Title: Bridging Complexity and Accessibility in Metabolomics with MetaboApps

Authors: Helena Mannochio-Russo[1,†], Wilhan D. Gonçalves Nunes[1,†], Haoqi Nina Zhao[1], Kine Eide Kvitne[1], Shipei Xing[1], Harsha Gouda[1], Julius Agongo[1], Ipsita Mohanty[1], Vincent Charron-Lamoureux[1], Prajit Rajkumar[1], Abzer K Pakkir Shah[2], Axel Walter[2,3,4], Rithi Krishnaraj[5], Yasin El Abiead[1], Patrick C. Ferreira[6], Simone Zuffa[1], Abubaker Patan[1], Andrés Mauricio Caraballo-Rodríguez[1], Wout Bittremieux[7], Daniel Petras[2,8], Mingxun Wang[5], Pieter C. Dorrestein[1,9,10,11]

Affiliations

[1] Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
[2] Functional Metabolomics Lab, CMFI Cluster of Excellence, University of Tuebingen, Tuebingen, Germany
[3] Department of Peptide-based Immunotherapy, Institute of Immunology, University and University Hospital Tübingen, Tübingen, Germany
[4] Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany
[5] Department of Computer Science and Engineering, University of California Riverside, Riverside, CA, USA
[6] Institute of Chemistry, University of Campinas, Campinas, SP, Brazil
[7] Department of Computer Science, University of Antwerp, 2020 Antwerpen, Belgium
[8] Department of Biochemistry, University of California Riverside, Riverside, CA, USA
[9] Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, 92093, USA
[10] Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA
[11] Department of Pharmacology, University of California San Diego, La Jolla, CA, 92093, USA

† Authors contributed equally

*Correspondence: pdorrestein@health.ucsd.edu and mingxun.wang@cs.ucr.edu

Disclosures: P.C.D. is an advisor and holds equity in Cybele, BileOmix, and Sirenas and is a scientific co-founder of, is an advisor to, and holds equity in Ometa, Enveda, and Arome with prior approval by the University of California, San Diego. P.C.D. also consulted for DSM animal health in 2023. M.W. is a co-founder of Ometa Labs LLC.

Main text:

Untargeted metabolomics is a powerful approach for exploring the chemical diversity and dynamics of biological systems. However, the types of questions that can be addressed depend not only on experimental design but also on the data processing and analysis workflows employed, many of which require advanced computational expertise. GNPS[1], now transitioning to its second major implementation (GNPS2), has evolved into an expandable platform that supports the integration of modular web applications designed to simplify and enhance downstream analysis. These apps, named MetaboApps, facilitate the post-processing of outputs of several GNPS workflows and help

make repository-scale metabolomics knowledge and other areas of metabolomics more accessible to a broader community.

Classical and feature-based molecular networking have become cornerstones of untargeted metabolomics, providing a framework for organizing and interpreting data[2,3]. These approaches produce structured data tables based on ion intensities, MS/MS spectral similarities, and annotations from spectral libraries. Within the GNPS/MassIVE ecosystem[1] (which also indexes Metabolights[4], Metabolomics Workbench[5], and more recently from NORMAN[6] and some Zenodo[7] deposits) each data entry is linked to a Mass Spectrometry Run Identifier (MRI) and/or Universal Spectrum Identifier (USI)[8,9], ensuring data provenance and enabling users to trace results back to the original raw data. In effect, molecular networking jobs transform complex mass spectrometry data into a format that is data science-ready. These tables are often visualized as molecular networks, offering a chemistry-centric view of the data.

However, to derive additional and deeper insights, these tables are frequently subjected to downstream analyses, whether to extract chemical patterns (e.g., using MassQL)[10], perform statistical analysis[11–14], or integrate the data into biological or multi-omics contexts[15–17]. While the developments of workflows within the GNPS environment (e.g., for molecular networks creation and compound annotation) have significantly improved accessibility within the metabolomics community, downstream analyses are still often carried out using scripting and custom in-house code. Thus, despite major advances in accessibility for core tasks, there is growing interest and demand for handling more complex tasks without the need to be an expert in high-level data science. These more advanced data analyses are oftentimes aligned with common challenges in metabolomics, such as the complexity of data manipulation, rapidly growing dataset sizes, and low annotation rate of detected features. To support this evolution, GNPS2 has been designed as an extensible platform, enabling the connection to MetaboApps: modular web applications tailored for diverse downstream analyses (**Figure 1a**). Each MetaboApp is designed with a specific purpose, facilitating new ways to build upon previous analyses, deepen result interpretation, and share findings, all without requiring a deep understanding of scripting or coding languages.

