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 to assess the potential uniqueness of dietary biomarkers to represent specific foods or food classes.
If there are multiple spectral matches possible within the scoring criteria, all spectral annotations are provided. For example, an MS/MS spectrum may match to spectra with the annotation of L-leucine, D-leucine, L-isoleucine, D-isoleucine as MS/MS-based mass spectrometry usually does not provide regiochemical and stereochemical information.

We performed additional MS/MS searches for 6 library standards and one unknown compound. These included: biocides fenamidone, spirotetramat, and enilconazole; the plant pigment cyanidin; Vitamin B5; 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\(^1\) and accordingly was detected in few samples (mushroom, spinach, and lettuce; Fig. 2a). For an interactive example of the results landing page see https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=16d14b8ef2d134fcabe227ddd6377db1b9.

The "view library hits" will reveal if any of the ~620,000 public GNPS reference spectra with structural annotations match the input data. The mirror plot can be used to assess how well the query matched the reference spectrum. 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\(^1\) and the foods sampled. However, this biocide is mainly used on citrus fruits and grapes and was detected in grapes, oranges, and cherries (Fig. 2a). Enilconazole is a fungicide mainly used on citrus fruits\(^1\) and was detected in 31% of samples classified as citrus (Fig. 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 (Fig. 2b), a plant pigment with reddish-purple color\(^1\), returned teas (47% prevalence) and fruits;
the highest prevalence was observed in raspberries (86%), blackberries (75%), and strawberries (100%). Vitamin B5, a ubiquitous metabolite, was detected in many samples, but had the highest prevalence in animal-based foods and fungi (Fig. 2c). We also searched for an antibiotic known to be used in farmed animals. Tetracycline (Fig. 2d) was detected in beef (5%) and poultry (22%). Finally, we searched for an unknown compound (Fig. 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 must 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 modes. 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 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 and therefore the use of low information MS/MS spectral entries are 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

Fig. 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.
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**

**Reference data**
The existing functionality of MASST\(^1\) 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 3500 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 WebProtege (Stanford University, California, USA), and can be viewed at https://webprotege.stanford.edu/

**Reporting of spectral matches**
To calculate the proportion of matches at every level of the ontology, each sample inherited all parent terms of its terminal label (e.g., tomato; parent label: red cherry tomato while another might be broadly classified as a red cherry tomato while another might be broadly classified as tomato).

**DATA AVAILABILITY**
The raw data used in this study are publicly available online at MassIVE (https://massive.ucsd.edu) under the accession MSV00084900. foodMASST jobs provided as examples can be viewed at: Domoic acid https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d77b006cbcf84879d789e60c4238899; Spiroretromat https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2784e2f286e432a2a43990d087fe3263; Enlichonazole https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f6579397e7d4ca7995d239a1a57d2a5; Farnamidone https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=1e1d48b4ef134fcabe227d637db1b9; Cyanidin https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b7eb71d28f23c48379d69e3e9618b9e; Pantothenic acid https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=6936a0bb89ec4a7aae79a3939fe08; Tetracycline https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=50212cb28c4f41aa35da99d5bf4e2; Unknown (m/z 457.257) https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d0777bac1e43ab5079f1afdf3d6011.

**CODE AVAILABILITY**
The open-source code for foodMASST is available on GitHub (https://github.com/mwang87/GNPS_MASSST and https://github.com/CCMS-UCSD/GNPS_Workflows).

Received: 5 August 2021; Accepted: 4 March 2022; Published online: 20 April 2022

**REFERENCES**

1. Wang, M. et al. Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).
2. Gauglitz, J. M. et al. Reference data based insights expand understanding of human metabolomes. *Preprint at bioRxiv* https://doi.org/10.1101/2020.07.08.194159 (2020).
3. Akimoto, N. et al. FlavonoidSearch: A system for comprehensive flavonoid annotation by mass spectrometry. *Sci. Rep.* **7**, 1–9 (2017).
4. Sakurai, N. et al. An application of a relational database system for high-throughput prediction of elemental compositions from accurate mass values. *Biometrika* **29**, 290–291 (2013).
5. Giacomoni, F. et al. Phytohub, an online platform to gather expert knowledge on polyphenols and other dietary phytochemicals. (2017) - Abstract. https://hal.archives-ouvertes.fr/hal-01697081/file/2017_Giacomoni_ICPH_Vu%40Rbec.pdf. Accessed April 6, 2022.
6. Afendi, F. M. et al. KNAPSACK family databases: Integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.* **53**, e1 (2012).
7. Wishart, D. S. et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
8. Horai, H. et al. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
9. MoNA – MassBank of North America. https://mona.fiehnlab.ucsd.edu/. Accessed April 6, 2022.
10. MassBank Europe. https://massbank.eu/MassBank/. Accessed April 6, 2022.
11. Lefebvre, K. A. & Robertson, A. Domoic acid and human exposure risks: A review. *Toxicon* **56**, 218–230 (2010).
12. MassBank Europe. https://massbank.eu/MassBank/. Accessed April 6, 2022.
13. Lefebvre, K. A. & Robertson, A. Domoic acid and human exposure risks: A review. *Toxicon* **56**, 218–230 (2010).
14. Pesticide National Synthesis Project. U.S. Geological Survey https://water.usgs.gov/ nawqa/pnsp/usage/maps/. Accessed April 6, 2022.
15. Chou, 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).
16. Granados-Chinchilla, F. & Rodriguez, 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).
17. Scheubert, K. et al. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat. Commun.* **8**, 1494 (2017).

**ACKNOWLEDGEMENTS**

This work was made possible with the support by the CCF foundation #675191, the U. S. National Institutes of Health R03 CA211211. Figures created with BioRender.com.

**AUTHOR CONTRIBUTIONS**
P.C.D. conceptualized the idea. J.M.G. and K.A.W. organized the data and metadata. P.C.D. conceptualized the idea. J.M.G. and K.A.W. organized the data and metadata. P.C.D. declares the following competing financial interests: on the scientific advisory board for Sirenas and Cybele; scientific advisor and co-founder of Enveda and Ometa LLC with approval by UC San Diego. MW declares the following competing financial interests: consultant for Sirenas and Founder of Ometa LLC. K.A.W., R.S., and J.M.G. declare no competing financial interests. All authors declare no competing non-financial interests.

**COMPETING INTERESTS**
P.C.D. declares the following competing financial interests: on the scientific advisory board for Sirenas and Cybele; scientific advisor and co-founder of Enveda and Ometa LLC with approval by UC San Diego. MW declares the following competing financial interests: consultant for Sirenas and Founder of Ometa LLC. K.A.W., R.S., and J.M.G. declare no competing financial interests. All authors declare no competing non-financial interests.

**ADDITIONAL INFORMATION**

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41538-022-00137-3.

**Correspondence** and requests for materials should be addressed to Pieter C. Dorrestein.

**Reprints and permission information** is available at http://www.nature.com/reprints

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
