

RESEARCH ARTICLE

Propagating annotations of molecular networks using *in silico* fragmentation

Ricardo R. da Silva^{1,2}, Mingxun Wang¹, Louis-Félix Nothias¹, Justin J. J. van der Hooft^{1,3}, Andrés Mauricio Caraballo-Rodríguez¹, Evan Fox⁴, Marcy J. Balunas⁵, Jonathan L. Klassen⁴, Norberto Peporine Lopes², Pieter C. Dorrestein^{1*}

1 Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, United States of America, **2** NPPNS, Department of Physic and Chemistry, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, SP, Brazil, **3** Bioinformatics Group, Department of Plant Sciences, Wageningen University, Wageningen, The Netherlands, **4** Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT, United States of America, **5** Division of Medicinal Chemistry, Department of Pharmaceutical Sciences, University of Connecticut, Storrs, CT, United States of America

* pdorrestein@ucsd.edu



OPEN ACCESS

Citation: da Silva RR, Wang M, Nothias L-F, van der Hooft JJJ, Caraballo-Rodríguez AM, Fox E, et al. (2018) Propagating annotations of molecular networks using *in silico* fragmentation. PLoS Comput Biol 14(4): e1006089. <https://doi.org/10.1371/journal.pcbi.1006089>

Editor: Avner Schlessinger, Icahn School of Medicine at Mount Sinai, UNITED STATES

Received: November 22, 2017

Accepted: March 13, 2018

Published: April 18, 2018

Copyright: © 2018 da Silva et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The raw data Euphorbia dendroides was retrieved from the public MassIVE datasets for Euphorbia dendroides (<ftp://massive.ucsd.edu/MSV000080502>), fecal (<ftp://massive.ucsd.edu/MSV000081120>) and a fungal data set (<ftp://massive.ucsd.edu/MSV000081671>). The CASMI data was downloaded from (<http://www.casmi-contest.org/2016/index.shtml>). The networks were calculated using GNPS web interface, and can be accessed with the following job IDs: for NIST library - <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=>

Abstract

The annotation of small molecules is one of the most challenging and important steps in untargeted mass spectrometry analysis, as most of our biological interpretations rely on structural annotations. Molecular networking has emerged as a structured way to organize and mine data from untargeted tandem mass spectrometry (MS/MS) experiments and has been widely applied to propagate annotations. However, propagation is done through manual inspection of MS/MS spectra connected in the spectral networks and is only possible when a reference library spectrum is available. One of the alternative approaches used to annotate an unknown fragmentation mass spectrum is through the use of *in silico* predictions. One of the challenges of *in silico* annotation is the uncertainty around the correct structure among the predicted candidate lists. Here we show how molecular networking can be used to improve the accuracy of *in silico* predictions through propagation of structural annotations, even when there is no match to a MS/MS spectrum in spectral libraries. This is accomplished through creating a network consensus of re-ranked structural candidates using the molecular network topology and structural similarity to improve *in silico* annotations. The Network Annotation Propagation (NAP) tool is accessible through the GNPS web-platform <https://gnps.ucsd.edu/ProteoSAFe/static/gnps-theoretical.jsp>.

Author summary

For genome analysis it is commonly accepted that one can hypothesize the function of genes based on sequence similarity, using annotated reference sequences. Once a homology hypothesis has been made based on reference annotations, it allows one to build hypothesis in terms of function and ultimately understand the underlying biology. In contrast, mass spectrometry (MS) can detect many molecules, as ions, yet we often cannot link a MS signal to a molecule. The reference libraries to annotate fragmented molecular

aa3386fd782e4875b6109fb32a93eb5a, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=229296bd89fc4eb19c4d4cb4e6d50744>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8ae2cbf410944c96a269e80670915851>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ee1233f263f94268ac62cd4cd358cd12>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=00851eb8fb2c4050b581ab898b9d228e>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1f905afc070241a58c74672b40333ac0>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e6663cf00af64a928def7a70108a2b19>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f2f4d89b7dbc4fc0ae78591b36f71585>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c26168effc244aedb95fbae7de289aaf>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7e45eb1bd3c34cc9a8a27bf20c182f4d>; NIST library subnetwork with cosine score < 0.7 <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8b0b5a467da916b81ab4f925a4f4b43>; for Fecal, Euphorbia dendroides extracts and Fungal dataset <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f0cab9c92247d44789900944a69874e8a>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ce2a564dbd704c0595494e04798b0233>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b753797b0dad4f1e84142dd59c84615b>; and finally for CASMI negative mode <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a3f02b1b648a43b6a210063a4ee2f787> and positive mode <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=231902c6d75f41df8403e454c96e8d4a>. The parameters can be accessed by cloning the job or at the link "Networking Parameters and Written Network Description". The corresponding NAP jobs can also be accessed through the web interface with the following job IDs: for NIST library - <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=29d517e67067476bae97a32f2d4977e0>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=d270e79876cb48deb6aab52a4fc647e>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=e2125577fe2646129becc248b96d42ba>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=81e01fe178d3424686079903d908b536>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=daa546b038604e5f83eaaaf811bd0313>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=61c8a0d01309408f8ecceb5b31dab1a8>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=60fe9f77b3d04789997bf19aa1a0a828>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=53f8494f9e8423697eebf4e98d287f0>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=c93a840100ec49b4bb3c12e5ed1e4790>,

data only cover a small portion of the known molecular space. The use of computational (*in silico*) fragmentation predictions from structural libraries offers a promising alternative. One of the weaknesses of the molecular annotation using such *in silico* approaches is that they currently annotate the molecules individually. However, molecular relationships, based on spectral similarity, can be used to enhance the structural hypothesis inferred from the annotation of molecules detected by mass spectrometry. We introduce an online tool called "Network Annotation Propagation" that uses a combination of molecular networks, based on spectral similarity, from which we infer molecular similarity, together with *in silico* fragmentation, to enable the scientific community to strengthen their MS annotations.

Introduction

One way to gain insight into the molecules of a biological sample is through mass spectrometry. Mass spectrometers are incredibly sensitive equipment capable, under specific conditions, of measuring attograms (10^{-18} g) of molecules in a sample [1]. In a targeted mass spectrometry analysis, such as airport security scans, for example, only molecular signatures of predetermined compounds (e. g. explosive components) are searched. In an untargeted mass spectrometry experiment, we do not set the mass spectrometer to weigh specific molecules only, instead, we have the potential to observe hundreds to thousands of ions from a single sample; but most experiments report only on one or a few dozen molecules and often within the limit of known pathways described in textbooks [2]. However, such pathways represent only a fraction of molecules that are detected.

Untargeted mass spectrometry is usually performed as follows: liquid chromatography based infusion is led into the instrument, then the ions are isolated inside the mass spectrometer, accelerated in a chamber filled with helium gas, for example, which results in thermal activation due to the collisions with the gas. When the ion is sufficiently activated, some of the molecule's bonds break and the resulting fragments can be observed. Untargeted mass spectrometry of laboratory cultured organisms, or even cells is already very complex, and this complexity increases in real world samples from plants, animals (including humans) and environments, as the molecules can come from the host cells, diet, their microbiome as well as any other environmental exposure. The way we convert the fragmentation spectrum of fragmented molecules, also referred to as an MS/MS, to identifications is through matching the observed MS/MS spectra to spectra from reference MS/MS libraries, termed spectral library search. The goal of the spectral library search is to find the best MS/MS match between an unknown spectrum and a known spectrum of a previously characterized molecule in the library. On average, through matching fragmented spectra with reference libraries, we can annotate 2% (average of spectral library matching in all GNPS datasets) of the data [3], although for well studied biological matrices, such as *Escherichia coli*, human cell lines, plasma or urine this may be as high as 10% [4].

To better understand this complexity, we must be able to increase our annotation rates of spectra across all projects in the public domain. Using spectral alignment to create molecular networks [5], molecules can be grouped into molecular families (inferred structural analogs) and annotations from known compounds can be propagated to connected spectra (neighbors) in the network (Fig 1A). While molecular networking can double or even triple the annotation rates [5], it requires a tedious process of manual inspection [6]. Reference libraries are commonly generated from commercially available standards, which resulted in very biased

<http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5fd60b02f8ab4274bf45fd5b715b5e0b>; NIST library subnetwork with cosine score < 0.7 <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=74e04164a8374929a4548655742c0a4f>; for Fecal, Euphorbia dendroides extracts and fungal <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=9b6bddc2ba154b1397a53c7f7933430a>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=8ae3aa45bfe449d7969975189b14f429>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=fb02f64992bb4a0fb46b0e4832e69597>; for CASMI positive <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=7c5a16eba2eb42f88647e7d21e57f1bc> and for negative mode <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=e9745e159f6f433d9efb71e9813df29a>. The parameters from NAP can be accessed by cloning the job through the web interface and Supplementary Table 11. The ProteoSAFe web interface can be found at <https://gnps.ucsd.edu/ProteoSAFe/static/gnps-theoretical.jsp>, under the workflow name NAP_CCMS and the code is available at github (https://github.com/DorresteinLaboratory/NAP_ProteoSAFe/).

Funding: The authors would like to acknowledge the scientific support of the Collaborative Mass Spectrometry Innovation Center and the Center for Microbiome Innovation, University of California - San Diego, and funding support from the US National Institutes of Health 1R03CA211211-01 and grant P41GM103484-07 from the National Institute of General Medical Sciences; and from the US National Science Foundation IOS-1656481. We further acknowledge Bruker and NIH grants GMS10RR029121 and R01 GM107550 for the support of the shared instrumentation infrastructure that enabled this work. We thank Madeleine Ernst for test and feedback on the NAP web interface. RRdS was initially supported by the São Paulo Research Foundation (Awards FAPESP-2015/03348-3, 2014/50265-3 and 14/01884-2). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

molecular libraries in the public domain as well as commercial libraries. While we have begun to capture annotations of spectra by the community from unpurified material, using the expertise of public data depositors [5], one of the most promising ways to improve annotations is through *in silico* matching to structural databases covering a broader range of the chemical space. Computational tools are able to capture and infer structural information from mass spectrometry data at a scale that dwarfs manual inspection rates. Although a variety of methods are used for *in silico* fragmentation matching [7], usually multiple candidate structures match a given query spectrum. Several *in silico* fragmentation methods have proposed different criteria to rank the most likely candidates, however, currently the correct structure is in average ranked among the first tens of candidates [8]. This means that an end user still has to wade through the top k matches and visually inspect the predictions, which is one of the reasons why such approaches are slowly adopted by experimentalists. In the last ten years the development of *in silico* fragmentation methods is experiencing a great improvement, in part because experimental data and reference libraries have become available in the public domain, providing training data for increasingly sophisticated algorithms to be developed [9–12]. The development of the next generation *in silico* fragmentation annotation methods hold the promise to increase the number of annotations and make this process much more efficient [13], however we often have multiple candidate structures predicted for each fragmentation spectrum and an end user does not have a good way to prioritize the candidate matches. Although there are other approaches to propagate annotations in a mass spectrometry experiment [14–18], since the introduction of molecular networking in 2012 by our labs, we and others have demonstrated that the concept of propagation through spectral alignments works extremely well [19–25].

