Data fusion discovery (DAFdiscovery) pipeline to aid compound annotation and bioactive compound discovery across diverse spectral data

Ricardo Moreira Borges[1*], Fernanda das Neves Costa[1], Fernanda O. Chagas[1], Andrew Magno Teixeira[1], Jaewon Yoon[2], Márcio Barczyszyn Weiss[2], Camila Manoel Crnkovic[2], Alan Cesar Pilon[3], Bruno C. Garrido[4], Luz Adriana Betancur[5], Abel M. Forero[6,7], Leonardo Castellanos[6], Freddy A. Ramos[6], Mônica T. Pupo[3], Stefan Kuhn[8]

*Corresponding authors: ricardo_mborges@ufrj.br

[1] Instituto de Pesquisas de Produtos Naturais Walter Mors, Universidade Federal do Rio de Janeiro, Brazil.

[2] Faculdade de Ciências Farmacêuticas, Universidade de São Paulo, Brazil.

[3] Faculdade de Ciências Farmacêuticas de Ribeirão Preto, Universidade de São Paulo, Brazil.

[4] Chemical Metrology Division, Organic Analysis Laboratory, Inmetro, Brazil.

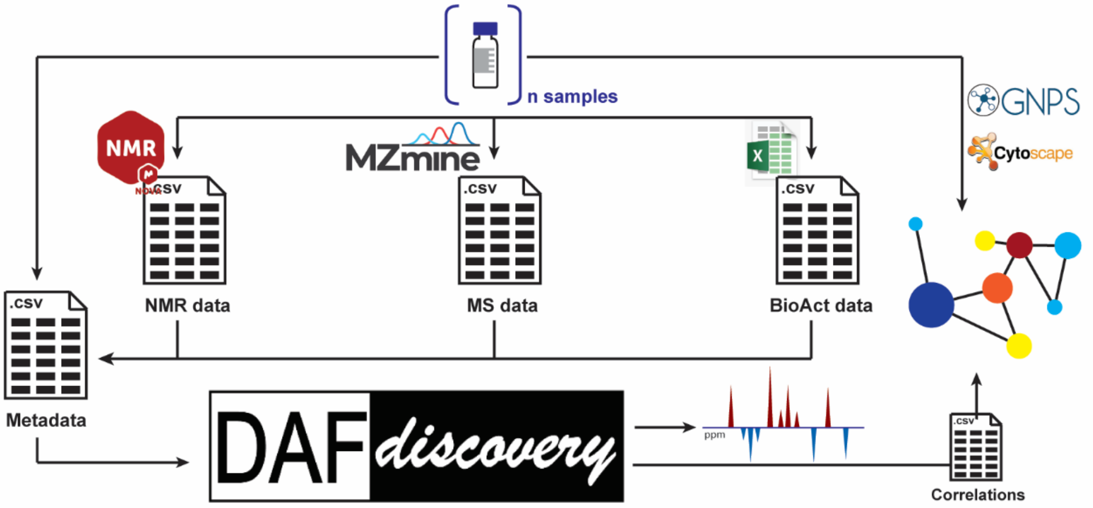
[5] Universidad de Caldas, Departamento de Química, Edificio Orlando Sierra, Bloque B, Sede Palogrande Calle 65 No. 26-10. Manizales, Caldas, Colombia.

[6] Universidad Nacional de Colombia, Sede Bogotá, Departamento de Química, Carrera 30 No. 45-03, Edificio de Química of 427, Bogotá, 111321, Colombia.

[7] Departamento de Química, Facultad de Ciencias and Centro de Investigacions Científicas Avanzadas (CI-CA) Universidade de A Coruña, 15071 A Coruña, Spain.

[8] School of Computer Science and Informatics, De Montfort University, United Kingdom.

Graphical Abstract



Highlights

1. DAFdiscovery is an easy-to-use platform designed to aid drug discovery based on natural products using a data fusion approach;
2. DAFdiscovery facilitates the use of Statistical Total Correlation (STOCSY) and Statistical HeteroSpectroscopy (SHY) methods to combine data from different datasets;
3. DAFdiscovery was proven to accelerate the process for determining compounds (or chemical features) highly correlated to a bioactivity readout;
4. Users are strongly encouraged to apply this method throughout their bioassay-guided fractionation studies.

Abstract

DAFdiscovery is a pipeline designed to help users combine NMR, MS and bioactivity data in a notebook-based application to accelerate annotation and discovery of bioactive compounds. It applies Statistical Total Correlation (STOCSY) and Statistical HeteroSpectroscopy (SHY) calculation in their data using an easy-to-follow Jupyter Notebook. Different case studies are presented for benchmarking, and the resultant outputs are shown to aid natural products identification and discovery. The goal is to encourage users to acquire MS and NMR data from their samples (in replicated samples and fractions when available) and to explore their variance to highlight MS features, NMR peaks, and bioactivity that might be correlated to accelerate bioactive compound discovery or for annotation-identification studies.

