foodMASST: A Mass Spectrometry Search Tool for Foods and Beverages.

Pieter Dorrestein (pdorrestein@health.ucsd.edu)
UCSD

Kiana West
UCSD

Robin Schmid
UCSD

Julia Gauglitz
UCSD

Mingxun Wang
UCSD

Brief Communication

Keywords: Spectrometry, Foods, Beverages

Posted Date: September 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-778710/v1

Abstract

There is a growing interest in unraveling the chemical complexity of our diets. To help the scientific community gain insight into the molecules present in foods and beverages that we ingest, we created foodMASST, a search tool for MS/MS spectra (of both known and unknown molecules) against a growing metabolomics food and beverage reference database. We envision foodMASST will become valuable for nutrition research and dietary biomarker discovery.

Introduction

Society is becoming more conscious of the foods we consume and more interested in understanding the compositions of specific dietary components. This sentiment is driven by the increasing awareness that diet influences our general health, including our microbiome, immune homeostasis, and even cognitive function. At the same time, our capacity to leverage untargeted metabolomics data, where many unidentified mass spectrometry signals are observed, has dramatically increased with the development of computational ecosystems such as GNPS/MassIVE[1]. There are now ~150-200 reported biomarkers of food intake[2], but it has still remained impossible to determine the uniqueness of these molecules across different foods and food groups.

The Mass Spectrometry Search Tool (MASST)[3] combined with a reference database of food metabolite data can be a powerful tool to understand the molecular landscape of foods. The MASST tool is a mass spectrometry search engine that identifies all data files in the GNPS/MassIVE untargeted metabolomics repository that contain a spectral match to a query MS/MS spectrum. We created a domain-specific MASST, called foodMASST (https://masst.ucsd.edu/foodmasst), to enable reporting of the search results in the context of foods and beverages (**Figure 1a**). As of Feb 2021, ~3,500 untargeted metabolomics files from different foods/beverages collected as part of the Global FoodOmics Project[4] (GFOP) have been deposited in MassIVE, a public mass spectrometry repository. Each food sample includes a classification according to a customized food ontology with additional metadata. FoodMASST utilizes this reference dataset to determine the food/beverage items containing a query spectrum and to contextualize the molecule's presence across foods. To increase usability by others in the community, we created a web interface to launch searches of known and unknown molecules with user-defined parameters and report the food information associated with the fragmentation data (= MS/MS) matches.

To start a foodMASST job, the parent mass and the MS/MS spectrum details are entered into the web interface. With the cloud based platform GNPS, all results are sharable with provenance and tied to user accounts. Once a job is completed, the results can be navigated through the landing page which displays several links to reports (**Figure 1b**). To provide additional context automatic spectral library search is performed against more than 30 public spectral libraries in the GNPS/MassIVE ecosystem including GNPS contributed libraries, Human Metabolome Database[5], all three Massbanks[6–8], and many others (for a list see https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp) to determine if molecules were known. "Dataset matches" navigates to all datasets (and files within those datasets) containing matches to the query spectrum. Reports specific to foodMASST can be found under “Foodomics Specific Analysis”. For each category in the food ontology, the proportion of matches are reported in “View Foodomics Specific Molecules” and visualized in “View Interactive Tree”. Metadata associated with the matching foods is reported in “View Matched Files”. For example, when the MS/MS spectrum for domoic acid, a potent neurotoxin from dinoflagellate blooms in the ocean[9], was searched (see data availability for job link), the only two matches obtained were associated with seafood - freshly caught mackerel (**Figure 1c**).

We performed additional representative MS/MS searches for 6 known molecules and one unknown. The includes: biocides fenamidone, spirotetramat, and enilconazole; the plant pigment cyanidin; Vitamin B_5; the antibiotic tetracycline; and an unknown molecule with a precursor *m/z* of 457.257. Fenamidone is a fungicide with low use in the US[10] and accordingly was detected in few samples (mushroom, spinach, and lettuce; **Figure 2a**). For an interactive example of the results landing page see https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=16d14b8efd134fcabe227dd6377db1b9. The “View Interactive Tree” link displays a visual representation of the food matches organized according to the GFOP ontology. Spirotetramat is an insecticide that also has low use in the US[10] and the foods sampled. However, this biocide is mainly used on citrus fruits and grapes and was detected in grapes, oranges, and cherries (**Figure 2a**). Enilconazole is a fungicide mainly used on citrus fruits[10] and was detected in 31% of samples classified as citrus (**Figure 2a)**. Interestingly, enilconazole is also used as an antifungal in veterinary medicine and was the only of the three biocides searched that was detected in a non-plant (goat cheese) sample. A search for cyanidin (**Figure 2b**), a plant pigment with reddish-purple color[11], returned teas (47% prevalence) and fruits; the highest prevalence was observed in raspberries (86%), blackberries (75%), and strawberries (100%). Vitamin B_5, a ubiquitous metabolite, was detected in many samples, but had the highest prevalence in animal-based foods and fungi (**Figure 2c**). We also searched for an antibiotic known to be used in farmed animals. Tetracycline[12] (**Figure 2d**) was detected in beef (5%) and poultry (22%). Finally, we searched for an unknown compound (**Figure 2e**) that was detected in an Alzheimer’s clinical cohort. The unknown had the highest detection rate in rice (27%) and oat (17%) samples, enabling the formulation that it may be associated with dietary habits. Links to the foodMASST jobs described above can be found in the data availability statement.