We therefore set out to explore if a large portion of the “dark matter” of metabolomics [3] experiments can be uncovered by combining molecular networking with existing *in silico* methods to improve the annotation rates and quality through automated propagation. An early hint of the utility of this approach was recently demonstrated. The Wolfender lab combined an *in silico* library using the *in silico* fragmentation method CFM-ID [9], and molecular networking results obtained through our community analysis platform Global Natural Product Social (GNPS) infrastructure, followed by subsequent manual inspection of the results. The Wolfender lab demonstrated that the combination of spectral networks and *in silico* fragmentation was an effective strategy for dereplication, a term used for identification of known molecules [26]. While this work explores the advantage of the combination of two annotation approaches, the *in silico* prediction method did not yet take advantage of the network topology. The topology should be taken in account, once, under the assumption that neighbor nodes in the spectral networks are structurally related, the *in silico* annotation of neighbor nodes should result in structurally related candidates. Here we show that the ranking of *in silico* annotations using the *in silico* prediction tool MetFrag [12, 27], can be improved using the topology of the molecular network through building consensus among the candidate structures from neighbors in the network. We further show this is effective even without any spectral library match. The approach is called Network Annotation Propagation (NAP). There are two scoring approaches utilized by NAP to re-rank candidates. When there is a spectral library match within a molecular family of the molecular network (connected component of a graph), NAP utilizes the MetFrag *in silico* prediction with the MetFusion [28] score to re-rank candidates (which we term *Fusion* scoring) (Fig 1B and 1E). The *in silico* fragmentation tool MetFusion combines the output of spectral library search and *in silico* fragmentation predictions by MetFrag to improve candidate structure ranking, by taking into account the structural similarity of all *in silico* candidate structures to spectral library candidate structures. In NAP, the MetFusion spectral library matches are replaced by the annotations of all direct neighbors in the

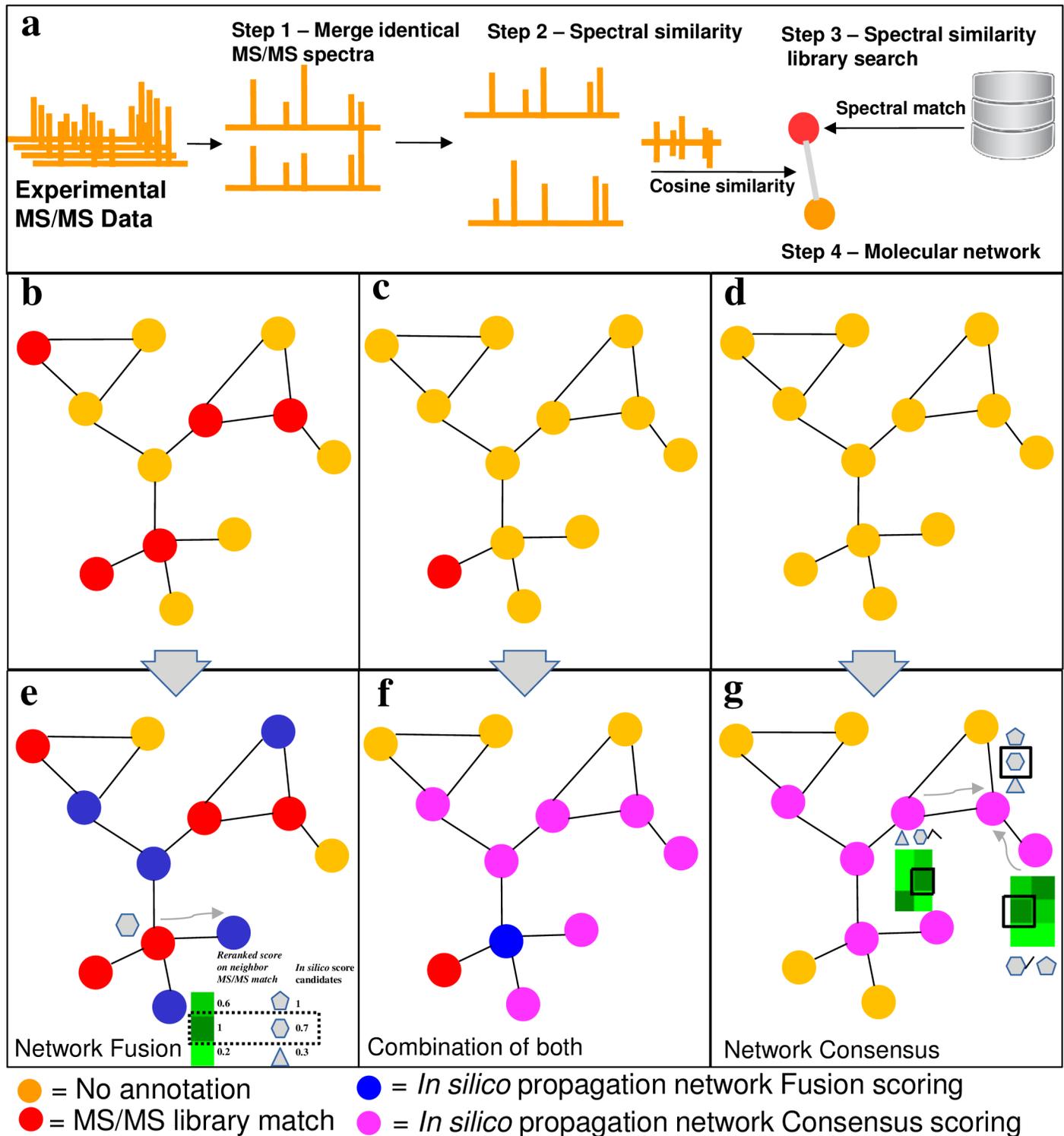


Fig 1. Representative scenarios of molecular networks obtained in an untargeted MS/MS experiment and possibilities for propagation. a) Introduction of molecular networking and library matching. b, c and d represent varying degree of spectral annotation in the network. e, f and g illustrate how the propagation of annotations can be used for each respective scenario (represented in the top panel). e) The *Fusion* scoring—The spectral library hit nodes (red) are employed to recalculate the score of candidate structures (grey shapes associated to nodes) for nodes having structure candidates from *in silico* fragmentation search (blue), based on their structural similarity (Represented by the green heatmaps, where darker green indicates a higher degree of similarity). f) and g) The *Consensus* scoring—a *Consensus* scoring can be used, based on the joint similarity of neighbors (pink nodes) for spectral library hits and *in silico* annotations (f), or *in silico* annotation only, when no library match is present (g).

<https://doi.org/10.1371/journal.pcbi.1006089.g001>

network. When there are no or very few spectral library matches (Fig 1C and 1D), a network consensus scoring will obtain the structural similarity from the candidate structures instead, exploiting the structural similarity of the *in silico* candidates (which we term *Consensus* scoring). This means that it is now possible to propagate annotations even when there are no spectral matches to reference MS/MS data.

After ranking improvements by considering neighboring nodes in the network, we show, using a spectral network of known spectra as benchmark data, that unsupervised cluster detection of candidate structures ranked using the network *Consensus* and *Fusion* scoring can find up to 81% of the correct compound substructures present in the first ranked candidate within a candidate structure list when reference libraries are considered while up to 63% correct compound substructures present in the first ranked candidate are found when no reference libraries matches are available within the molecular network. Additionally, because it takes significant computational resources for both data storage and to compute the propagations, NAP has been implemented as a ProteoSAFe workflow for High Performance Computing (HPC) onto the GNPS analysis infrastructure at the UCSD Center for Computational Mass Spectrometry.

Results

Construction of network Fusion and network Consensus in NAP

The resulting molecular clusters of molecular networking can be categorized into three different scenarios. First, clusters where we have a large number of matches to reference MS/MS spectra (Fig 1B), secondly, clusters with one or very few spectral matches by connected component (Fig 1C), and finally clusters with no spectral matches (Fig 1D). This range of scenarios requires different solutions. NAP was designed to handle all three scenarios. The starting point for NAP is the construction of a molecular network from data containing MS/MS spectra on the GNPS web-platform. Nodes in the network correspond to clusters of similar fragmentation spectra and edges represent spectral similarity between any two given nodes [5] (Fig 1A). Molecular networks allow the generation of hypothesis regarding the structural relationship of compounds connected in the network through annotation propagation. In parallel to molecular network construction, each node (a consensus spectrum) is subjected to spectral library search and *in silico* fragmentation search using MetFrag (Fig 1E–1G). The re-ranking of *in silico* candidate structures is calculated by the weighted contribution of MetFrag's score with a "spectral summary" (see Methods section), as previously described in MetFusion [28] and therefore we refer to this approach as network *Fusion*. We extended the *Fusion* principle to regions of the network where there are no library hits, where the *in silico* candidate structures of a given node can be re-ranked, by finding which candidate structure for this node maximizes the structural similarity to its neighbor node's candidate structures (Fig 1G). The *Fusion* scoring has a strict requirement that a library match has to be available as a direct neighbor (Fig 1B). However, in the case of sparse library matches in a molecular network or none at all (Fig 1C and 1D), we can apply network *Consensus* scoring (Fig 1G). The consensus scoring is built from the *n-first* neighbor candidates, where *n-first* is a user selected parameter. It is also possible to combine network *Fusion* scoring with network *Consensus* scoring through NAP (Fig 1F). In that scenario network *Consensus* scoring is calculated after network *Fusion* scoring and uses its re-ranked scores, instead of MetFrag's, to propagate the spectral annotations to more distant neighboring nodes.

Benchmarking NAP with a standard library

To benchmark NAP, we have created a molecular network with a subset of 5,467 MS/MS [M+H]⁺ spectra from NIST17 library that are structurally unique and have spectral similarity

(cosine score ≥ 0.6) to at least one spectrum in the subset (Fig A in S1 Text). A further subset (1,734 spectra), consisting of a network with only edges having a cosine score < 0.7 was selected in order to ensure that the annotation propagation will not be biased by structural identity [28], and the re-ranking can be performed with varying degree of structural similarity to neighbor nodes. This validation is important for analogs that share a substructure that is captured by spectral similarity, and helps to show that the propagation can be useful in that scenario (Figs B and C in S1 Text). NAP is built on top of *in silico* fragmentation performed with MetFrag, which searches for biologically relevant small molecule in a structure database including GNPS, HMDB (Human Metabolome Database) [29], SUPER NATURAL II [30], ChEBI (Chemical Entities of Biological Interest) [31] and DNP (Dictionary of Natural Products). In total this represents 367,204 unique small molecules (based on the first block of InChIKeys).

First, we created ranked candidate lists for each spectrum (node in the network) individually with MetFrag, which were used as the base ranking for propagation. Departing from candidate ranking at single spectrum level we wanted to assess the impact of network *Consensus* scoring on the results. In order to assess the influence of the *n-first* parameter on the network *Consensus* results, we applied the *Consensus* ranking method and varied the *n-first* parameter from 1 to 20 (Fig 2A and Fig D in S1 Text). Because network *Consensus* does not consider spectral library matches, the resulting ranking is based on the re-ranking of structures obtained with MetFrag based on the propagation of all direct neighbors' structural similarity. If neighbor nodes deriving from highly similar spectra not grouped in the networking process (Fig 1A—Step 1) are present, and structurally similar candidates are present in the candidate list, *Consensus* scoring can help to obtain structurally related candidates (Fig E in S1 Text). By adjusting the *n-first* parameter, it is possible to observe the relationship between the number of neighbor structures considered with the ranking position of the correct structure of the node being re-ranked (S1 Data).

The ranking of known molecular structure with MetFrag alone at single spectrum level had a mean ranking position of 14.7 with a median of 5 (Fig 2B). If we only consider the top 5 ranked candidates of all direct neighbors with *Consensus* (*n-first* 5), we observe an average ranking improved to 10.9 and median to 2. Further improvement was observed increasing the *n-first* parameter to 10, improving mean to 10.2 and median to 2. Similarly, for *n-first* parameters 15 and 20 we observed means of 9.8, 9.6 and median of 2 and 2, respectively (Fig 2A, Fig D in S1 and S2 Data). Thus, the *Consensus* scoring was able to annotate the correct structure in median on the top two candidates searching a reasonable large structure database derived from biological sources (367,204 molecules). Looking at Fig 2A one could conclude that a higher the *n-first* parameter always improves results, which is not necessarily the case, for example, inspecting unique candidates ranked in the first positions by different parameter numbers shows that *n-first* 1 had more unique correct compounds better ranked than *n-first* 5 to 20 (Fig F in S1 Text). The effect of the *n-first* parameter depends on the average number of candidates obtained as well as the number of connections the nodes have, as each node takes information from all directly connected nodes. Although we offer a default value of 10 *n-first*, that parameter has to be adjusted for each study, and ideally, manually curated.