INTRODUCTION

The traditional approach for bioactive compound discovery in natural products research comprises a series of incremental fractionations to obtain isolated and purified compounds for bioactivity assay. The well-known "bioassay-guided fractionation fractionation" [1, 2] dominated that field and led to the discovery of avermectin,[3] camptothecin, taxol, [4] and artemisin.[5] Nowadays, the evolving 'omics' technologies are driving many studies in natural products to apply metabolomics approaches and methods with the goal of identifying bioactive compounds in early stages, and to avoid those labor-intensive fractionation steps that may reach replicated results.[2] Thus, the analysis of the raw extract and preliminary fractions within a mixture analysis scheme has been used to identify possible active compounds. This approach would also aid prioritizing promising samples among a sample set or a bank of extracts.[6]

The main analytical techniques for chemical data acquisition are Mass Spectrometry (MS) and Nuclear Magnetic Resonance (NMR). Typically, the users choose the analytical technique according to the study design considering sensitivity, selectivity and sample availability as well as physicochemical properties of the sample itself. However, the complementary aspect of each technique must be emphasized specially for high-confidence compound identification. In the case of bioactive compounds search, these analytical data must be integrated with the results from bioassays. Here, the term bioactivity data will be used to define the set of values obtained from bioassays (e.g., % inhibition, IC50, LC50, etc). Different approaches were suggested in the literature on how to integrate chemical data with bioassay results to reach bioactive compounds in early research stages.[6, 7, 8] Within these efforts, multivariate analysis (i.e., PCA, PLS-DA) of the chemical data of samples classified into active and inactive in a given bioassay is the straightforward method to follow.[9, 10]

NP Analyst is a recent tool developed as a webserver to integrate MS metabolomics data with a series of bioactivity assays such as antimicrobial panels.[7] Through NP Analyst users can integrate data from a big library of samples to prioritize those pointed out as promising; for instance, the parameter consistency of bioactivity of each MS features present. Developed by the same group (Linington Lab, SFU), MADByTE is a tool that searches for sharing TOCSY spin systems (in combination with HSQC peaks) from NMR data to create networks that can be used to visualize samples according to their characteristics and biological profiles.[8] Alternatively, some studies apply multivariate methods such as PLS-DA to classify active versus inactive samples and extract features of interest from the loadings output. This method has been proved beneficial for selection of promising features and compounds of interest. However, by categorizing samples as a binary classification (active or inactive) classification using ranges of the biological assay readout, one tends to ignore the range of variations in these data. Such natural variation of the biological assay readouts, which is related to compound concentrations, must be explored as valuable data.

Statistical Total Correlation Spectroscopy (STOCSY) is performed by calculating the covariance and the correlation between the peaks (data points) across different samples of a dataset to highlight the highly correlated with high positive covariance peaks assuming they are from the same molecular structure.[11] Covariance stands for the combined variability between different variables and correlation stands for the linearity of that combined variability of those different variables; essentially the parameters covariation and correlation try to describe how different variables (in this case, features) behave in analogous manners. Thus, STOCSY is initiated by the selection of a peak (a driver peak) to yield covariance and correlation values between this driver peak and all other variables across the different samples. This takes into consideration the fact that under quantitative parameters, NMR peaks from different nuclei from the same molecule will have specific intensities (according to the number of magnetic active spins) and all of those peaks will vary with high correlation between different samples according to their concentrations. STOCSY calculations have been applied to several cases in metabolomics studies to facilitate the biomarker identification stage.[11] The application of STOCSY like calculations in different datasets, such as NMR and MS, has gained more attention over the last five years. Statistical HeteroSpectroscopy (SHY) was first proposed by Crockford, Maher, Ahmadi, Barrett, Plumb, Wilson and Nicholson [12] for the combination of data acquired from NMR and MS to deliver more informative results. Their goal was to identify biomarkers using data combined from NMR and MS. Generally, the relative variation of a peak in NMR will be highly correlated with the variation of a peak in MS if they refer to the same compound. The authors effectively correlated UPLC-MS and NMR data from drug metabolites and proved an unexpected unreported metabolite derivative of disopyramide and many others from ibuprofen. Hao, Liebeke, Sommer, Viant, Bundy and Ebbels [13] have shown that the correlation between direct infusion mass spectrometry (DIMS) and NMR was “surprisingly successful in linking structurally related signals” indicating the value of this approach even with the possibility of ion suppression.

In this sense, our study introduces a notebook-based application called DAFdiscovery (Data Fusion-based Discovery) to assist STOCSY/SHY users to combine their NMR data with MS and/or bioactive data in an easy-to-follow Jupyter Notebook. DAFdiscovery enables users with no previous programming skills to fuse data from different sources (NMR, MS, and/or Bioassay) to apply a STOCSY/SHY function (written in Python and adapted from Robinette, Brüschweiler, Schroeder and Edison [14]), to produce statistically relevant NMR spectral plot and a correlation result for the probed MS features.

RESULTS AND DISCUSSION

DAFdiscovery was developed to improve dissemination of STOCSY and SHY calculations to natural product scientists, enabling data fusion and discovery of compounds of interest through correlation calculations. A tutorial is presented as Supporting Information (S1). Briefly, STOCSY runs from the data submitted and correlation and covariance are calculated according to a selected driver. This driver must be selected by the users and it can be an MS feature or an NMR peak of interest or the bioactivity readout. SHY is similar to STOCSY but this acronym is used when data is produced by different technical sources. The choice of using Jupyter Notebook with Python as the main platform for this method was due to their availability. Thus, this method does not require any strong programming capabilities from prospective users. Intentionally, it was written valuing readability to be amenable and also stimulating to users interested in gaining programming skills.

The use of .csv (comma-separated values) files enables users to apply their own processing methods and software of choice for data processing. To note, .csv files are text-based files where information (values or text) is separated by commas. This implementation was developed using .csv files exported from MZMine2 and MNova using a Metadata file to organize filenames in their respective order (please refer to https://github.com/RicardoMBorges/DAFdiscovery/wiki/Tutorial-for-DAFdiscovery to more details on). Different bioactivity assays can be used as entries as the bioactivity data since the input is also a .csv file, and so it is assay-agnostic. Note that DAFdiscovery does not accept as input a matrix with rows containing only zeros (e.g., MS feature with no peaks across samples) since it is based on correlation calculations; it is not possible to solve a division where a denominator is zero (with Python this will result in 'NaN' meaning 'not a number'). The input files require samples in columns where each column header represents one of the sample names (or filename) and the rows represents the feature numbers (MS, NMR, or bioactivity readout). A metadata .csv file is used to organize samples and their respective file names for reordering. Figure 1 illustrates the input file requirements.