The creation of these web apps is enabled by Streamlit, a widely used open-source Python framework popular in the bioinformatics community[18]. Similar to the Streamlit-based web apps developed for OpenMS (primarily used by the proteomics community[19], but also includes UmetaFlow[20] for metabolite quantification and identification) this framework is now being leveraged to support complex data analyses by GNPS users across a wide range of disciplines, including environmental science, natural products research, exposomics, metabolomics, toxicology, pharmacometabolomics, drug discovery, agriculture, and other biotechnological and clinical applications. Streamlit enables integration with essential Python packages such as Pandas[21] for data manipulation and Plotly for interactive visualization, facilitating dynamic, flexible, and intuitive data exploration. This makes it especially valuable for users without a coding background, but also provides experienced programmers a platform for rapid, interactive data analysis and figure generation, all supported by detailed documentation. Through enabling MetaboApps, the GNPS2 ecosystem provides a template that simplifies the integration of Streamlit-based applications into its infrastructure and supports the inclusion of “Downstream analysis” buttons directly within

analysis jobs. This allows users to launch MetaboApps and generate results using the GNPS2 output tables or using the job task ID, a blockchain-based identifier that contains all provenance information and results from a previous GNPS2 job. Depending on the application, secondary or even tertiary analyses can be performed with a single click. This streamlined approach reduces the need to manually track or transfer data for follow-up analysis, guarantees consistency during multiprocessing steps, and is especially valuable as data volumes grow and cross-repository analyses become more common.

Although a diverse range of MetaboApps have been developed to address the growing demands for more accessible downstream analyses in metabolomics, this is only the beginning. While 11 applications are currently integrated into GNPS2 workflows, including both published tools and those still under development (**Supplementary Table S1**), and with an estimated 3,000 users per month, we anticipate the development of many more. Future MetaboApps may extend beyond direct metabolomics data analysis to support broader aspects of metabolomics research, such as structure enumeration tools or applications for interpreting biological assay results. The current set already spans multiple analytical needs, from pattern-based spectral queries to ontology and enrichment, statistical and multiomics analyses, and repository-scale investigations (**Figure 1b**). The PostMN MassQL app enables users to apply MassQL queries directly to spectral files from classical and feature-based molecular networking to detect specific fragmentation patterns within a study. Users can select pre-defined queries from the MassQL compendium[10] or input custom queries to target specific compound classes. Results can be visualized by highlighting only the nodes in molecular networks that match the query criteria. For example, applying queries designed for bile acids to a Feature-Based Molecular Networking job of 1,993 fecal samples from the American Gut Project[22] revealed MS/MS signatures of mono- to tetrahydroxylated bile acids, including compounds without library matches (**Figure 1c**). This is particularly valuable, as it allows researchers to infer the chemical class of a compound even in the absence of a direct spectral match. The Multi-step MassQL app builds on this by enabling sequential queries, which proved to be particularly useful for distinguishing closely related compounds like bile acid isomers[23]. While this implementation was specifically developed for bile acids, we envision that other classes of molecules could also benefit from this approach.

Another category of apps supports annotation enrichment and contextualization. The Reference Data-Driven app summarizes molecular networking output data using ontology annotations and associated metadata[24], generating figures that help users understand chemical distributions across samples or experimental groups. It currently enables tracking MS/MS data from food sources and is readily extensible to other types of metadata applications. Two more specialized applications of this approach are the Drug Readout and Food Readout apps. Drug Readout allows users to annotate molecular features with standardized pharmacologic vocabularies, enabling users to interpret potential drug exposures across various ontology levels[25]. The Food Readout app simplifies complex data tasks for diet readout using machine learning trained features and merges them with intensities. The CMMC Dashboard facilitates the visualization of enrichment analyses performed using the Collaborative Microbial Metabolite Center knowledgebase (CMMC-kb)[26], providing plots and interactive interfaces that integrate information on metabolite sources, microbial origins[27], biological activity information, and metadata

associations. These tools are particularly valuable for interpreting complex datasets in the context of biological or environmental exposure. To support deeper statistical analyses, the FBMN Stats app offers a wide range of downstream statistical tools tailored for FBMN outputs, including data normalization, PCA, ANOVA, and PERMANOVA, all with interactive and downloadable visualizations[13]. The Chemical Proportionality app identifies and prioritizes putative biochemical transformations, scoring metabolite pairs based on anti-correlated abundance profiles across conditions[28]. In addition, these MetaboApps can extend to additional functionalities and go beyond the metabolomics field. With CorrOmics, users can perform multi-omics integration, correlating microbial and metabolite features to uncover biologically meaningful associations and generate network files suitable for further exploration in more specialized software visualization like Cytoscape. In the context of repository-scale analyses, the Conjugated Metabolome Explorer facilitates the discovery of novel metabolite conjugates in public mass spectrometry repositories by leveraging reverse spectral matching, offering a hypothesis-generation tool for metabolite biotransformation studies. Meanwhile, the Reverse Metabolomics app allows users to visualize the distribution of compounds across body sites, organisms, or disease contexts using the MASST results[29], enabling researchers to derive biological insights from repository-scale outputs.

In summary, MetaboApps offers a modular solution to challenges in metabolomics data analysis, which is especially important as data volumes continue to grow and analyses span multiple repositories. By integrating intuitive web interfaces directly within the GNPS2 infrastructure, these tools empower users of all backgrounds to extract additional insights from complex datasets without the requirement of computational expertise. This growing ecosystem reflects a strong commitment to open science and accessibility, and as new needs arise, it provides a robust foundation for continued innovation.

