput Screening of Plant Chemodiversity. Encyclopedia of Analytical Chemistry.

15. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al.. A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol*. 30:918–202012;

16. Pluskal T, Castillo S, Villar-Briones A, Orešič M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics*. 11:3952010;

17. Myers OD, Sumner SJ, Li S, Barnes S, Du X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Anal Chem*. 89:8696–7032017;

18. Gaudry A, Huber F, Nothias L-F, Cretton S, Kaiser M, Wolfender J-L, et al.. MEMO: Mass Spectrometry-Based Sample Vectorization to Explore Chemodiverse Datasets. *Front Bioinform*. Frontiers Media SA; 2022; doi: 10.3389/fbinf.2022.842964.

19. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al.. Sharing and community curation of mass spectrometry data with GNPS. *Nat Biotechnol*. 34:828–372016;

20. Probst D, Reymond J-L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J Cheminform*. 12:122020;

21. Neske A, Ruiz Hidalgo J, Cabedo N, Cortes D. Acetogenins from Annonaceae family. Their potential biological applications. *Phytochemistry*. 174:1123322020;

22. Allard P-M, Péresse T, Bisson J, Gindro K, Marcourt L, Pham VC, et al.. Integration of Molecular Networking and In-Silico MS/MS Fragmentation for Natural Products Dereplication. *Anal Chem*. 88:3317–232016;

23. Rutz A, Dounoue-Kubo M, Ollivier S, Bisson J, Bagheri M, Saesong T, et al.. Taxonomically Informed Scoring Enhances Confidence in Natural Products Annotation. *Front Plant Sci*. 10:13292019;

24. Allard PM, Péresse T, Bisson J, Gindro K, Marcourt L, Pham VC, et al.. Integration of molecular networking & in-silico MS/MS fragmentation: a novel dereplication strategy in natural products chemistry. Planta Medica.

25. Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, et al.. The LOTUS initiative for open natural products research: knowledge management through Wikidata. Manubot; 2021.

26. Probst D, Reymond J-L. SmilesDrawer: Parsing and Drawing SMILES-Encoded Molecular Structures Using Client-Side JavaScript. *J Chem Inf Model*. 58:1–72018;

27. Probst D, Reymond J-L. FUn: a framework for interactive visualizations of large,

high-dimensional datasets on the web. *Bioinformatics*. 34:1433–52018;

28. Capecchi A, Probst D, Reymond J-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J Cheminform*. 12:432020;

29. Kim HW, Wang M, Leber CA, Nothias L-F, Reher R, Kang KB, et al.. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J Nat Prod*. 84:2795–8072021;

30. Hodgson H, De La Peña R, Stephenson MJ, Thimmappa R, Vincent JL, Sattely ES, et al.. Identification of key enzymes responsible for protolimonoid biosynthesis in plants: Opening the door to azadirachtin production. *Proc Natl Acad Sci U S A*. 116:17096–1042019;