As proof of concept, 5 case studies have been used to test DAFdiscovery. No interpretation was made for each data used as a case study. As aforementioned, DAFdiscovery was developed to combine MS, NMR, and bioactivity data, or any combination of two of them; also, NMR data can be used alone. Thus, the pipeline is separated into 5 options: (Option 1) fusion of NMR, MS, and Bioassay data; (Option 2) fusion of NMR and MS data; (Option 3) fusion of NMR and Bioassay data; (Option 4) fusion of MS and Bioassay data; and (Option 5) for NMR alone.

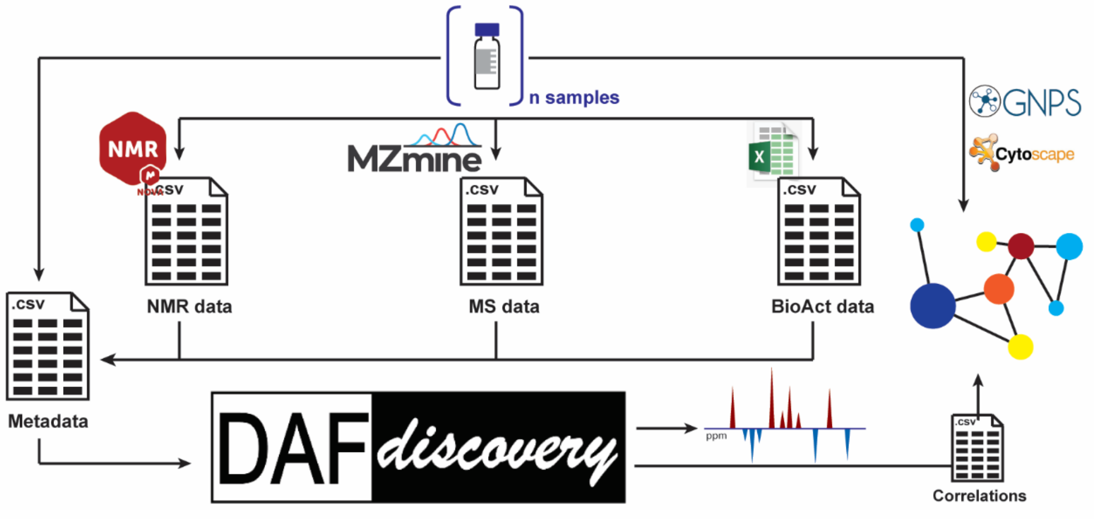


Figure 1. Pipeline designed for DAFdiscovery highlighting the use of .csv files (Metadata, NMR data, MS data, and BioAct data) to achieve a STOCSY-modulated NMR spectra and the indexed correlation results to color-code GNPS-FBMN networks using Cytoscape.

A pipeline designed for data fusion-based discoveries was developed to disseminate the use of the STOCSY function on data acquired from different sources, including bioactivity readouts. It accepts data exported as flat .csv files so that users can continue with their preferred data processing tools of choice (e.g., MNova and MZMine). This pipeline was written in Jupyter Notebook and Python to guarantee a user-friendly approach. Notebooks and files are available at https://github.com/RicardoMBorges/DAFdiscovery. A full tutorial with basic information on how to run Jupyter Notebook is available at https://github.com/RicardoMBorges/DAFdiscovery/wiki/Tutorial-for-DAFdiscovery.

Case I. The application of STOCSY calculation within NMR metabolomics is well accepted and there is no novelty to this demonstration. This option was added into this method as an obvious application use. Thus, as a case study, the NMR data downloaded from Metabolights (MTBLS2052 – "Tissue, urine and serum NMR metabolomics dataset from a 5/6 nephrectomy rat model of chronic kidney disease"[15]) was processed and submitted to DAFdiscovery. To demonstrate the application of STOCSY, the peak at δ_H 3.98 was selected as the driver to reveal the statistically pure spectra of hippuric acid with the highly correlated peaks at δ_H 7.56, 7.64, and 7.84 (Figure 2-A). Another clear result is the acquired spectra of citrate with highly correlated peaks at δ_H 2.55 and 2.70 (Figure 6-B).