Next we set out to assess the other approaches in NAP, the impact of network *Fusion* scoring, and from *Consensus* scoring (using as base ranking the *Fusion* scores) with *n-first* parameter fixed to 10. The network *Fusion* score had 29.0% increase in the first position ranking and overall 18.4% increase on rankings better than MetFrag alone for correct annotation among the top twenty candidates (Fig 2B and S3 Data and S4 Data). If *Consensus* scoring is being applied it means one or more neighbors possess *in silico* fragmentation candidates, and those are used to re-rank the candidate list of the node being processed, under the assumption that

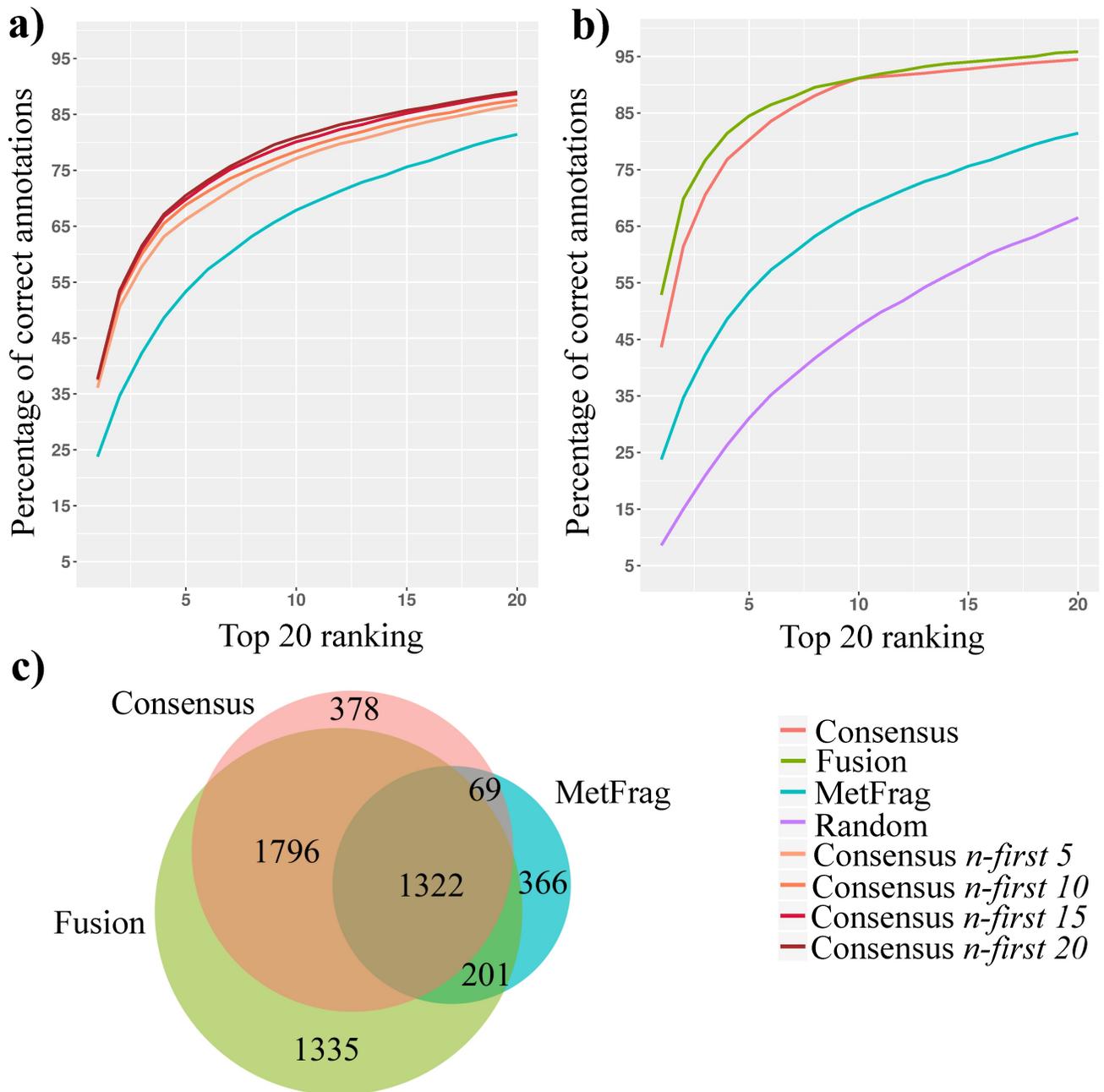


Fig 2. NAP re-ranking assessment using the 5,467 NIST17 [M+H]⁺ benchmark data set that have known nearest neighbors in the molecular network. a) The impact of setting the *n-first* parameter on percentage of correct annotations for network *Consensus* scoring. Where *n-first* indicates the *n* number of top ranked candidate structures (from 5 to 20) considered from the neighbor nodes during the *Consensus* scoring. b) Percentage of correct annotations ranking for each method. c) Number of spectra with improved ranking of correct annotations, that is, for which *Fusion* or *Consensus* scoring ranked the correct structure better than *MetFrag*.

<https://doi.org/10.1371/journal.pcbi.1006089.g002>

nodes connected by spectral similarity should possess structurally related first ranked candidates. For instances where the direct neighbor was previously re-ranked by *Fusion* scoring, the *Consensus* scoring can take advantage of this previous ranking, as the correct structure is more likely to be ranked among the top *n-first* candidates (Fig 1F). The network *Consensus* scoring had a 19.8% increase in the first position ranking and overall 13.0% increase on rankings better

than MetFrag alone for correct annotation among the top 20 candidates, when propagating is performed after *Fusion* scoring (Fig 2B and S3 Data and S4 Data).

The subset with edges of cosine score < 0.7 had very similar results to those described above for the complete network (S5 Data and Fig G in S1 Text), showing that the propagation is efficient under varying degree of spectral similarity. Overall, the average ranking of the correct annotation improved from 14.7 for MetFrag alone (and rank 23 for random assignment) to 4.7 for network *Fusion* and 6.3 for network *Consensus* (using previous *Fusion* ranking). The median improvement ranged from 5 for MetFrag alone (and rank 12 for random assignment) to 1 for network *Fusion* and 2 for network *Consensus* (Fig 2A and S4 Data and S5 Data). Both network *Fusion* and network *Consensus* scoring also have a larger number of unique (best ranking observed only for one approach—*Fusion*, *Consensus* or MetFrag) correct rankings when compared to MetFrag (Fig 2C). Interestingly, although 32.8% of improved annotations were found by network *Fusion* and network *Consensus* overlap, network *Consensus* found 6.9% annotations with better ranking than network *Fusion*, while network *Fusion* had 24.4% better annotations (Fig 2C). Those results suggest that, even in a network with nodes having spectral library matches, the use of neighbor candidate structures can provide complementary information for ranking.

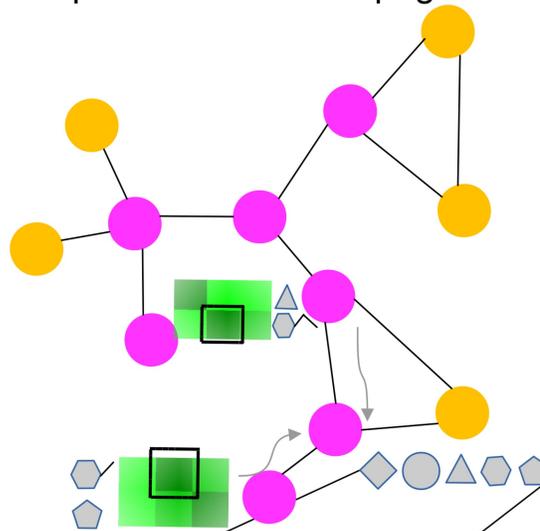
Clustering candidate structures

To further validate the approach we tested whether NAP annotations were correct at substructure level. In order to group candidate structures based on their structural similarity and dynamically assign groups inside candidate lists we used the Dynamic Tree Cut method [32] for dynamic branch cutting and unsupervised group detection on the result of hierarchical clustering (Fig 3). Grouping the structurally related candidates revealed that 84% of the structures assigned as first candidate by network *Fusion* scoring were contained in the same ClassyFire [33] chemical *Class* taxonomic classification of the known true structure (Correct Class), compared to 44% for MetFrag only (Fig 3, S6 Data). The ClassyFire classification provides a hierarchical classification that, similarly to unsupervised grouping, is based on structural similarity, however, specific features of the compound classes can be captured by nested classifications, from *Super Class* to *Class*, for example (See Methods section for user manual link). For network *Consensus* 78% were contained in the correct *Class* (66% considering *Consensus* without *Fusion*). When the structural similarity group is considered with multiple ClassyFire chemical classes, 84% of the structures assigned as first candidate with network *Fusion* and 76% for network *Consensus* scoring (67% considering *Consensus* without *Fusion*) were contained in the same structure similarity group of the known true structure (Correct substructure), compared to 42% for MetFrag only (Fig 3) suggesting that *in silico* methods can benefit from chemical classification as the best ranked structure is more likely to share a substructure with correct structure and belong to the same chemical class.

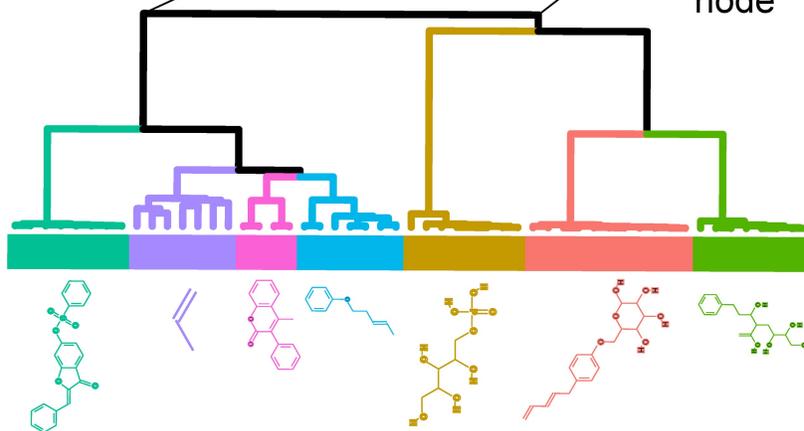
Single spectrum searches

We tested the NAP on its ability to annotate the challenge spectra from the Critical Assessment of Small Molecule Identification (CASMI) contest 2016 (<http://www.casmi-contest.org/2016/index.shtml>). The CASMI contest aims at benchmarking computational tools in untargeted mass spectrometry. The CASMI 2016 consisted of 146 spectra in positive ion mode, and 81 spectra in negative ion mode. The CASMI spectra were combined with all public spectral MS/MS libraries (after removal of CASMI library) to form one molecular network (one for each acquisition mode). The public MS/MS libraries were added as NAP needs a spectral network as input and will only be able to propagate an annotation if two or more similar spectra are

Step 1 - Anotation Propagation



Step 2 - Cluster candidate structures for every single node



Retrieve structures inside each group

Aurone flavonoids
 Isoflavonoids
 Organooxygen compounds
 Benzopyrans
Benzene and substituted derivatives
 Benzopyrans
 Benzopyrans

Retrieve classes inside same group

Step 3 - Apply clustering to whole network and compare known structure to group that contains the first rank from MetFrag, Fusion and Consensus scoring

Correct Class attribution

MetFrag - 44%
 Fusion - 82%
 Consensus - 78%

Correct Group attribution

MetFrag - 42%
 Fusion - 81%
 Consensus - 76%

Fig 3. Schematic representation of the molecular structure candidates clustering by structural similarity and dynamic cluster assignment from the 5467 NIST17 [M+H]⁺ unique compound spectra with ClassyFire chemical taxonomy. The structurally related group of candidate structures, detected by unsupervised clustering, containing the first candidates ranked by network *Fusion* and network *Consensus* often contains the maximum common substructure shared between *in silico* candidates (for candidates inside the group defined by clustering) and the known structure in the validation dataset (numbers shown in the bottom). The structurally related groups are highlighted by colors, inside each group we also show class assignments. The classes/structures that were predicted by unsupervised clustering were compared with the known compound.

<https://doi.org/10.1371/journal.pcbi.1006089.g003>

connected nodes in the network. After molecular networking, only the CASMI spectra and their connected nodes were retained. The filtering resulted in 136 positive and 50 negative spectra with at least one putative analog connected in the spectral library network. This means not all data from CASMI 2016 had a related structure in the public libraries, especially for negative mode that has fewer MS/MS references in the public domain. The resulting network contained 884 and 175 nodes for positive and negative modes respectively (Fig H in [S1 Text](#)). In the network, 33 positive and 4 negative spectra from CASMI were connected to at least one other challenge spectrum. The networks were composed of 107 and 44 molecular families (or connected components) for positive and negative nodes, respectively (Fig H in [S1 Text](#)). NAP was able to annotate 129 structures (6 structures had spectral analogs with names only and no structural information annotated with InChI or SMILES), 113 correct structures were ranked in first position, 10 on second, 4 in third, 1 in fourth and 1 in tenth place. Overall NAP, using reference library annotations (*Fusion* ranking), had 54 better rankings than MetFrag alone ([S7 Data](#)). When reference libraries were not directly used (*Consensus* ranking) NAP was able to annotate 130 structures (5 structures had spectral analogs with no structural information annotated), 70 correct structures in first position, 19 on second, 6 in third, 10 in fourth and 25 from fifth to tenth place. Overall network *Consensus* had 36 better rankings than MetFrag alone ([S7 Data](#)). For negative mode, NAP using *Fusion* scoring had one structure not annotated, 44 in first place, 4 in second and 1 in fifth place, overall 19 better rankings than MetFrag alone ([S8 Data](#)). For *Consensus* scoring in negative mode, we found one structure not annotated, 31 in first place, 7 in second and 11 from third to tenth place, overall 8 better rankings than MetFrag alone ([S8 Data](#)). Remarkably, even though we considered only 186 out of 227 challenge spectra, the strategy of network annotation propagation using MetFrag annotations had a comparable performance to the best performing employed in category 3 of the CASMI 2016 [8], highlighting that molecular networking and re-ranking based on structural information can complement such existing *in silico* fragmentation methods. The cumulated numbers (indicated below between parentheses) of correct annotation with the top rank in positive and negative ionization modes were, for NAP using network *Fusion* (157), for the network *Consensus* scoring (using previous *Fusion* ranking) (101), for MS-Finder [34] (159) and CFM-ID [157]. Note that NAP strategy with the spectral library network made from public spectral libraries can be computationally intensive, especially when large number of spectra are used. For that reason, we implemented NAP interface in the GNPS interface that uses high performance computing. Step by step instructions on how to use NAP is provided in the supporting information (Supplementary Tool Manual) and online (<https://gnps.ucsd.edu/ProteoSAFe/static/gnps-theoretical.jsp>).