There are some precautions one has to take to prevent the over-interpretation of the results. Additionally, there are limitations with the presented foodMASST approach that are not specific to foodMASST but rather general to MS/MS spectral matching based on untargeted metabolomics. For example, mass spectrometry can be collected in positive and negative ion mode. The reference data is currently limited to positive ion mode and thus molecules only ionizable in negative mode cannot be used, however this infrastructure can easily accommodate negative ion mode if the community chooses to provide such reference data. Another caveat is that two different molecules, especially structurally related isomers, can have nearly identical MS/MS spectra. Another common feature of mass spectrometry is that molecules may be ionized as different adducts (e.g., H^+, Na+, K^+, NH_4^+). It is common that in an untargeted metabolomics experiment to have multiple adducts for each molecule. We encourage searching all adducts that have MS/MS information as it is impossible to get informative MS/MS matching when there are 1, 2, or 3 fragment ions. Such searches provide too little structural information to be reliable[13] and therefore the use of low information MS/MS spectral entries is discouraged. In general, the more ions and tighter the mass tolerances used for the search the less likely spurious matches are obtained.

The user may also be interested in structural analogs of related molecules in different foods as they are likely to have similar biological activities. Distributions of analogs can be discovered by searching in analog mode and reporting the neighbors of the searched spectrum in a molecular network. Analog searches will also allow improved discovery of MS/MS matches collected on different instruments or with different instrument settings.

The GFOP reference dataset will continue to grow. The community can contribute to the database that foodMASST uses by depositing LC-MS/MS-based metabolomics data, with the food-specific metadata, into GNPS/MassIVE followed by correspondence with the authors who will inspect the contributed data and add it to the existing database. We anticipate foodMASST will provide valuable insight for unknown MS/MS signals relevant to clinical studies and known signals being considered as dietary biomarkers. More broadly, the enhancement of MASST for domain-specific reporting using well-curated reference datasets will undoubtedly prove useful for many research areas.

Methods

The existing functionality of MASST[3] was utilized to create a workflow wherein MS/MS matches identified within the GFOP food reference dataset are reported and contextualized. The GFOP reference dataset contains untargeted metabolomics data acquired from over 3,500 food and beverage samples encompassing both human and animal dietary components. Each sample is associated with metadata to describe its characteristics, source, and preparation. Samples were also organized according to a custom ontology which, at the highest level, distinguishes between plant- or animal-based foods, algae, fungi, supplements, minerals, and animal feeds. The ontology is stored and managed using WebProtégé (Stanford University, California, USA), and can be viewed at https://webprotege.stanford.edu/#projects/23f59b5a-4c29-41df-b2d8-a2ea7282d912/edit/Classes. Each sample was labeled using the most specific ontology term based on the metadata provided (e.g., one sample might be classified as a red cherry tomato while another might be broadly classified as tomato).

To calculate the proportion of matches at every level of the ontology, each sample inherited all parent terms of its terminal label (e.g., terminal label: red cherry tomato; parent labels: cherry tomato, tomato, berry, fleshy fruit, fruit, plant). For every ontology term, the number of samples with that label and a spectral match to the query was divided by the total number of samples with that label. The GFOP ontology was combined with the foodMASST results and visualized in a tree structure using the D3.js library and code adapted from Rob Schmuecker (https://bl.ocks.org/robschmuecker).