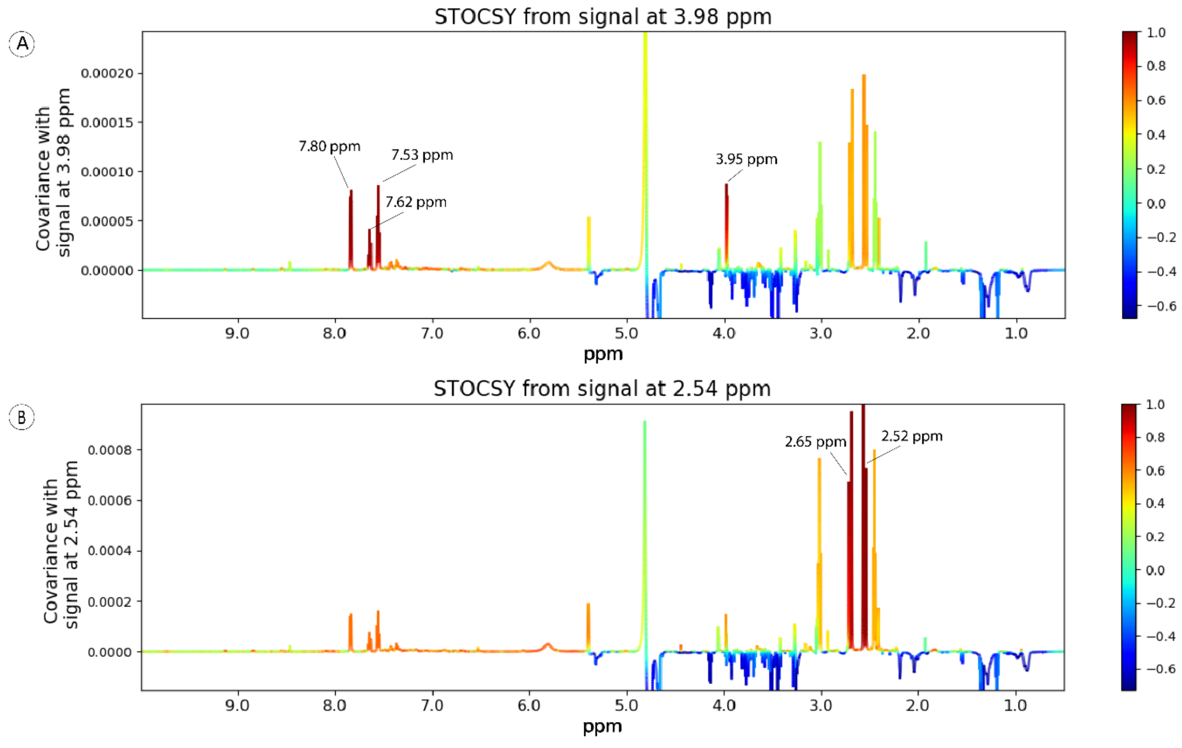


Figure 2. Output scheme from DAFdiscovey on ^{1}H NMR data. The NMR STOCSY plot shows NMR resonance highly correlated to the driver peak at δ_H 3.98 (A) and δ_H 2.54 (B). (A) The statistically pure spectrum of hippuric acid is highlighted in bright red. (B) The statistically pure spectrum of citrate is highlighted in bright red. Data downloaded from Metabolights (MTBLS2052 – "Tissue, urine and serum NMR metabolomics dataset from a 5/6 nephrectomy rat model of chronic kidney disease"[15]).

Case II. LC-MS data (MS level 1) and cytotoxicity assay using the *Artemia salina* model assay (unpublished data) of a set of 6 prefractions, analyzed in duplicates, derived from a crude extract of a cyanobacteria strain is demonstrated to describe Option 4 (MS + Bioactivity correlations). In such cases where NMR data is not available, the NMR STOCSY plot will not be of any use. Instead, a scatterplot with retention time vs *m/z* is used to visualize MS features highly correlated with the bioactivity readouts (Figure 3). Here, the bioactivity results across the samples are chosen as the driver peak for the STOCSY calculation and the goal is to spot highly correlated MS-features with positive covariance. In other words, to promote MS-features that varies similarly with the bioactivity results. The users may focus their attention on these larger and bright-red colored marks to explore MS-features positively and highly correlated MS-features (data not published). By plotting this view as a scatterplot of retention time vs *m/z*, users might even identify different adducts if it is the case.

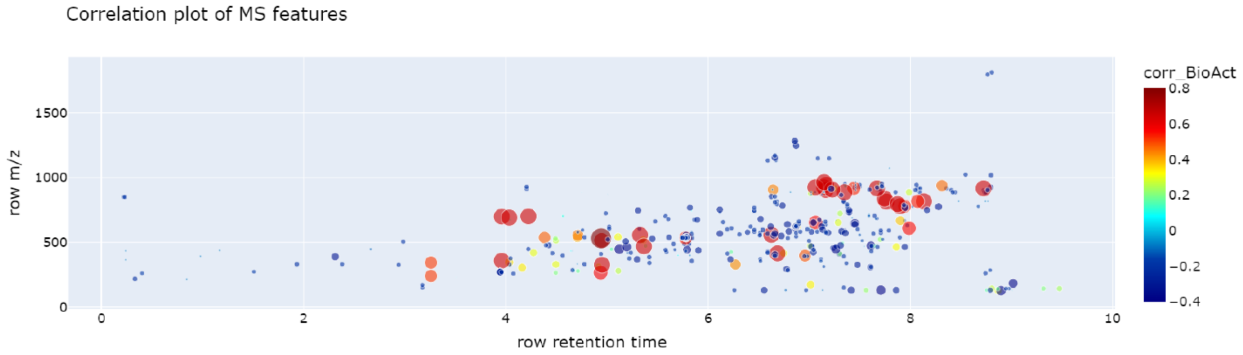


Figure 3. Output scheme from DAFdiscovey on a set of LC-MS and bioactivity data. The retention time vs *m/z* scatterplot shows MS features highly correlated to the bioactivity readouts (chosen as the driver). Unpublished data provided by Professor Camila Manoel Crnkovic (USP, Brazil).

For natural products research, definition of the chemical space and dereplication are key, and MS based analysis is the approach most commonly followed. There are several reasons for that: very high sensitivity and selectivity to detect higher ranges of compounds; its use as a high-throughput analytical instrumentation is easier than NMR for example; and MS-based databases are more comprehensive than their NMR counterparts.[16]

Case III. As another example, GC-MS data and the bioactivity effect against the larval population of *Aedes aegypti* [17] is also demonstrated to describe Option 4 (MS + Bioactivity correlations) (Figure 4-A). The .csv file with the correlations output from STOCSY indexed according to each MS-feature was imported as a node attribute to the available molecular network to produce a simplified view of the chemical space analyzed (Figure 4-B). To validate this approach, LC_{50} (µg/ml) of the annotated compounds was compared to the MS-features highlighted from the present pipeline. As expected, the MS-features at 26.63 min (annotated as an isomer of himachalene), 24.59 min (annotated as an isomer of himachalene), 25.69 min (annotated as longifolene), and 7.91 min (annotated as an isomer of eucalyptol) min were found to be positively correlated to the LC_{50} values obtained from the analyzed samples. In addition, the MS-features at 15.45 min (annotated as citronellol), 17.25 (annotated as citral) and 27.18 min (unknown) were found to be negatively correlated to the LC_{50} values; these indeed have shown lower LC_{50} results. Here, again the bioactivity results across the samples are chosen as the driver peak for the STOCSY calculation, but the bioactivity is reported as LC_{50} and it will have lower values when the active compound is more concentrated. Thus, the goal here is to spot highly correlated MS-features with negative covariance.