Assessing the utility of NAP with metabolomics data sets

Above NAP has only been tested with known reference standards but not against typical data sets that are encountered in untargeted metabolomics experiments. We now set out to test NAP against previously published public fecal and plant metabolomics data sets from GNPS (MassIVE IDs: MSV000081120 and MSV000080502) [35–37] and a new fungal data set

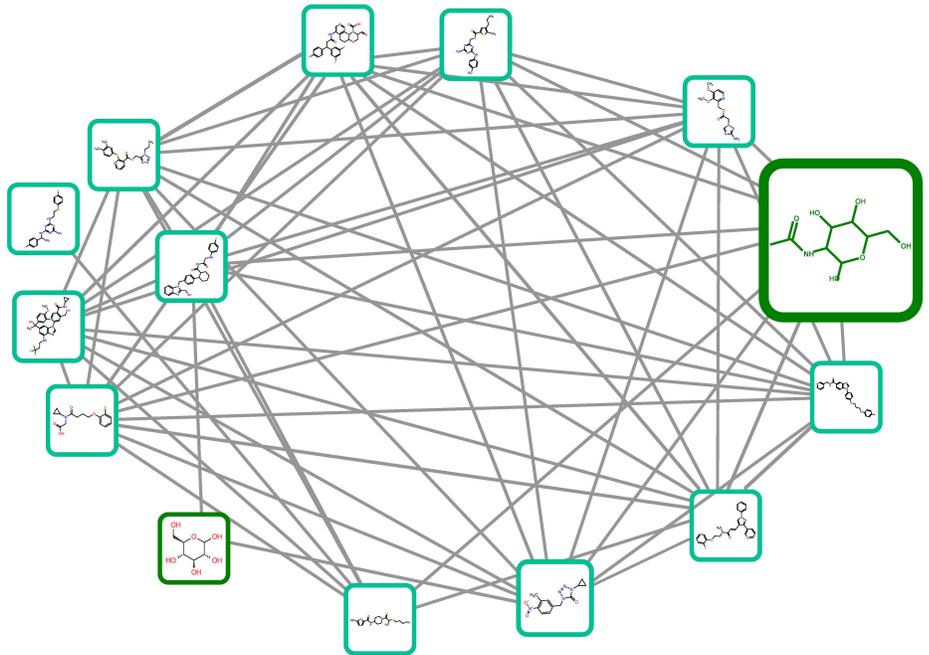
(MSV000081671). We chose these sets because authors of these projects had annotated them extensively. Fecal samples, analyzed with a Thermo Q Exactive instrument, represent complex mixtures full of small molecules from various backgrounds including endogenous compounds like amino acids, sugars, drugs and other xenobiotics, and food derived compounds. When analyzing molecular networks derived of such sample types, although the data sets have been inspected in great detail, there are still many molecular families [38, 39] that currently have few library matches or none at all.

We therefore set out to test NAP using the structural databases of natural products collection described above as well as PubChem. One molecular family that consisted of 14 nodes included two reference library matches: one to N-acetylgalactosamine and one to glucose from the GNPS-EMBL-MCF spectral library (Fig 4). Both library matches initially result in level 3 annotation according to the guidelines forwarded by the metabolomics society in 2007 [40]. Manual inspection revealed that the node with parent mass 222.110 is indeed consistent with N-acetylgalactosamine (Fig I in S1 Text) but N-acetylglucosamine is a likely candidate as well because they exhibit similar fragment losses (Fig I in S1 Text) consistent with the level 3 annotation. A comparison with standards would be required to achieve level 1 annotation [40]. Manual inspection of the data could not support the match to glucose (Fig J in S1 Text) and we would consider this to be a false annotation as i) with the search parameters used we do not have a match with FDR of 1% [41] and ii) glucosamine also results in mass fragments typical for sugars. In total, five propagated nodes presented N-acetylglucosamine containing structures and two additional structures had sugar and acetate structural features, albeit one structure had two acetates and the sugars are not in a cyclic configuration and are of different size.

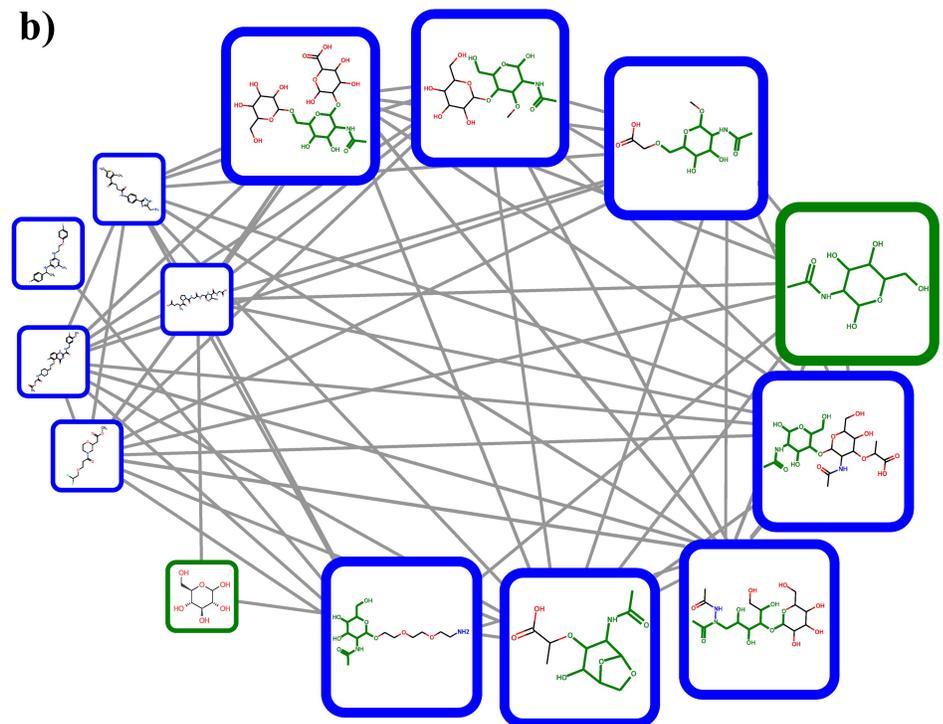
Upon matching the MS/MS spectrum of the node with precursor m/z of 294.118, each of the candidate hits matched by MzCloud (www.mzcloud.org) contained N-acetylglucosamine (or N-acetylgalactosamine) (Fig 4 and Fig L in S1 Text). Further inspection of the fragmentation of N-acetylglucosamine revealed that the specific N-acetyl containing fragments with m/z 96.0444 and 84.0444 (both $[M+H]^+$) are present across all members of the subnetwork. Some nodes did not return a N-acetylglucosamine containing structure within the top-10 candidates, it is likely that for those precursor masses no N-acetylglucosamine containing candidates are present in the reference structure library. Thus, the key take away from this NAP result is that this fecal data set contains a putative N-acetylated sugar family of molecule. Finally, it is worth noting that none of the top candidates found by MetFrag alone contained N-acetylglucosamine substructures, indicating how the propagation of a library match within a molecular network positively contributes to candidate ranking (Fig 4).

To further highlight the potential of NAP to aid structural annotation, we also processed a previously described data set [36] from extracts of the plant *Euphorbia dendroides*. In Fig 5, we illustrated how the network propagation from known molecular structures with available MS/MS spectra can improve the candidate ranking of the neighboring nodes. The reference MS/MS spectra are obtained from 23 compounds isolated from that extract and subsequently identified by NMR [36, 37], some of which were already available in the GNPS spectral library. The GNPS network revealed that 18 previously identified molecules in this dataset were annotated. For one molecular family we observed two spectral library matches, both belonging to the phorboid diterpene esters, while none of the other nodes in the network were annotated with spectral library search (Fig 5A). Among the candidates proposed by MetFrag (Fig 5B) with the structure bio-database described above, none of the top ranked were phorboids diterpene esters, while 11 out of 14 nodes were annotated as phorboid derivatives by network *Fusion* based re-ranking (Fig 5C). When *Consensus scoring* was used without taking in account the library matches, 8 out of 14 nodes were correctly annotated (Fig M in S1 Text).

a)



b)



 Spectral match on GNPS
 MetFrag top ranked candidates
 NAP Consensus top ranked candidates

Fig 4. N-acetyl-sugar metabolite family: Annotations of library match to N-acetylgalactosamine propagates through network. N-acetylglucosamine containing top-ranked candidates are represented in larger boxes. a) Result from MetFrag b) Result from network *Consensus*. Highlighted in green is the maximum common substructure (MCSS) of each node to the reference library (green border node) for the seven N-acetylglucosamine related molecules.

<https://doi.org/10.1371/journal.pcbi.1006089.g004>

The third data set was obtained from fungus gardens raised by the ant *Trachymyrmex septentrionalis*. A 37-nodes molecular family was selected as illustrative example of the network annotation propagation workflow for this data set (Fig 6). GNPS library matches suggested oxygenated steroid derivatives present in this cluster. Manual verification allowed confirmation of the annotation for ergosterol peroxide (m/z 429.336) with a parent accurate mass (error 0.4 ppm), while an ergosterol derivative was suggested for the m/z 415.357 node (Fig 6). We used NAP to search the DNP database for related structures. *Consensus* scoring provided additional structural annotation for 32 nodes related to ergosterol peroxides, while MetFrag provided top ranked annotations for only 18 spectra (S9 Data). Fig 6 shows example annotations for a cluster of nodes that reproduce the theoretical scenario illustrated in Fig 1F, where spectral library annotations can be propagated to direct neighbor nodes, that aid in the propagation to more distant nodes with *Consensus* scoring.

Discussion

One of the exciting developments in the annotation of highly complex samples derived from an organism or environment, is to take advantage of the relatedness of molecules co-occurring in the samples, which are often substrates and products of biochemical transformations. The benefits of using chemical relatedness to improve LC-MS/MS-based annotation using expected biotransformations was demonstrated by different approaches in the metabolomics field [17, 42–48]. Herein, we show that annotation propagation using molecular networking improves the annotations of neighboring nodes, harnessing the expected relationships of compounds detected in the sample. The process of propagating is often guided by the mass difference between the precursor ion masses, for example, a difference of 14.0157 Da may correspond to methyl functionalizations, but in many cases the fragmentation spectra alone is not sufficient to establish with confidence the position of the modification and type of modification, and only a partial annotation (isomers) or a molecular class annotation can be proposed. A 14.0157 could be the result of methylation, or substitution of fatty acids (propionate vs butyrate) or amino acids (e.g. Gly vs Ala) due to catalytic promiscuity in the biosynthesis of the molecule. The use of *in silico* fragmentation can provide further insights on structure annotation when used to initiate the manual propagation starting from spectral library annotation [26]. It is important to mention that the connectivity of the network can change if we change the networking parameters, especially the parameters ‘Min Pairs Cos’ (the Minimum cosine score that must occur between a pair of consensus MS/MS) and the ‘Minimum Matched Fragment Ions’. This is one of the reasons for which in the ‘Benchmarking NAP with a standard library’ section we split the dataset in lower spectral similarity only (0.6 to 0.7 cosine score range) and reanalyzed it, having comparable result from the complete dataset (0.6 to 0.9 cosine score range). It is also worth mentioning that improvements on data acquisition as well as preprocessing can drastically optimize the networking, with for example, less redundant nodes and higher quality fragmentation spectra, and consequently optimize results from NAP [24, 49, 50].