Declarations

Data and Code Availability: Raw data are available on MassIVE (https://massive.ucsd.edu/) under the accession MSV000084900. The open-source code is available on GitHub (https://github.com/mwang87/GNPS_MASST and https://github.com/CCMS-UCSD/GNPS_Workflows). foodMASST jobs provided as examples can be viewed at: Domoic acid https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d77bb006bcbf48f79d789e60c4238899; Spirotetramat https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2784cf02f6864321a24390d00f7e3263; Enilconazole https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f6759397ec7d4ca7995d239a1a57d2a5; Fenamidone https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=16d14b8efd134fcabe227dd6377db1b9; Cyanidin https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b7eb71d28f52438c87d969c3e9618b9e; Pantothenic acid https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=6936ab0b89ec4a7aaec79a3a935f9e08; Tetracycline https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e5021c2bd8cc4f41ae35da99d5bfe4f2; Unknown (m/z=457.257) https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=da0777bdc1eb43aeb60791faf7d36011.

Acknowledgments: This work was made possible with the support by the CCF foundation #675191, the U.S. National Institutes of Health 1U19AG063744, R01 GM107550, P41 GM103484 and R03 CA211211.

Author Contributions: PCD conceptualized the idea. JMG and KAW organized the data and metadata. KAW, JMG, RS, MW, PCD performed analysis. MW created the website. KAW, RS, MW developed the code. KAW and PCD wrote the manuscript.

Declaration of Interests: PCD is on the scientific advisory board for Sirenas and Cybele. PCD is a scientific advisor and co-founder of Enveda and Ometa LLC with approval by UC San Diego. MW is a consultant for Sirenas and Founder of Ometa LLC.

References

1. Wang, M. *et al.* Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).

2. Barabási, A.-L., Menichetti, G. & Loscalzo, J. The unmapped chemical complexity of our diet. *Nature Food* **1**, 33–37 (2019).

3. Wang, M. *et al.* Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).

4. Gauglitz, J. M. *et al.* Reference data based insights expand understanding of human metabolomes. *bioRxiv* 2020.07.08.194159 (2020) doi:10.1101/2020.07.08.194159.

5. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).

6. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).

7. *MoNA - MassBank of North America* https://mona.fiehnlab.ucdavis.edu/.

8. *MassBank Europe* https://massbank.eu/MassBank/.

9. Lefebvre, K. A. & Robertson, A. Domoic acid and human exposure risks: a review. *Toxicon* **56**, 218–230 (2010).

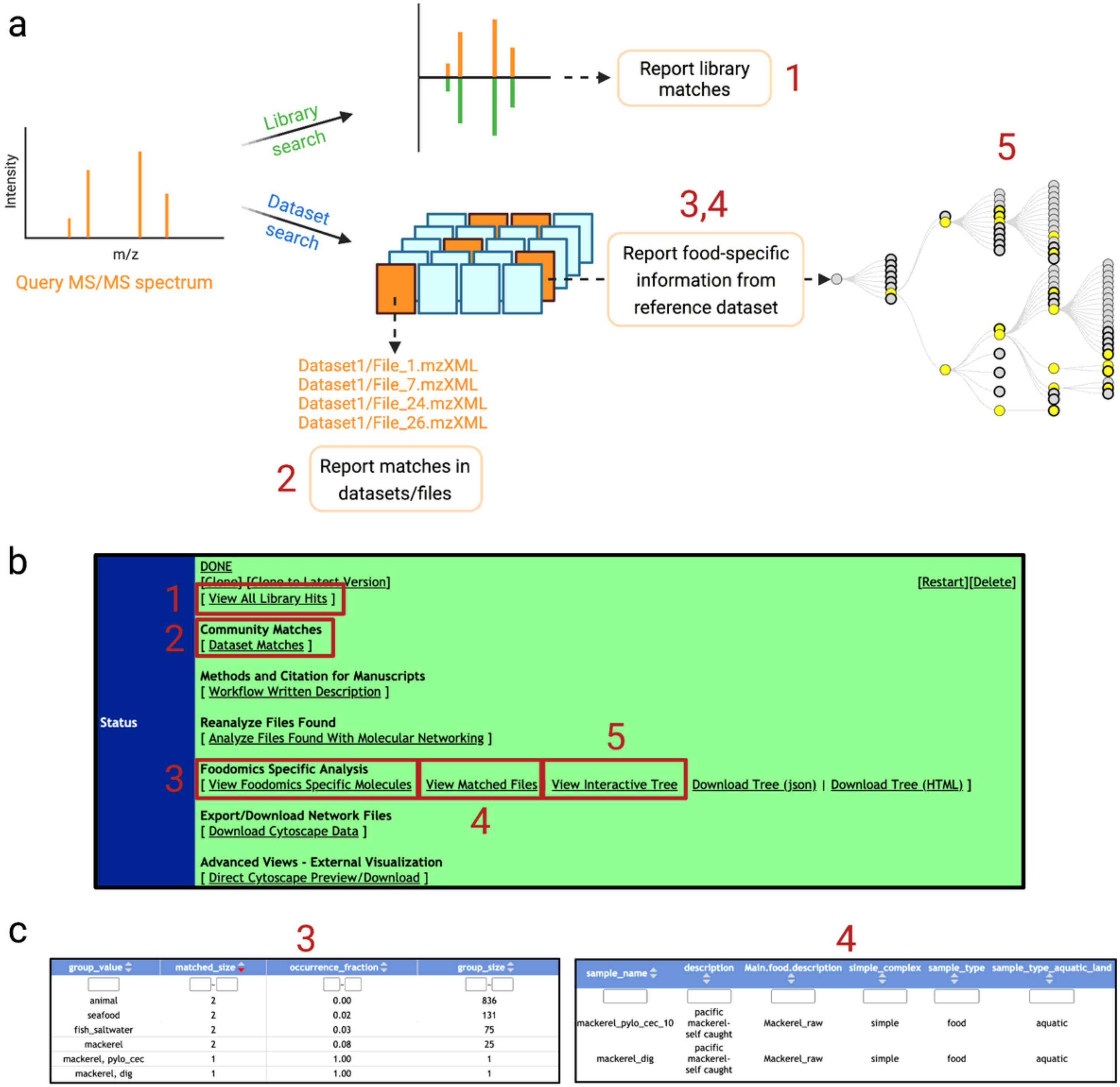
10. Pesticide National Synthesis Project. *U.S. Geological Survey* https://water.usgs.gov/nawqa/pnsp/usage/maps/.