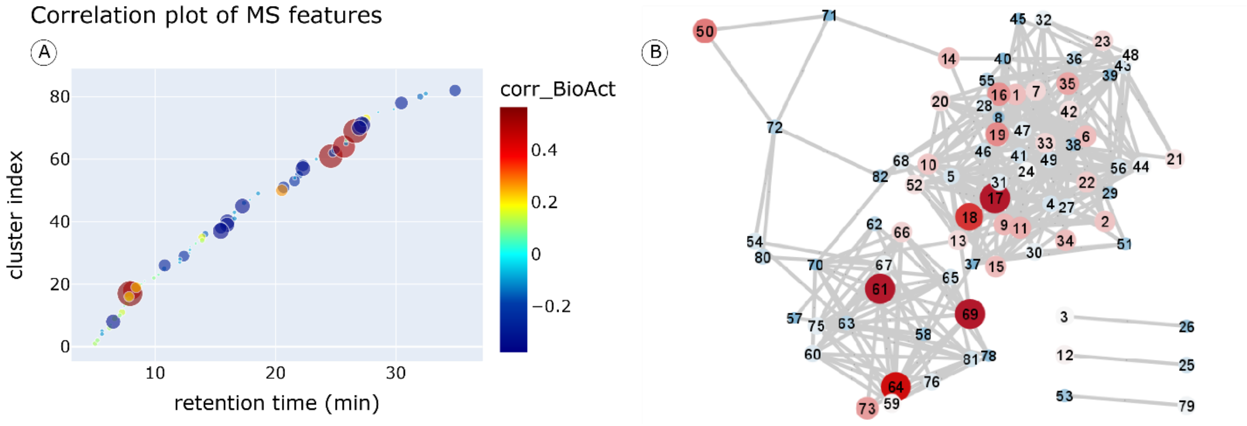


Figure 4. Output scheme from DAFdiscovey on a set of GC-MS and bioactivity data. The retention time vs scan number scatterplot shows MS features highly correlated to the bioactivity readouts (A). Molecular network from [17] where the nodes were color coded according to the resulting correlation output from STOCSY using the larvicidal effect against *A. aegypti* (B). Published data [17], provided by Dr. Alan Pilon (USP, Brazil).

Case IV. As part of the efforts to discover antimicrobial compounds from Actinobacteria strains, a sample set created through four growth media of the *Streptomyces* sp. PNM-9 was analyzed using NMR-based metabolic profiling and antibacterial activity assay against *Burkholderia glumae* [9]. The authors used a principal component analysis (PCA) and an orthogonal projection to latent structures discriminant analysis (OPLS-DA) method to identify metabolic differences and distinguish samples between active and inactive. Using this approach and the bioactivity (reported as the minimal inhibitory concentration, MIC) was selected as driver, specific ^{1}H-NMR peaks were highly correlated and negatively covaried. The same dataset was submitted to DAFdiscovery using the raw values for the antibacterial assay (without the classification into active or inactive) to reach equivalent results mainly for peaks at δ_H 7.47 and δ_H 5.62 corresponding to the active phenylethyl amides as pointed out in the original paper. Highly correlated peaks (with negative covariance since MIC will indicate lowest concentration of and active compound that prevents bacterial growth) are identified as peaks of interest (Figure 5).

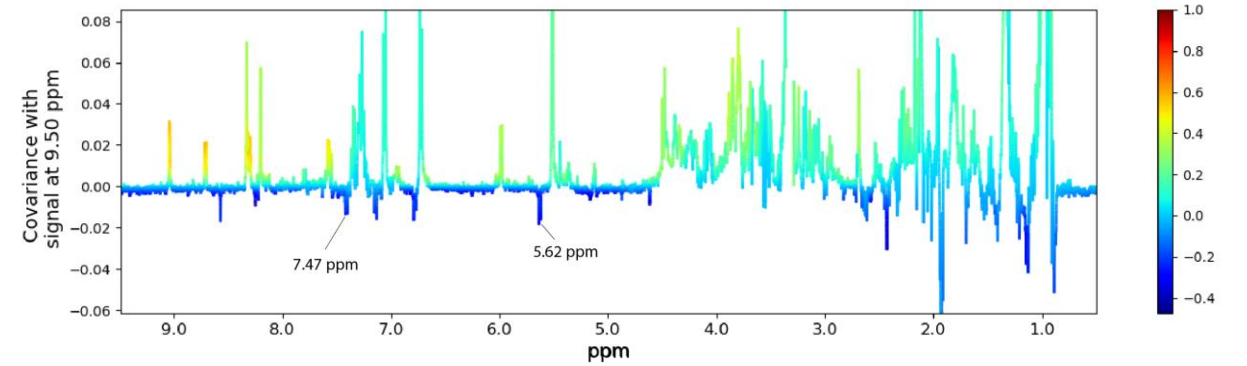


Figure 5. Output NMR STOCSY plot shows NMR resonance highly correlated to the antibacterial assay. Published data [9] provided by Professors Freddy A. Ramos and Monica T. Pupo, and collaborators.