The results show that NAP improves the rank of correct candidate structure generated by an *in silico* tool. The performance of NAP *Fusion* and *Consensus* scores were evaluated with two reference standard datasets (NIST library and CASMI 2016 challenges), and NAP’s use

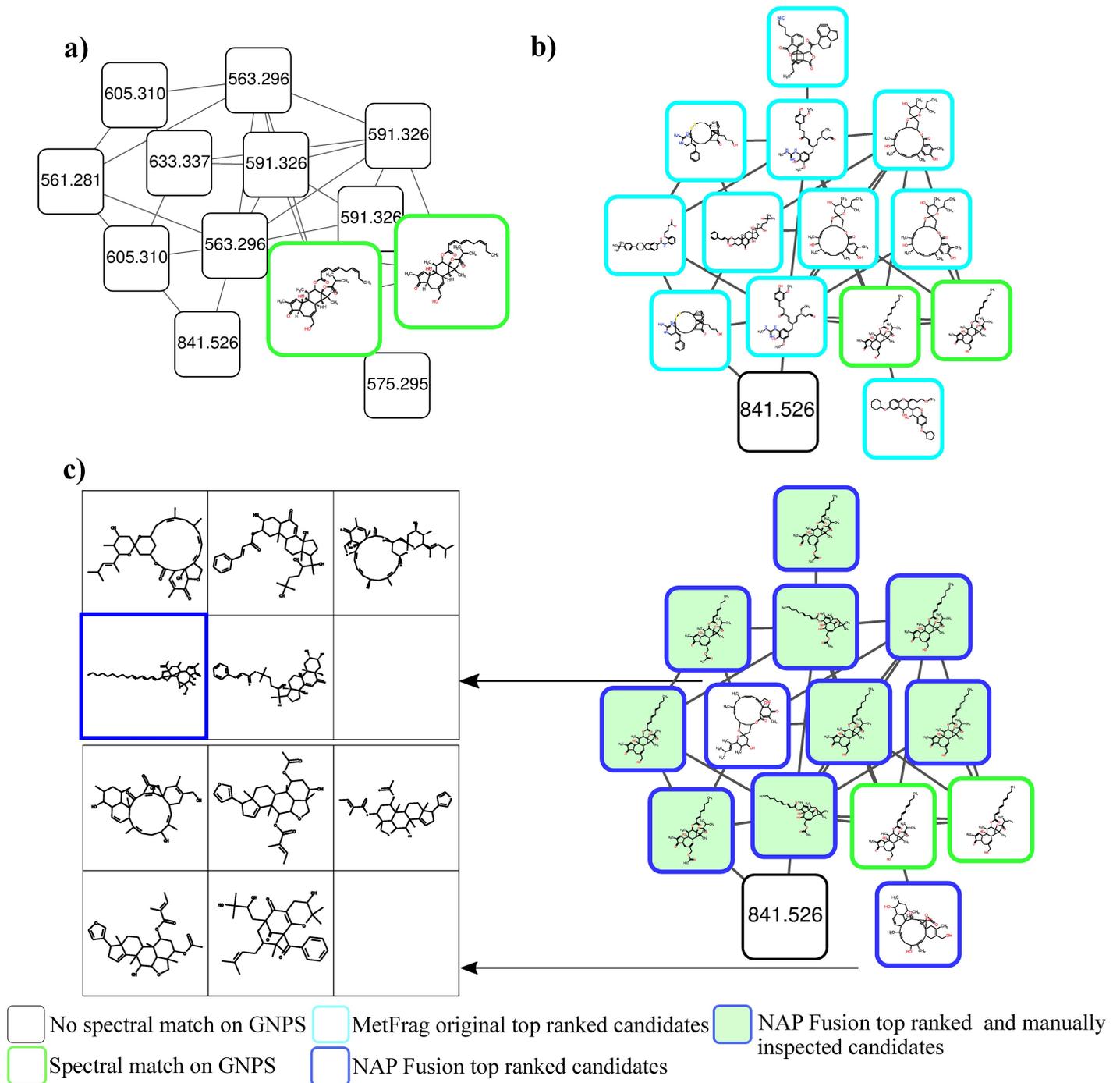


Fig 5. Network annotation of the *E. dendroides* plant extracts with NAP and visualized in Cytoscape with ChemViz2 plug-in. a) Result of the spectral library annotation using public spectral libraries available on GNPS. b) MetFrag annotation with top scoring molecules from bio-database. c) NAP annotation with top scoring matches using NAP network *Fusion* scoring, showing candidate lists associated to two nodes.

<https://doi.org/10.1371/journal.pcbi.1006089.g005>

was highlighted with three experimental datasets (fecal samples, plant extracts and fungus garden samples). Overall, both NAP scores improve the rank of the correct structure from MetFrag. Moreover, when considering the compound class with ClassyFire, we observed that NAP

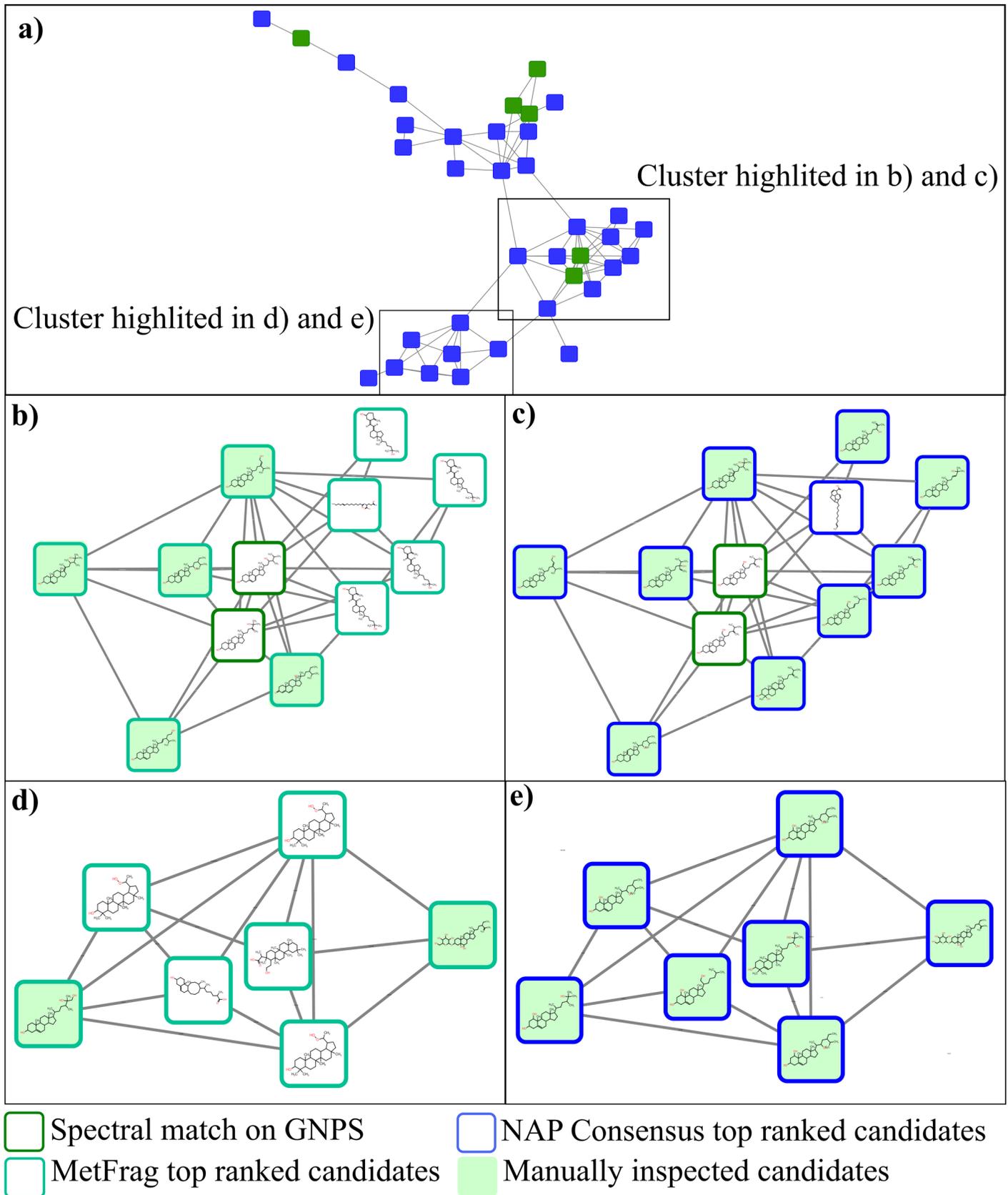


Fig 6. Network annotation of the *T. septentrionalis* fungus gardens extracts with NAP and visualized in Cytoscape with ChemViz2 plug-in. a) Group of nodes in which the annotation can be directly propagated from spectral library matches (top) and other in which the inspection of neighbor candidate structures can improve the annotation (bottom). Blue background nodes represent the presence of candidate structures from *in silico* annotation and green background represents candidates from spectral library annotation. b) MetFrag top ranked candidates. c) NAP consensus top ranked candidates. d) MetFrag top ranked candidates. e) NAP consensus top ranked candidates.

<https://doi.org/10.1371/journal.pcbi.1006089.g006>

improves the assignment of the correct compound structure in the first position from 29.0% with *Fusion* and 19.8% with *Consensus* to a better class assignment associated with the scoring to 38% with *Fusion* and 34% with *Consensus* (Fig 3), compared to *in silico* fragmentation performed alone at single spectrum level. With structure similarity clustering and class assignment one can begin to understand if there is an association between higher correct annotation rates for specific classes or structural motifs. This means that even when the actual structure may be incorrectly ranked, or absent from the database, the re-ranking annotated many structures with similar structural motifs instead. This is important as this would support a level 3 annotation and will allow end users to make informed decisions regarding the structural hypothesis of the molecules that could be detected by mass spectrometry. There is an association between lower annotation rates for specific classes or structural motifs, such as flavonoids, as there are often many isomeric structures possible for flavonoids and flavonoids do annotate within a structural family (Fig N in S1 Text and S10 Data).

One of the NAP's challenges relies in the selection of meaningful structure databases. Previous studies have shown that large databases such as PubChem contain a high number of synthetic molecules that have distinct molecular features from molecules typically produced by living organisms [51]. For that reason, *in silico* fragmentation methods have used dedicated databases to search candidates [52], or used methods that improve natural product likeness in candidate ranking [10]. Therefore, we have used the largest naturally occurring small molecule databases, and offer the possibility for the user to select between popular databases (GNPS, HMDB, SUPER NATURAL, ChEBI) or to upload their own custom database. The PubChem library is also available to users, as an alternative for instances where no candidate structure is found in smaller targeted structural libraries, such as DNP, MarinLit or AntiBase for example [53]. We also provide additional code to guide users on database formatting from a list of InChI or SMILES (See Methods section).

We have used the combinatorial fragmentation approach (MetFrag) because of its generality and usability reported in recent studies [12, 54], but the network propagation method can be extended to other classes of *in silico* fragmentation approaches. We expect that the use of NAP has the potential to improve the performance of other *in silico* fragmentation annotation approaches [9–11, 55]. Tools such as CFM-ID can predict a fragmentation spectrum from a structure, and offer the flexibility to create *in silico* spectral libraries. However, the ability to retrain the classification model and to regenerate new *in silico* spectral libraries is essential to keep pace with the growing number of public spectral libraries. Additionally, further developments in NAP should integrate LC-MS processing tools in order to annotate the adduct type and predict the molecular formula with confidence. The integration of those tools will limit the search of candidates in structure databases to only those having the likely molecular formula(s) [56] and/or substructures [17], which will improve the ranking of the *in silico* tools, and the performance of NAP annotation. Additionally, alternative ways to propagate information or select neighbor candidates should be tested to improve propagation results [57].

The fecal, plant and fungal data sets revealed many expected molecules, the acetylated-saccharide family, the phorboid ester family and the sterols family, respectively. Sterol derivatives are common fungal metabolites, and have been isolated from medicinal fungi [58] and soil fungi [59–64]. The bioactivity range for this class of compounds include antibiosis [61] and

anti-cancer, as it has been described for ergosterol peroxide [58]. Since sterols are fungal metabolites, their identification from this dataset is consistent. Now that the presence of ergosterol peroxide and oxygenated sterols has been observed in fungal gardens, and considering their potent biological activities [58, 61], their ecological role in the *T. septentrionalis* fungal gardens symbiotic system need to be elucidated.