11. Khoo, H. E., Azlan, A., Tang, S. T. & Lim, S. M. Anthocyanidins and anthocyanins: colored pigments as food, pharmaceutical ingredients, and the potential health benefits. *Food Nutr. Res.* **61**, 1361779 (2017).

12. Granados-Chinchilla, F. & Rodríguez, C. Tetracyclines in Food and Feedingstuffs: From Regulation to Analytical Methods, Bacterial Resistance, and Environmental and Health Implications. *J. Anal. Methods Chem.* **2017**, 1315497 (2017).

13. Scheubert, K. *et al.* Significance estimation for large scale metabolomics annotations by spectral matching. *Nat. Commun.* **8**, 1494 (2017).

Figures



Status

DONE
[Clone] [Clone to Latest Version]
[View All Library Hits]
[Restart][Delete]

Community Matches
[Dataset Matches]

Methods and Citation for Manuscripts
[Workflow Written Description]

Reanalyze Files Found
[Analyze Files Found With Molecular Networking]

Foodomics Specific Analysis
[View Foodomics Specific Molecules | View Matched Files | View Interactive Tree | Download Tree (json) | Download Tree (HTML)]

Export/Download Network Files
[Download Cytoscape Data]

Advanced Views - External Visualization
[Direct Cytoscape Preview/Download]

group_value	matched_size	occurrence_fraction	group_size
animal	2	0.00	836
seafood	2	0.02	131
fish_saltwater	2	0.03	75
mackerel	2	0.08	25
mackerel, pylo_cec	1	1.00	1
mackerel, dig	1	1.00	1

sample_name	description	Main.food.description	simple_complex	sample_type	sample_type_aquatic_land
mackerel_pylo_cec_10	pacific mackerel-self caught	Mackerel_raw	simple	food	aquatic
mackerel_dig	pacific mackerel-self caught	Mackerel_raw	simple	food	aquatic

Figure 1

foodMASST workflow and reports. a) A foodMASST query will search both library MS/MS spectra and metabolomics datasets that have been deposited in GNPS/MassIVE. Library spectra that match the query are reported (1) as potential annotations if the molecule is unknown. Public datasets, and files within datasets, containing spectral matches to the query are also reported (2) to provide context about where the molecule is observed. Additional reporting based on matches within the food reference dataset can be viewed and downloaded under “Foodomics Specific Analysis”. The percentage of samples containing a spectral match for each category of the food ontology are tabulated (3). Metadata for each food item with matches is reported (4). A visual summary of the food ontology and matches in each category can be viewed in the browser (5) or downloaded.