Case V. A sample set that included 10 samples of essential oils from *Melaleuca alternifolia* and *Melaleuca rhaphiophylla* for a study of species differentiation was submitted to GC-MS and ^{13}C NMR for data acquisition.[18] This data was used in a previous paper that demonstrated a similar fused data approach but the pipeline was processed using Matlab. The MS data was processed using MZMine2 for mass detection, ion chromatogram building, chromatographic deconvolution, spectral deconvolution, and peak alignment. The NMR data was processed using MNova for chemical shift reference, phase correction, normalization, and alignment. Both data sets were exported as .csv to be used as input for DAFdiscovery. Figure 6 shows the NMR peaks (Figure 6-A) and the MS-features (Figure 6-B) highly correlated with the peak at δ_C 71.77 (chosen as driver for STOCSY). Thus, the results became constrained to just a few NMR peaks and MS features to be analyzed for compound identification. Moreover, the correlation results for the MS features can be indexed into the molecular network (Figure 6-C) calculated from the same MS processed data for visualization.

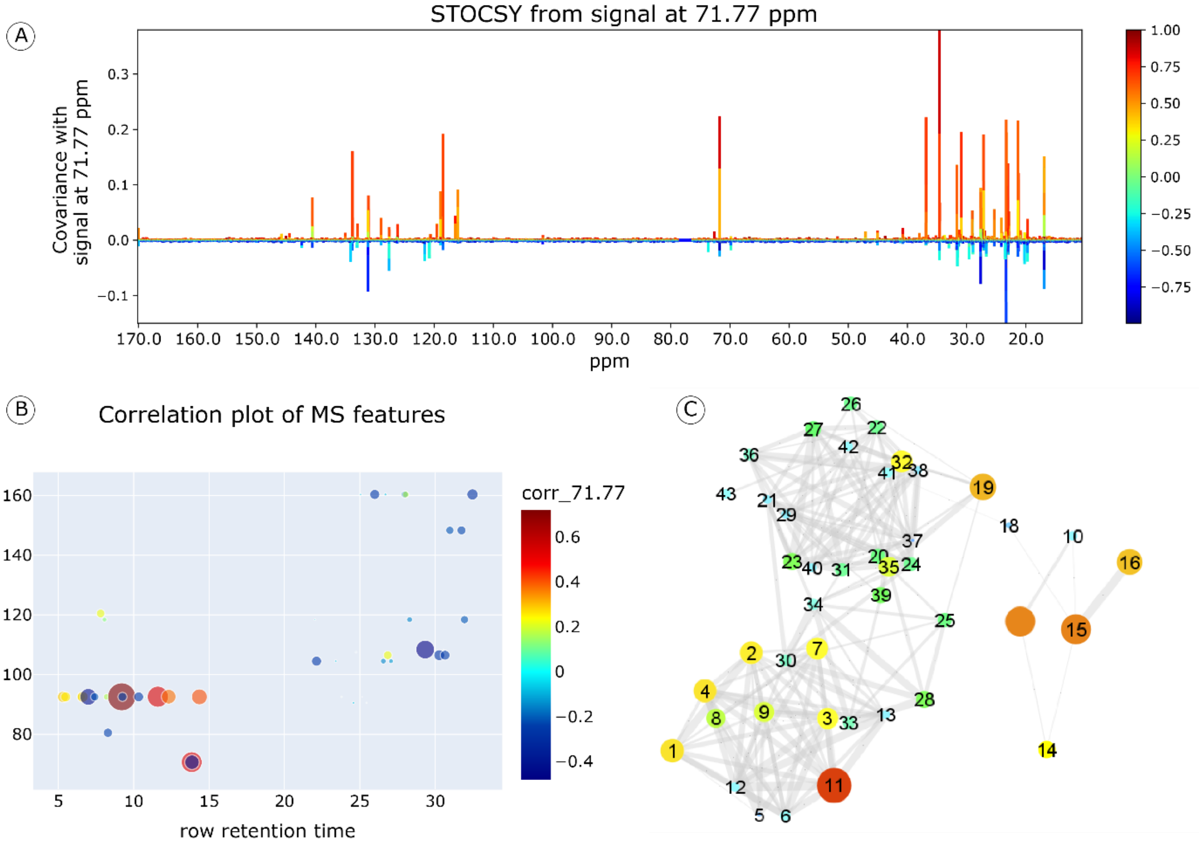


Figure 6. Output scheme from DAFdiscovey on a set of GC-MS and ^{13}C NMR data. The NMR STOCSY plot (A) shows NMR resonance highly correlated to the driver peak at δ_C 71.77. The retention time vs *m/z* scatterplot (B) highlights (size and color) MS features highly correlated to the same driver peak at δ_C 71.77. And the molecular network (C) calculated with the processed GC-MS data where the nodes were color-coded according to their correlation values to the driver peak at δ_C 71.77. Published data [18] provided by Professor Ricardo Moreira Borges (UFRJ, Brazil).

To validate this pipeline, different case studies were shown to exemplify diverse combinations among NMR, MS, and bioactivity data. The resulting outputs are present for each combination and they demonstrate how convenient this approach can be for compound identification and drug discovery within natural products research. The authors do strongly encourage users to apply this method on their traditional bioguided fractionation approach for bioactive natural products discovery and to acquire as much data as possible from every single fraction to create variance. Users can explore biological variance from different extracts or chromatographic variances created by fractionation using DAFdiscovery to accelerate bioactive compound discovery or for annotation-identification studies.

EXPERIMENTAL SECTION

General DAFdiscovery pipeline. The choice of Jupyter Notebook (and Python) for the pipeline development was made since it is an open-source product, free, and has good visualization schemes. Hopefully, this will also encourage students in their early careers to start developing programming skills. A tutorial for installation of Anaconda Jupyter Notebook and to download DAFdiscovery is available in the Supporting Information.