We expect that network annotation propagations will increase with the deposit of new reference MS/MS spectra in searchable public spectral libraries and extensions of structural databases with potential candidate structures. *In silico* libraries have already been a part of annotations in Metlin, a metabolomics search engine, since 2005 [65], and more recently Metlin provides CFM-ID *in silico* predicted reference spectra as part of their search engine. LIPID MAPS [66] and LipidBlast [67] use predicted spectra for lipids and mzCloud uses *ab initio* predicted spectra. *In silico* propagated annotations are not yet a part of such search engines. When these become part of public reference libraries, it is critical that provenance of annotation is retained so that users can decide to rely on the annotation or not. Moreover, when associated to *in silico* fragmentation, annotation propagation has the potential to improve the structural hypothesis for many MS/MS annotations (both qualitative and quantitatively), especially if those are validated as *in silico* library entries by experts, and reused during future spectral library searches. To facilitate that process, NAP is integrated into the GNPS web-platform and the functionality allowing the user to add those expert curated putative annotations will become a part of GNPS public libraries, with provenance clearly indicated for the user. We anticipate that *in silico* network based propagation will be one key approach to fill in the dark matter of metabolomics annotations.

Methods

Experimental data generation

Trachymyrmex septentrionalis fungus garden samples were collected from across the Eastern USA. Authorization for collecting samples were previously obtained from the corresponding state department: State of New Jersey Department of Environmental Protection Division of Parks and Forestry State Park Service unnumbered Letter of Authorization; North Carolina Division of Parks and Recreation Scientific Research and Collecting Permit 2015_0030; Florida Department of Agriculture and Consumer Services unnumbered Letter of Authorization; Georgia Department of Natural Resources State Parks & Historic Sites Scientific Research and Collection Permit 032015; Department of Natural Resources Wildlife Resources Division unnumbered Letter of Authorization. Samples were extracted with 2:1 dichloromethane/methanol 3 times and dried under nitrogen. Samples were resuspended in 100% methanol containing 2 μ M sulfamethazine as internal standard and LC-MS/MS analysis was performed in an UltiMate 3000 UPLC system (Thermo Scientific) using a Kinetex 1.7 mm C18 reversed phase UHPLC column (50 X 2.1 mm) and Maxis Q-TOF mass spectrometer (Bruker Daltonics) equipped with ESI source. The column was equilibrated with 5% solvent B (LC-MS grade acetonitrile, 0.1% formic acid) for 1 min, followed by a linear gradient from 5% B to 100% B in 8 min, held at 100% B for 2 min. Then, 100%–5% B in 0.5 min and maintained at 5% B for 2.5 min at a flow rate of 0.5 mL/min throughout the run. MS spectra were acquired in positive ion mode in the range of 100–2000 m/z. A mixture of 10 mg/mL of each sulfamethazine, sulfamethizole, sulfachloropyridazine, sulfadimethoxine, amitriptyline, and coumarin-314 was run after every 24 injections for quality control. An external calibration with ESI-L Low Concentration Tuning Mix (Agilent technologies) was performed prior to data collection and internal calibrant Hexakis(1H,1H,3H-tertrafluoropropoxy)phosphazene was used throughout the runs. The capillary voltage of 4500 V, nebulizer gas pressure (nitrogen) of 2 bar, ion source

temperature of 200°C, dry gas flow of 9 L/min source temperature. Spectral rate of 3 Hz for MS1 and 10 Hz for MS/MS, total cycle time range of 0.83 sec consisted of one full MS scan and up to 5 MS/MS scans; MS/MS active exclusion parameter was enabled, set to 2 and to release after 30 s, precursor ion was reconsidered for MS/MS if current intensity/previous intensity ratio >2; CID energies for MS/MS data acquisition were used as in Table 1:

Basic stepping function was used to fragment ions at 50% and 125% of the CID calculated for each m/z from the above table with timing of 50% for each step. Similarly, basic stepping of collision RF of 550 and 800 Vpp with a timing of 50% for each step and transfer time stepping of 57 and 90 μs with a timing of 50% for each step was employed. The mass of internal calibrant was excluded from the MS/MS list using a mass range of m/z 921.5–923.5. The data were deposited in the online repository namely MassIVE (<ftp://massive.ucsd.edu/MSV000081671>).

Structure database construction

Structures were downloaded from their respective databases GNPS (<http://gnps.ucsd.edu/ProteoSAFe/gnpslibrary.jsp?library=all>), HMDB (<http://www.hmdb.ca/downloads>), SUPER NATURAL (http://bioinf-applied.charite.de/supernatural_new/) and ChEBI (<https://www.ebi.ac.uk/chebi/downloadsForward.do>). The Dictionary of Natural Products (DNP) structures were downloaded manually using the institutional subscription. All structures were classified by ClassyFire [33] taxonomy classification using an *in house* script available at https://github.com/DorresteinLaboratory/NAP_ProteoSAFe/. An initial user manual is also available for NAP parameter setting, including ClassyFire classification based database candidates' selection.

Network calculation and data availability

The NIST17 library .msp file containing 574,826 spectra from various instruments, acquisition modes and adduct types was parsed with an *in house* script to recover all [M+H]⁺ unique compound spectra. From 11,331 spectra recovered, a subset of 5,467 NIST17 [M+H]⁺ unique compound presented at least one analog in the networking conditions described below and were retained for validation. The raw data was retrieved from the public MassIVE datasets for *Euphorbia dendroides* (<ftp://massive.ucsd.edu/MSV000080502>), fecal (<ftp://massive.ucsd.edu/MSV000081120>) and a fungal data set (<ftp://massive.ucsd.edu/MSV000081671>). The CASMI data was downloaded from (<http://www.casmi-contest.org/2016/index.shtml>). The networks were calculated using GNPS web interface, and can be accessed with the following job IDs: for NIST library - <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=>

Table 1. CID energies for MS/MS data acquisition.

Type	Mass	Width	Collision	Charge State
Base	100.00	4.00	22.00	1
Base	100.00	4.00	18.00	2
Base	300.00	5.00	27.00	1
Base	300.00	5.00	22.00	2
Base	500.00	6.00	35.00	1
Base	500.00	6.00	30.00	2
Base	1000.00	8.00	45.00	1
Base	1000.00	8.00	35.00	2
Base	2000.00	10.00	50.00	1
Base	2000.00	10.00	50.00	2

<https://doi.org/10.1371/journal.pcbi.1006089.t001>

aa3386fd782e4875b6109fb32a93eb5a, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=229296bd89fc4eb19c4d4cb4e6d50744>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8ae2cbf410944c96a269e80670915851>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ee1233f263f94268ac62dc4cd358cd12>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=00851eb8fb2c4050b581ab898b9d228e>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1f905afc070241a58c74672b40333ac0>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e6663cf00af64a928def7a70108a2b19>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f2f4d89b7dbc4fc0ae78591b36f71585>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c26168effc244aedb95fba7de289aaf> <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7e45eb1bd3c34cc9a8a27bf20c182f4d>; NIST library subnetwork with cosine score < 0.7 <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8b0b5a467da9416b81ab4f925a4f4b43>; for Fecal, *Euphorbia dendroides* extracts and Fungal dataset <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f0cabc92247d44789900944a69874e8a>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ce2a564dbd704c0595494e04798b0233>, <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b753797b0dad4f1e84142dd59c84615b>; and finally for CASMI negative mode <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a3f02b1b648a43b6a210063a4ee2f787> and positive mode <http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=231902c6d75f41df8403e454c96e8d4a>. The parameters can be accessed by cloning the job or at the link “Networking Parameters and Written Network Description”. The corresponding NAP jobs can also be accessed through the web interface with the following job IDs: for NIST library - <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=29d517e67067476bae97a32f2d4977e0>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=d270e79876cb48deb6aab52a4fc647e>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=e2125577fe2646129becc248b96d42ba>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=81e01fe178d3424686079903d908b536>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=daa546b038604e5f83eaafb811bd0313>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=61c8a0d01309408f8ecceb5b31dab1a8>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=60fe9f77b3d04789997bf19aa1a0a828>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=53f8494ff9e8423697eebf4e98d287f0>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=c93a840100ec49bdbb3c12e5ed1e4790>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=5fd60b02f8ab4274bf45fd5b715b5e0b>; NIST library subnetwork with cosine score < 0.7 <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=74e04164a8374929a4548655742c0a4f>; for Fecal, *Euphorbia dendroides* extracts and fungal <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=9b6bddd2ba154b1397a53c7f7933430a>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=8ae3aa45bfe449d7969975189b14f429>, <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=fb02f64992bb4a0fb46b0e4832e69597>; for CASMI positive <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=7c5a16eba2eb42f88647e7d21e57f1bc> and for negative mode <http://proteomics2.ucsd.edu/ProteoSAFe/status.jsp?task=e9745e159f6f433d9efb71e9813df29a>. The parameters from NAP can be accessed by cloning the job through the web interface and [S11 Data](#).

Implementation and availability

The workflow was implemented using R and Python languages. The MetFrag 2.3 command line tool was downloaded from <http://c-ruttikies.github.io/MetFrag/>. The scoring *Fusion* used was $S_c = \alpha * f_c + (1 - \alpha) \sum_{j=1}^M sig(m_j * t_{cj})$ previously described in MetFusion as: where c represents each MetFrag candidate, f_c the MetFrag score and the ‘spectral summary’ represents the spectral library search cosine scores m_j for all neighbor nodes j , the chemical similarity t_{cj} between MetFrag candidate c and each neighbor node result j . The sig represents the sigmoid

function. We used as default the optimized parameters $\alpha = 0.3$, $\beta = -9$ and $\gamma = 0.6$ [28]. The *Consensus* scoring uses the same scoring function, but instead of using the structural similarity of the spectral library search from the neighbor it uses the maximum structural similarity of up to n first candidates of each neighbor node.

The molecular fingerprints and structural similarity were calculated with the fingerprint R package (<https://cran.rstudio.com/web/packages/fingerprint/index.html>), using CDK version 1.5.13. We have used the 'extended' fingerprint type and tanimoto similarity for *Fusion* and *Consensus* scoring. The same fingerprint type and the dissimilarity calculated as $1 - \text{tanimoto}$ similarity were used for clustering.

To group candidate structures we used the Dynamic Tree Cut method for dynamic branch cutting available in the dynamicTreeCut R package (<https://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/BranchCutting/>). Before clustering we create a low dimensional (up to 10 dimensions) projection with Multidimensional Scaling of tanimoto dissimilarity matrices, and then cluster the new coordinates using hierarchical clustering with euclidean distance and 'ward' grouping method. After the creation of a dendrogram, the same is subjected to automated group detection (dynamicTreeCut) using the following parameters: `minClusterSize = 2`, `method = "hybrid"`, `deepSplit = 2`.

The Workflow was implemented using ProteoSAFe (<https://bix-lab.ucsd.edu/display/PS/XML+Configuration+Overview>). The ProteoSAFe web interface can be found at <http://proteomics2.ucsd.edu/ProteoSAFe/>, under the workflow name NAP_CCMS and the code is available at github (https://github.com/DorresteinLaboratory/NAP_ProteoSAFe/).

Evaluation

To create the evaluation, results of each analysis from NAP containing the original MetFrag, *Fusion* and *Consensus* scores were compared. Ranking ties among ranking methods were considered overlap between the methods. When ties were found inside candidate lists, all candidates were considered to have the same position (e. g. 1, 2, 2, 2, 3, . . . instead of 1, 2, 2, 2, 5, . . .). The random candidate assignment was determined by sampling from the list of candidates from each spectrum following a uniform distribution.

Supporting information

S1 Text. Supplementary Figures for model validation and manual inspection.
(DOC)

S1 Data. Assessment of n first parameter impact on *Consensus* scoring.
(XLS)

S2 Data. Mean and median ranking position for MetFrag and subsequent *Consensus* re-ranking of candidate structures.
(XLS)

S3 Data. General overview of NIST benchmarking dataset annotation results.
(XLS)

S4 Data. Mean, median and first position ranking position for MetFrag *Fusion* and *Consensus* re-ranking of candidate structures.
(XLS)

S5 Data. Detailed information of NIST benchmark selection of spectra with cosine score < 0.7 in the network.
(XLS)

S6 Data. Detailed information of all NIST benchmark selection.

(XLS)

S7 Data. Detailed information of evaluated CASMI 2016 positive mode data.

(XLS)

S8 Data. Detailed information of evaluated CASMI 2016 negative mode data.

(XLS)

S9 Data. Manually inspected clustered spectra annotation for fungal garden dataset.

(XLS)

S10 Data. Class error association test for NIST benchmark dataset.

(XLS)

S11 Data. NAP parameters for each dataset described in the manuscript.

(XLS)

Acknowledgments

We thank Madeleine Ernst for test and feedback on the NAP web interface.

Author Contributions

Conceptualization: Ricardo R. da Silva, Norberto Peporine Lopes, Pieter C. Dorrestein.

Data curation: Louis-Félix Nothias, Justin J. J. van der Hooft, Andrés Mauricio Caraballo-Rodríguez.

Formal analysis: Ricardo R. da Silva.