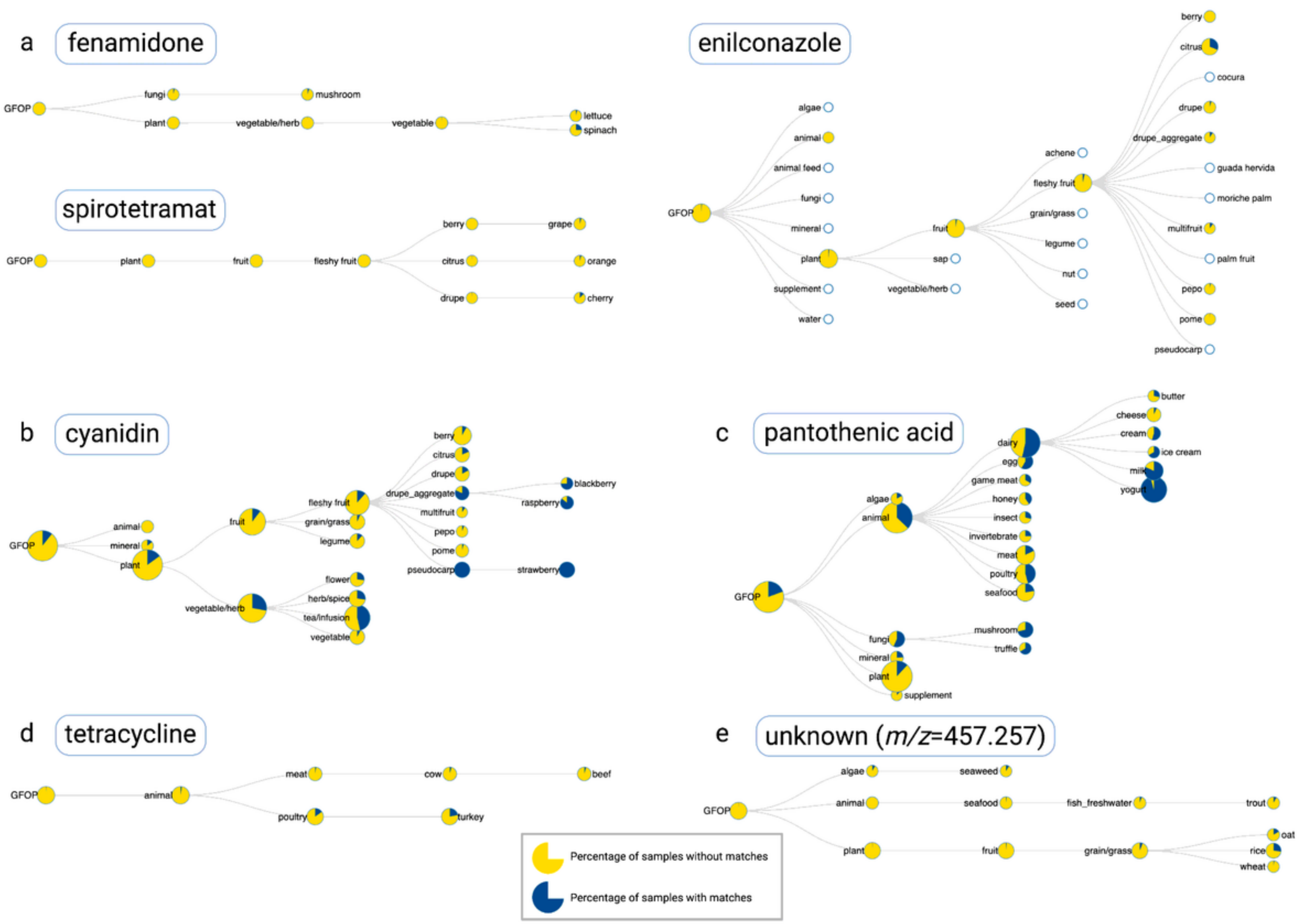


Figure 2

Examples of foodMASST results. a) Three biocides were queried against the Global FoodOmics reference dataset to determine their presence among sampled foods. b) Cyanidin, a plant pigment responsible for reddish-purple color, was observed in the expected categories such as teas, blackberries, raspberries and strawberries. c) Vitamin B5 was observed across many food categories, but was most prevalent in animal and fungi samples. d) Tetracycline, an antibiotic commonly administered to livestock, was detected only in beef and turkey samples. e) An unknown molecule detected in biospecimens from Alzheimer’s patients may be related to the consumption of oats or rice. Nodes are scaled according to the total number of samples classified for that ontology term or any of its descendants. Pies represent the percentages of samples with (blue) and without (yellow) matches to the query spectrum.