The main part of this pipeline is the STOCSY function that calculates the covariance and correlation between a certain feature across the samples with all other features present. The output of STOCSY is a 1D NMR spectrum with signal intensities modulated by the covariance and color-coded according to the correlation. In addition, a .csv file with the correlation results for each MS feature is produced when MS data is submitted. This MS derived .csv file is intended to be used as a node attribute table in Cytoscape to color-code a molecular network created following the GNPS-FBMN workflow [19]. If the user submits NMR, MS, and bioactivity data from the same sample set, the proposed pipeline will highlight NMR peaks and MS nodes highly correlated with the bioactivity results. Finally, .csv files exported from well-known processing software are used as input because the authors recognize the efficiency of the existing data processing methods for both MS and NMR and to make it vendor-agnostic.

The Pipeline. The users start with a .csv Metadata file with the columns 'Ordered_Samples', 'Ordered_NMR_filename' (when NMR data is used), 'Ordered_MS_filename' (when MS data is used), and 'Ordered_BioAct_filename' (when bioassay data is used) (Figure S1-S4). This file must contain the exact filename used in the previous data processing step. The processing step for each technique is independent of DAFdiscovery which will use only the exported .csv file from each specific dataset. The pipeline will adopt the order given in the Metadata file to reorder the data files from each specific technique according to the filename given by the expected .csv files. Then, the STOCSY function is applied and the resulting statistical NMR spectra, and the MS feature list with the correlations are stored for interpretation.

DAFdiscovery accepts data from NMR, MS, and bioassays to perform data fusion of every given possibility. Thus, users can choose to run STOCSY on: (1) NMR, MS, and Bioassay; (2) NMR and MS; (3) NMR and Bioassay; (4) MS and Bioassay; and (5) NMR only. DAFdiscovery can be downloaded here: https://github.com/RicardoMBorges/DAFdiscovery. A tutorial for installation is available here:

https://github.com/RicardoMBorges/DAFdiscovery/wiki/Tutorial-for-DAFdiscovery. In order to keep files organized, each project should be kept in their own "Project_" directory. Within the pipeline, the first input is the definition of the project directory name and all the produced files and images will be saved inside it.

Data. Five different datasets were used as case studies to demonstrate the use of DAFdiscovery. Case I represents a dataset solely with NMR data from a metabolomics study. It was downloaded from Metabolights (MTBLS2052 – "Tissue, urine and serum NMR metabolomics dataset from a 5/6 nephrectomy rat model of chronic kidney disease"[15]). Case II is an unpublished data of a screening study for cyanobacterial secondary metabolites (analyzed by LC-MS) potentially active in the *Artemia salina* model assay. This dataset was provided and coordinated with Professor Camila Manoel Crnkovic (USP, Brazil). Case III represents a dataset comprising GC-MS and a bioassay against the larval population of *Aedes aegypti* [17]. This dataset was provided and coordinated with Dr. Alan Pilon (USP, Brazil). Case IV demonstrated the use of NMR and antibacterial activity assay against *Burkholderia glumae* of an Actinobacteria strain grown in different media conditions.[9] This dataset was provided by Professors Freddy A. Ramos and Monica T. Pupo. Case V represents the combination of MS and NMR data for confidence annotation of compounds. This data was acquired from essential oils from *Melaleuca alternifolia* and *Melaleuca rhaphiophylla* for a study of species differentiation and it was provided by Professor Ricardo Moreira Borges (UFRJ, Brazil).

ACKNOWLEDGMENTS

The authors acknowledge Walter Mors Institute of Research on Natural Products, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) for the grant 210.489/2019 APQ-1, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) for the grant 2020/07710-7. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001. Brazilian federal funding agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) is also acknowledged.

REFERENCES

(1) Weller, M. G. A unifying review of bioassay-guided fractionation, effect-directed analysis and related techniques. *Sensors (Basel)* **2012**, *12* (7), 9181-9209. DOI: 10.3390/s120709181 PubMed.
(2) Lautié, E.; Russo, O.; Ducrot, P.; Boutin, J. A. Unraveling Plant Natural Chemical Diversity for Drug Discovery Purposes. *Frontiers in Pharmacology* **2020**, *11*, Review. DOI: 10.3389/fphar.2020.00397.
(3) Campbell, W. C. History of avermectin and ivermectin, with notes on the history of other macrocyclic lactone antiparasitic agents. *Curr Pharm Biotechnol* **2012**, *13* (6), 853-865. DOI: 10.2174/138920112800399095 From NLM.
(4) Wall, M. E.; Wani, M. C. Camptothecin and taxol: from discovery to clinic. *J Ethnopharmacol* **1996**, *51* (1-3), 239-253; discussion 253-234. DOI: 10.1016/0378-8741(95)01367-9 From NLM.