Funding acquisition: Norberto Peporine Lopes, Pieter C. Dorrestein.

Project administration: Mingxun Wang, Norberto Peporine Lopes.

Resources: Evan Fox, Marcy J. Balunas, Jonathan L. Klassen.

Software: Ricardo R. da Silva, Mingxun Wang.

Supervision: Norberto Peporine Lopes, Pieter C. Dorrestein.

Validation: Louis-Félix Nothias, Justin J. J. van der Hooft, Andrés Mauricio Caraballo-Rodríguez.

Visualization: Louis-Félix Nothias, Justin J. J. van der Hooft, Andrés Mauricio Caraballo-Rodríguez.

Writing – original draft: Ricardo R. da Silva, Mingxun Wang, Louis-Félix Nothias, Justin J. J. van der Hooft, Andrés Mauricio Caraballo-Rodríguez, Marcy J. Balunas, Jonathan L. Klassen, Norberto Peporine Lopes, Pieter C. Dorrestein.

References

1. Khedr A, El-Hay SSA, Kammoun AK. Liquid chromatography-tandem mass spectrometric determination of propofol in rat serum and hair at attogram level after derivatization with 3-bromomethyl-propylpiperazine. *J Pharm Biomed Anal* [Internet]. 2017 Feb 5 [cited 2018 Jan 26]; 134:195–202. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27915197> <https://doi.org/10.1016/j.jpba.2016.11.051> PMID: 27915197

2. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* [Internet]. 1999 Jan 1 [cited 2012 Jul 10]; 27(1):29–34. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=148090&tool=pmcentrez&rendertype=abstract> PMID: 9847135
3. da Silva RR, Dorrestein PC, Quinn RA. Illuminating the dark matter in metabolomics. *Proc Natl Acad Sci* [Internet]. 2015 Oct 1 [cited 2015 Oct 2]; 201516878. Available from: <http://www.pnas.org/content/early/2015/09/30/1516878112.extract>
4. Dealing with the Unknown: Metabolomics and Metabolite Atlases. *J Am Soc Mass Spectrom* [Internet]. 2010 Sep 1 [cited 2017 Oct 31]; 21(9):1471–6. Available from: <http://www.sciencedirect.com/science/article/pii/S1044030510002692> <https://doi.org/10.1016/j.jasms.2010.04.003> PMID: 20452782
5. Wang M, Carver JJ, Phelan V V, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* [Internet]. 2016 Aug 9 [cited 2017 Mar 11]; 34(8):828–37. Available from: <http://www.nature.com/doifinder/10.1038/nbt.3597> <https://doi.org/10.1038/nbt.3597> PMID: 27504778
6. Demarque DP, Crotti AEM, Vessecchi R, Lopes JLC, Lopes NP. Fragmentation reactions using electrospray ionization mass spectrometry: an important tool for the structural elucidation and characterization of synthetic and natural products. *Nat Prod Rep* [Internet]. 2016 Mar [cited 2017 Nov 8]; 33(3):432–55. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26673733> <https://doi.org/10.1039/c5np00073d> PMID: 26673733
7. Hufsky F, Scheubert K, Böcker S. New kids on the block: novel informatics methods for natural product discovery. *Nat Prod Rep* [Internet]. 2014 Jun [cited 2017 Mar 12]; 31(6):807. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24752343> <https://doi.org/10.1039/c3np70101h> PMID: 24752343
8. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K, et al. Critical Assessment of Small Molecule Identification 2016: automated methods. *J Cheminform* [Internet]. 2017 Dec 27 [cited 2017 Apr 27]; 9(1):22. Available from: <http://jcheminf.springeropen.com/articles/10.1186/s13321-017-0207-1> <https://doi.org/10.1186/s13321-017-0207-1> PMID: 29086042
9. Allen F, Greiner R, Wishart D. Competitive fragmentation modeling of ESI-MS/MS spectra for putative metabolite identification. *Metabolomics* [Internet]. 2015 Feb 5 [cited 2017 Aug 24]; 11(1):98–110. Available from: <http://link.springer.com/10.1007/s11306-014-0676-4>
10. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc Natl Acad Sci U S A*. 2015;
11. Brouard C, Shen H, Dührkop K, d'Alché-Buc F, Böcker S, Rousu J. Fast metabolite identification with Input Output Kernel Regression. *Bioinformatics* [Internet]. 2016 Jun 15 [cited 2017 Apr 26]; 32(12):i28–36. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27307628> <https://doi.org/10.1093/bioinformatics/btw246> PMID: 27307628
12. Ruttkies C, Schymanski EL, Wolf S, Hollender J, Neumann S. MetFrag relaunched: incorporating strategies beyond *in silico* fragmentation. *J Cheminform* [Internet]. 2016 Jan 29 [cited 2016 Jan 30]; 8(1):3. Available from: <http://www.jcheminf.com/content/8/1/3/abstract>
13. Böcker S, Kind T, Torbašinović H, Obrenović S, Mehta SS, Tsugawa H, et al. Searching molecular structure databases using tandem MS data: are we there yet? *Curr Opin Chem Biol* [Internet]. 2017 Feb [cited 2017 May 28]; 36(3):1–6. Available from: <http://linkinghub.elsevier.com/retrieve/pii/S1367593116301922>
14. Steuer R, Kurths J, Fiehn O, Weckwerth W. Observing and interpreting correlations in metabolomic networks. *Bioinformatics* [Internet]. 2003 May 22 [cited 2017 Mar 18]; 19(8):1019–26. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12761066> PMID: 12761066
15. Ridder L, Wagener M. SyGMA: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem* [Internet]. 2008 May 19 [cited 2017 Mar 14]; 3(5):821–32. Available from: <http://doi.wiley.com/10.1002/cmdc.200700312> <https://doi.org/10.1002/cmdc.200700312> PMID: 18311745
16. Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, et al. Predicting Network Activity from High Throughput Metabolomics. Ouzounis CA, editor. *PLoS Comput Biol* [Internet]. 2013 Jul 4 [cited 2013 Jul 5]; 9(7):e1003123. Available from: <http://dx.plos.org/10.1371/journal.pcbi.1003123> <https://doi.org/10.1371/journal.pcbi.1003123> PMID: 23861661
17. van der Hoof JJJ, Wandy J, Barrett MP, Burgess KE V, Rogers S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc Natl Acad Sci U S A* [Internet]. 2016 Nov 29 [cited 2017 Mar 12]; 113(48):13738–43. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27856765> <https://doi.org/10.1073/pnas.1608041113> PMID: 27856765
18. Aguilar-Mogas A, Sales-Pardo M, Navarro M, Tautenhahn R, Guimerà R, Yanes O. iMet: A computational tool for structural annotation of unknown metabolites from tandem mass spectra. 2016 Jul 14 [cited 2017 Feb 23]; Available from: <http://arxiv.org/abs/1607.04122>

19. Bouslimani A, Porto C, Rath CM, Wang M, Guo Y, Gonzalez A, et al. Molecular cartography of the human skin surface in 3D. *Proc Natl Acad Sci U S A* [Internet]. 2015 Apr 28 [cited 2015 Nov 13]; 112(17):E2120–9. Available from: <http://www.pnas.org/content/112/17/E2120> <https://doi.org/10.1073/pnas.1424409112> PMID: 25825778
20. Garg N, Wang M, Hyde E, da Silva RR, Melnik A V, Protsyuk I, et al. Three-Dimensional Microbiome and Metabolome Cartography of a Diseased Human Lung. *Cell Host Microbe* [Internet]. 2017; Available from: <https://www.sciencedirect.com/science/article/pii/S1931312817304055>
21. Olivon F, Allard P-M, Koval A, Righi D, Genta-Jouve G, Neyts J, et al. Bioactive Natural Products Prioritization Using Massive Multi-informational Molecular Networks. *ACS Chem Biol* [Internet]. 2017 Oct 20 [cited 2017 Oct 31]; 12(10):2644–51. Available from: <http://pubs.acs.org/doi/abs/10.1021/acscchembio.7b00413> <https://doi.org/10.1021/acscchembio.7b00413> PMID: 28829118
22. Chagas FO, Caraballo-Rodríguez AM, Dorrestein PC, Pupo MT. Expanding the Chemical Repertoire of the Endophyte *Streptomyces albospinus* RLe7 Reveals Amphotericin B as an Inducer of a Fungal Phenotype. *J Nat Prod* [Internet]. 2017 May 26 [cited 2017 Oct 31]; 80(5):1302–9. Available from: <http://pubs.acs.org/doi/abs/10.1021/acs.jnatprod.6b00870> <https://doi.org/10.1021/acs.jnatprod.6b00870> PMID: 28375005
23. Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, et al. Mass spectral molecular networking of living microbial colonies. *Proc Natl Acad Sci U S A* [Internet]. 2012 Jun 26 [cited 2013 Aug 6]; 109(26):E1743–52. Available from: <http://www.pnas.org/content/early/2012/05/08/1203689109> <https://doi.org/10.1073/pnas.1203689109> PMID: 22586093
24. Olivon F, Roussi F, Litaudon M, Touboul D. Optimized experimental workflow for tandem mass spectrometry molecular networking in metabolomics. *Anal Bioanal Chem* [Internet]. 2017 Sep [cited 2017 Nov 7]; 409(24):5767–78. Available from: <http://link.springer.com/10.1007/s00216-017-0523-3> <https://doi.org/10.1007/s00216-017-0523-3> PMID: 28762069
25. de Oliveira G, Carnevale Neto F, Demarque D, de Sousa Pereira-Junior J, Sampaio Peixoto Filho R, de Melo S, et al. Dereplication of Flavonoid Glycoconjugates from *Adenocalymma imperatoris-maximiliani* by Untargeted Tandem Mass Spectrometry-Based Molecular Networking. *Planta Med* [Internet]. 2016 Nov 2 [cited 2017 Nov 7]; 83(7):636–46. Available from: <http://www.thieme-connect.de/DOI/DOI?10.1055/s-0042-118712> <https://doi.org/10.1055/s-0042-118712> PMID: 27806406
26. 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* [Internet]. 2016 Feb 16 [cited 2016 Mar 2]; Available from: <http://www.ncbi.nlm.nih.gov/pubmed/26882108>
27. Wolf S, Schmidt S, Müller-Hannemann M, Neumann S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* [Internet]. 2010 Jan [cited 2013 Aug 8]; 11(1):148. Available from: <http://www.biomedcentral.com/1471-2105/11/148>
28. Gerlich M, Neumann S. MetFusion: integration of compound identification strategies. *J Mass Spectrom* [Internet]. 2013 Mar [cited 2014 Feb 4]; 48(3):291–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23494783> <https://doi.org/10.1002/jms.3123> PMID: 23494783
29. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, et al. HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res* [Internet]. 2013 Jan 1 [cited 2017 Mar 11]; 41(D1):D801–7. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks1065>
30. Banerjee P, Erehman J, Gohlke B-O, Wilhelm T, Preissner R, Dunkel M. Super Natural II—a database of natural products. *Nucleic Acids Res* [Internet]. 2015 Jan [cited 2017 Jul 26]; 43(Database issue):D935–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/25300487>
31. Hastings J, de Matos P, Dekker A, Ennis M, Harsha B, Kale N, et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* [Internet]. 2013 Jan [cited 2015 Jun 20]; 41(Database issue):D456–63. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3531142&tool=pmcentrez&rendertype=abstract> <https://doi.org/10.1093/nar/gks1146> PMID: 23180789
32. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* [Internet]. 2008 Mar 1 [cited 2013 Feb 10]; 24(5):719–20. Available from: <http://bioinformatics.oxfordjournals.org/cgi/content/long/24/5/719> <https://doi.org/10.1093/bioinformatics/btm563> PMID: 18024473
33. Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J Cheminform* [Internet]. 2016 Dec 4 [cited 2016 Nov 22]; 8(1):61. Available from: <http://jcheminf.springeropen.com/articles/10.1186/s13321-016-0174-y>
34. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, et al. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Anal Chem* [Internet]. 2016 Jul 15 [cited 2016 Jul 16]; Available from: <http://www.ncbi.nlm.nih.gov/pubmed/27419259>