(5) Kingston, D. G. I.; Cassera, M. B. Antimalarial Natural Products. *Prog Chem Org Nat Prod* **2022**, *117*, 1-106. DOI: 10.1007/978-3-030-89873-1_1 From NLM.
(6) Olivon, F.; Allard, P. M.; Koval, A.; Righi, D.; Genta-Jouve, G.; Neyts, J.; Apel, C.; Pannecouque, C.; Nothias, L. F.; Cachet, X.; et al. Bioactive Natural Products Prioritization Using Massive Multi-informational Molecular Networks. *ACS Chem Biol* **2017**, *12* (10), 2644-2651. DOI: 10.1021/acschembio.7b00413. Nothias, L. F.; Nothias-Esposito, M.; da Silva, R.; Wang, M.; Protsyuk, I.; Zhang, Z.; Sarvepalli, A.; Leyssen, P.; Touboul, D.; Costa, J.; et al. Bioactivity-Based Molecular Networking for the Discovery of Drug Leads in Natural Product Bioassay-Guided Fractionation. *J Nat Prod* **2018**, *81* (4), 758-767. DOI: 10.1021/acs.jnatprod.7b00737. Wolfender, J. L.; Litaudon, M.; Touboul, D.; Queiroz, E. F. Innovative omics-based approaches for prioritisation and targeted isolation of natural products - new strategies for drug discovery. *Nat Prod Rep* **2019**, *36* (6), 855-868. DOI: 10.1039/c9np00004f.
(7) Lee, S.; van Santen, J. A.; Farzaneh, N.; Liu, D. Y.; Pye, C. R.; Baumeister, T. U. H.; Wong, W. R.; Linington, R. G. NP Analyst: An Open Online Platform for Compound Activity Mapping. *ACS Central Science* **2022**. DOI: 10.1021/acscentsci.1c01108.
(8) Egan, J. M.; van Santen, J. A.; Liu, D. Y.; Linington, R. G. Development of an NMR-Based Platform for the Direct Structural Annotation of Complex Natural Products Mixtures. *J Nat Prod* **2021**. DOI: 10.1021/acs.jnatprod.0c01076.
(9) Betancur, L. A.; Forero, A. M.; Vinchira-Villarraga, D. M.; Cardenas, J. D.; Romero-Otero, A.; Chagas, F. O.; Pupo, M. T.; Castellanos, L.; Ramos, F. A. NMR-based metabolic profiling to follow the production of anti-phytopathogenic compounds in the culture of the marine strain Streptomyces sp. PNM-9. *Microbiol Res* **2020**, *239*, 126507. DOI: 10.1016/j.micres.2020.126507.
(10) Borges, D. G. L.; Echeverria, J. T.; de Oliveira, T. L.; Heckler, R. P.; de Freitas, M. G.; Damasceno-Junior, G. A.; Carollo, C. A.; Borges, F. A. Discovery of potential ovicidal natural products using metabolomics. *PLoS One* **2019**, *14* (1), e0211237. DOI: 10.1371/journal.pone.0211237.
(11) Robinette, S. L.; Lindon, J. C.; Nicholson, J. K. Statistical spectroscopic tools for biomarker discovery and systems medicine. *Anal Chem* **2013**, *85* (11), 5297-5303. DOI: 10.1021/ac4007254. Sands, C. J.; Coen, M.; Ebbels, T. M.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Data-driven approach for metabolite relationship recovery in biological 1H NMR data sets using iterative statistical total correlation spectroscopy. *Anal Chem* **2011**, *83* (6), 2075-2082. DOI: 10.1021/ac102870u. Smith, L. M.; Maher, A. D.; Cloarec, O.; Rantalainen, M.; Tang, H.; Elliott, P.; Stamler, J.; Lindon, J. C.; Holmes, E.; Nicholson, J. K. Statistical correlation and projection methods for improved information recovery from diffusion-edited NMR spectra of biological samples. *Anal Chem* **2007**, *79* (15), 5682-5689. DOI: 10.1021/ac0703754.
(12) Crockford, D. J.; Maher, A. D.; Ahmadi, K. R.; Barrett, A.; Plumb, R. S.; Wilson, I. D.; Nicholson, J. K. 1H NMR and UPLC-MSE Statistical Heterospectroscopy: Characterization of Drug Metabolites (Xenometabolome) in Epidemiological Studies. *Analytical Chemistry* **2008**, *80*, 6835-6844.
(13) Hao, J.; Liebeke, M.; Sommer, U.; Viant, M. R.; Bundy, J. G.; Ebbels, T. M. Statistical Correlations between NMR Spectroscopy and Direct Infusion FT-ICR Mass Spectrometry Aid Annotation of Unknowns in Metabolomics. *Anal Chem* **2016**, *88* (5), 2583-2589. DOI: 10.1021/acs.analchem.5b02889.
(14) Robinette, S. L.; Brüschweiler, R.; Schroeder, F. C.; Edison, A. S. NMR in metabolomics and natural products research: two sides of the same coin. *Accounts of chemical research* **2012**, *45* (2), 288-297. DOI: 10.1021/ar2001606 From NLM.
(15) Hanifa, M. A.; Skott, M.; Maltesen, R. G.; Rasmussen, B. S.; Nielsen, S.; Frøkiær, J.; Ring, T.; Wimmer, R. Tissue, urine and blood metabolite signatures of chronic kidney disease in the 5/6 nephrectomy rat model. *Metabolomics* **2019**, *15* (8), 112. DOI: 10.1007/s11306-019-1569-3.
(16) Sindelar, M.; Patti, G. J. Chemical Discovery in the Era of Metabolomics. *J Am Chem Soc* **2020**, *142* (20), 9097-9105. DOI: 10.1021/jacs.9b13198. Sorokina, M.; Steinbeck, C. Review on natural products databases: where to find data in 2020. *J Cheminform* **2020**, *12* (1), 20. DOI: 10.1186/s13321-020-00424-9.

(17) Pilon, A. C.; Del Grande, M.; Silvério, M. R. S.; Silva, R. R.; Albernaz, L. C.; Vieira, P. C.; Lopes, J. L. C.; Espindola, L. S.; Lopes, N. P. Combination of GC-MS Molecular Networking and Larvicidal Effect against Aedes aegypti for the Discovery of Bioactive Substances in Commercial Essential Oils. *Molecules* **2022**, *27* (5). DOI: 10.3390/molecules27051588.

(18) Borges, R. M.; Resende, J. V. M.; Pinto, A. P.; Garrido, B. C. Exploring correlations between MS and NMR for compound identification using essential oils: A pilot study. *Phytochem Anal* **2022**. DOI: 10.1002/pca.3107.

(19) Nothias, L. F.; Petras, D.; Schmid, R.; Duhrkop, K.; Rainer, J.; Sarvepalli, A.; Protsyuk, I.; Ernst, M.; Tsugawa, H.; Fleischauer, M.; et al. Feature-based molecular networking in the GNPS analysis environment. *Nat Methods* **2020**, *17*, 905–908. DOI: 10.1038/s41592-020-0933-6.