35. van der Hooft JJJ, Wandy J, Young F, Padmanabhan S, Gerasimidis K, Burgess KE V., et al. Unsupervised Discovery and Comparison of Structural Families Across Multiple Samples in Untargeted Metabolomics. *Anal Chem* [Internet]. 2017 Jul 18 [cited 2017 Oct 31]; 89(14):7569–77. Available from: <http://pubs.acs.org/doi/abs/10.1021/acs.analchem.7b01391> <https://doi.org/10.1021/acs.analchem.7b01391> PMID: 28621528
36. Esposito M, Nim S, Nothias L-F, Gallard J-F, Rawal MK, Costa J, et al. Evaluation of Jatrophone Esters from *Euphorbia* spp. as Modulators of *Candida albicans* Multidrug Transporters. *J Nat Prod* [Internet]. 2017 Feb 24 [cited 2017 Oct 31]; 80(2):479–87. Available from: <http://pubs.acs.org/doi/abs/10.1021/acs.jnatprod.6b00990> <https://doi.org/10.1021/acs.jnatprod.6b00990> PMID: 28106996
37. Esposito M, Nothias L-F, Nedev H, Gallard J-F, Leyssen P, Retailleau P, et al. *Euphorbia dendroides* Latex as a Source of Jatrophone Esters: Isolation, Structural Analysis, Conformational Study, and Anti-CHIKV Activity. *J Nat Prod* [Internet]. 2016 Nov 23 [cited 2017 Oct 31]; 79(11):2873–82. Available from: <http://pubs.acs.org/doi/abs/10.1021/acs.jnatprod.6b00644> <https://doi.org/10.1021/acs.jnatprod.6b00644> PMID: 27786472
38. Nguyen DD, Wu C-H, Moree WJ, Lamsa A, Medema MH, Zhao X, et al. MS/MS networking guided analysis of molecule and gene cluster families. *Proc Natl Acad Sci U S A* [Internet]. 2013 Jul 9 [cited 2017 Jul 30]; 110(28):E2611–20. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23798442> <https://doi.org/10.1073/pnas.1303471110> PMID: 23798442
39. Melnik A V., da Silva RR, Hyde ER, Aksenov AA, Vargas F, Bouslimani A, et al. Coupling Targeted and Untargeted Mass Spectrometry for Metabolome-Microbiome-Wide Association Studies of Human Fecal Samples. *Anal Chem* [Internet]. 2017 Jul 18 [cited 2017 Jul 30]; 89(14):7549–59. Available from: <http://pubs.acs.org/doi/abs/10.1021/acs.analchem.7b01381> <https://doi.org/10.1021/acs.analchem.7b01381> PMID: 28628333
40. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, et al. Proposed minimum reporting standards for chemical analysis. *Metabolomics* [Internet]. 2007 Sep 12 [cited 2013 Mar 12]; 3(3):211–21. Available from: <http://link.springer.com/10.1007/s11306-007-0082-2> <https://doi.org/10.1007/s11306-007-0082-2> PMID: 24039616
41. Scheubert K, Hufsky F, Petras D, Wang M, Nothias L-F, Dührkop K, et al. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat Commun* [Internet]. 2017 Dec 14 [cited 2018 Jan 24]; 8(1):1494. Available from: <http://www.nature.com/articles/s41467-017-01318-5> <https://doi.org/10.1038/s41467-017-01318-5> PMID: 29133785
42. Breitling R, Ritchie S, Goodenowe D, Stewart ML, Barrett MP. Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics* [Internet]. 2006 Jul [cited 2010 Jun 30]; 2(3):155–64. Available from: <http://www.springerlink.com/content/x01808376167kk12> <https://doi.org/10.1007/s11306-006-0029-z> PMID: 24489532
43. Rogers S, Scheltema RA, Girolami M, Breitling R. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics* [Internet]. 2009 Feb; 25(4):512–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19095699> <https://doi.org/10.1093/bioinformatics/btn642> PMID: 19095699
44. Silva RR, Jourdan F, Salvanha DM, Letisse F, Jamin EL, Guidetti-Gonzalez S, et al. ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics* [Internet]. 2014 Feb 3 [cited 2014 Apr 7]; 30(9):1336–7. Available from: <http://bioinformatics.oxfordjournals.org/content/30/9/1336.abstract.html?etoc> <https://doi.org/10.1093/bioinformatics/btu019> PMID: 24443383
45. Daly R, Rogers S, Wandy J, Jankevics A, Burgess KE V., Breitling R. MetAssign: Probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics* [Internet]. 2014 Jun 9 [cited 2014 Jun 11]; btu370–. Available from: <http://bioinformatics.oxfordjournals.org/content/early/2014/06/09/bioinformatics.btu370.short?rss=1>
46. Jourdan F, Breitling R, Barrett MP, Gilbert D. MetaNetter: inference and visualization of high-resolution metabolomic networks. *Bioinformatics* [Internet]. 2008 Jan [cited 2010 Aug 23]; 24(1):143–5. Available from: <http://bioinformatics.oxfordjournals.org/cgi/content/abstract/24/1/143> <https://doi.org/10.1093/bioinformatics/btm536> PMID: 18003642
47. Gaquerel E, Kuhl C, Neumann S. Computational annotation of plant metabolomics profiles via a novel network-assisted approach. *Metabolomics* [Internet]. 2013 Feb 23 [cited 2013 Mar 4]; Available from: <http://link.springer.com/10.1007/s11306-013-0504-2>
48. Li D, Baldwin IT, Gaquerel E. Navigating natural variation in herbivory-induced secondary metabolism in coyote tobacco populations using MS/MS structural analysis. *Proc Natl Acad Sci U S A* [Internet]. 2015 Jul 13 [cited 2015 Jul 18]; 1503106112–. Available from: <http://www.pnas.org/content/early/2015/07/09/1503106112.abstract?sid=2cedf13b-76e6-4132-af96-4bdf1da1fe3d>
49. Olivon F, Grellet G, Roussi F, Litaudon M, Touboul D. MZmine 2 Data-Preprocessing To Enhance Molecular Networking Reliability. *Anal Chem* [Internet]. 2017 Aug 11 [cited 2017 Nov 17]; 89(15):7836–

40. Available from: <http://pubs.acs.org/doi/10.1021/acs.analchem.7b01563> <https://doi.org/10.1021/acs.analchem.7b01563> PMID: 28644610
50. Kind T, Tsugawa H, Cajka T, Ma Y, Lai Z, Mehta SS, et al. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom Rev* [Internet]. 2017 Apr [cited 2017 Apr 28]; Available from: <http://doi.wiley.com/10.1002/mas.21535>
51. Peironcely JE, Reijmers T, Coulier L, Bender A, Hankemeier T. Understanding and classifying metabolite space and metabolite-likeness. Willson RC, editor. *PLoS One* [Internet]. 2011 Jan [cited 2014 Apr 2]; 6(12):e28966. Available from: <http://dx.plos.org/10.1371/journal.pone.0028966> <https://doi.org/10.1371/journal.pone.0028966> PMID: 22194963
52. Wang Y, Kora G, Bowen BP, Pan C. MIDAS: A Database-Searching Algorithm for Metabolite Identification in Metabolomics. *Anal Chem* [Internet]. 2014 Aug 26 [cited 2014 Aug 28]; Available from: <http://dx.doi.org/10.1021/ac5014783>
53. Mohamed A, Nguyen CH, Mamitsuka H. Current status and prospects of computational resources for natural product dereplication: a review. *Brief Bioinform* [Internet]. 2015 Jul 7 [cited 2016 Mar 14]; 17(2):309–21. Available from: <http://bib.oxfordjournals.org/content/17/2/309.abstract?etoc> <https://doi.org/10.1093/bib/bbv042> PMID: 26153512
54. Blaženović I, Kind T, Torbašinović H, Obrenović S, Mehta SS, Tsugawa H, et al. Comprehensive comparison of *in silico* MS/MS fragmentation tools of the CASMI contest: database boosting is needed to achieve 93% accuracy. *J Cheminform* [Internet]. 2017 Dec 25 [cited 2017 May 28]; 9(1):32. Available from: <http://jcheminf.springeropen.com/articles/10.1186/s13321-017-0219-x>
55. Ridder L, van der Hooft JJ, Verhoeven S. Automatic Compound Annotation from Mass Spectrometry Data Using MAGMa. *Mass Spectrom* [Internet]. 2014 Jul 2 [cited 2015 Apr 30]; 3(Special_Issue_2):S0033–S0033. Available from: https://www.jstage.jst.go.jp/article/massspectrometry/3/Special_Issue_2/3_S0033/_article
56. Böcker S, Letzel MC, Lipták Z, Pervukhin A. SIRIUS: decomposing isotope patterns for metabolite identification. *Bioinformatics* [Internet]. 2009 Jan 15 [cited 2013 Aug 8]; 25(2):218–24. Available from: <http://bioinformatics.oxfordjournals.org/content/25/2/218.full> <https://doi.org/10.1093/bioinformatics/btn603> PMID: 19015140
57. Picart-Armada S, Thompson WK, Buil A, Perera-Lluna A. diffuStats: an R package to compute diffusion-based scores on biological networks. *Bioinformatics* [Internet]. 2017 Oct 5 [cited 2017 Dec 28]; Available from: <http://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btx632/4347202>
58. Wu Q-P, Xie Y-Z, Deng Z, Li X-M, Yang W, Jiao C-W, et al. Ergosterol Peroxide Isolated from *Ganoderma lucidum* Abolishes MicroRNA miR-378-Mediated Tumor Cells on Chemoresistance. Kalpana G V., editor. *PLoS One* [Internet]. 2012 Aug 30 [cited 2017 Nov 4]; 7(8):e44579. Available from: <http://dx.plos.org/10.1371/journal.pone.0044579> <https://doi.org/10.1371/journal.pone.0044579> PMID: 22952996
59. Bioactive azaphilones from the fungus *Penicillium multicolor* CM01. *Phytochem Lett* [Internet]. 2016 [cited 2017 Nov 8]; 16:56–60. Available from: <http://www.sciencedirect.com/science/article/pii/S1874390016300301>
60. Qiao M-F, Yi Y-W, Deng J. Steroids from an Endophytic Eurotium rubrum Strain. *Chem Nat Compd* [Internet]. 2017 Jul 20 [cited 2017 Nov 8]; 53(4):678–81. Available from: <http://link.springer.com/10.1007/s10600-017-2089-x>
61. Mitsuguchi H, Seshime Y, Fujii I, Shibuya M, Ebizuka Y, Kushiro T. Biosynthesis of Steroidal Antibiotic Fusidanes: Functional Analysis of Oxidosqualene Cyclase and Subsequent Tailoring Enzymes from *Aspergillus fumigatus*. *J Am Chem Soc* [Internet]. 2009 May 13 [cited 2017 Nov 8]; 131(18):6402–11. Available from: <http://pubs.acs.org/doi/abs/10.1021/ja8095976> <https://doi.org/10.1021/ja8095976> PMID: 19415934
62. Dupont S, Lemetals G, Ferreira T, Cayot P, Gervais P, Beney L. ERGOSTEROL BIOSYNTHESIS: A FUNGAL PATHWAY FOR LIFE ON LAND? *Evolution (N Y)* [Internet]. 2012 Sep [cited 2017 Nov 8]; 66(9):2961–8. Available from: <http://doi.wiley.com/10.1111/j.1558-5646.2012.01667.x>
63. Weete JD, Abril M, Blackwell M, Parks L, M Sancholle M. Phylogenetic Distribution of Fungal Sterols. Butler G, editor. *PLoS One* [Internet]. 2010 May 28 [cited 2017 Nov 8]; 5(5):e10899. Available from: <http://dx.plos.org/10.1371/journal.pone.0010899> <https://doi.org/10.1371/journal.pone.0010899> PMID: 20526375
64. Ondeyka JG, Jayasuriya H, Herath KB, Guan Z, Schulman M, Collado J, et al. Steroidal and Triterpenoidal Fungal Metabolites as Ligands of Liver X Receptors. *J Antibiot (Tokyo)* [Internet]. 2005 Sep [cited 2017 Nov 4]; 58(9):559–65. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16320760> <https://doi.org/10.1038/ja.2005.76> PMID: 16320760

65. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN: a metabolite mass spectral database. *Ther Drug Monit.* 2005; 27(6):747–51. PMID: [16404815](https://pubmed.ncbi.nlm.nih.gov/16404815/)
66. Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, et al. LMSD: LIPID MAPS structure database. *Nucleic Acids Res [Internet]*. 2007 Jan 3 [cited 2017 Nov 8]; 35(Database):D527–32. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkl838> <https://doi.org/10.1093/nar/gkl838> PMID: [17098933](https://pubmed.ncbi.nlm.nih.gov/17098933/)
67. Kind T, Liu K-H, Lee DY, DeFelice B, Meissen JK, Fiehn O. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat Methods [Internet]*. 2013 Aug [cited 2015 Mar 3]; 10(8):755–8. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3731409&tool=pmcentrez&rendertype=abstract> <https://doi.org/10.1038/nmeth.2551> PMID: [23817071](https://pubmed.ncbi.nlm.nih.gov/23817071/)