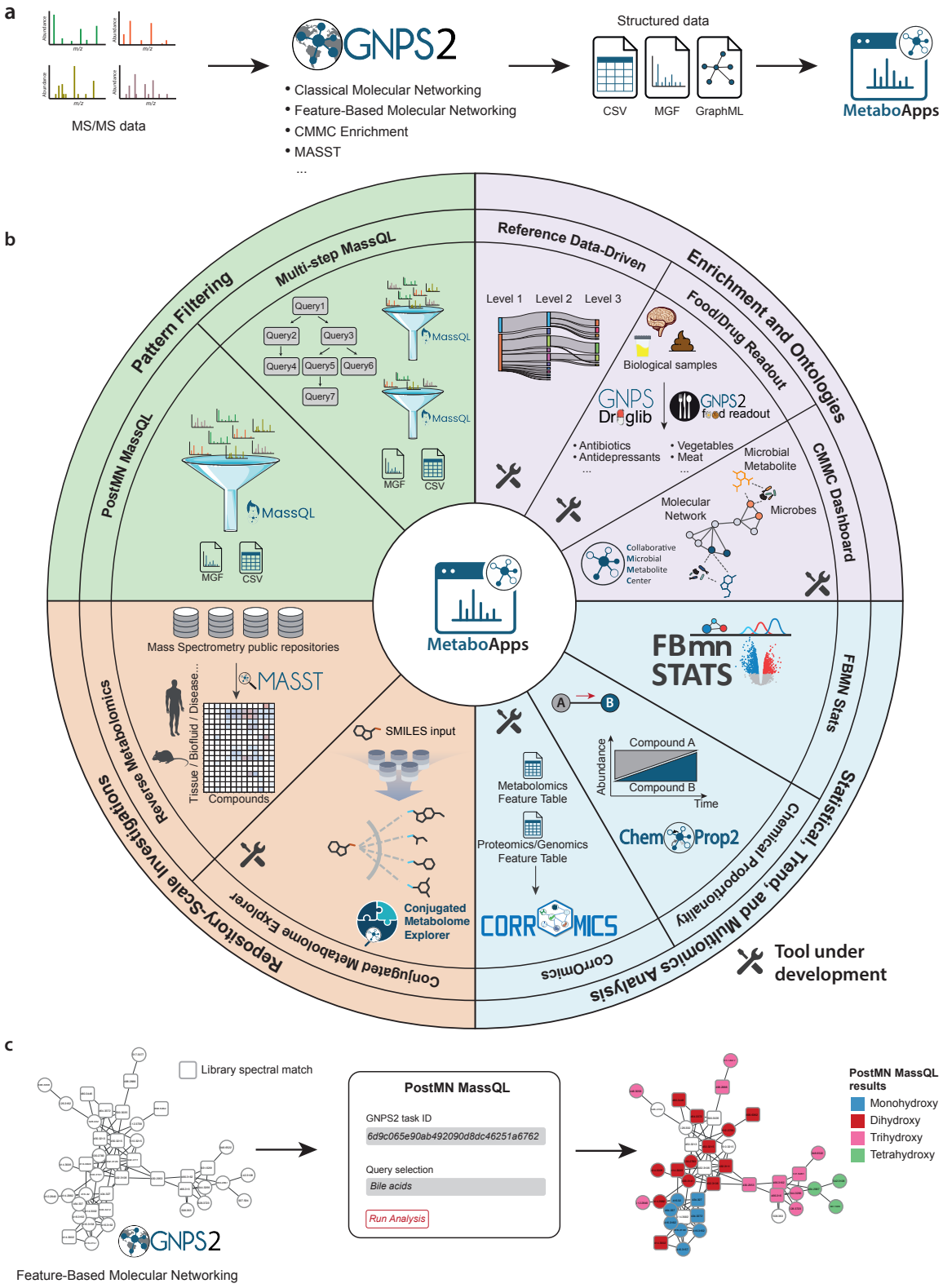

Figure 1. Overview of the MetaboApps integrated into GNPS2. a) Representation of the workflow: MS/MS data are analyzed through GNPS2 workflows, and the resulting structured data (e.g., CSV, MGF, GraphML) files can be directly fed into MetaboApps. **b)** Major categories of

MetaboApps, each addressing a specific type of downstream analysis. 1) Pattern filtering apps like PostMN MassQL and Multi-step MassQL enable spectral pattern-based queries using the MassQL language. 2) Enrichment and ontologies apps, such as Reference Data-Driven, Food/Drug Readout, and CMMC Dashboard, contextualize molecular networking results into biological metadata and reference ontologies. 3) Statistical, Trend, and Multiomics analyses tools like FBMN stats, Chemical Proportionality, and CorrOmics support statistical analyses and integration with other omics. 4) Repository-Scale Investigations applications, such as Reverse Metabolomics and the Conjugated Metabolome Explorer, allow exploration and hypothesis generation using mass spectrometry data from public repositories. **c)** Representative example of how MetaboApps can be used. In this case, the task ID of a Feature-Based Molecular Networking GNPS2 job of a public dataset (MSV000080673[22]) was analyzed using the PostMN MassQL app with a predefined set of bile acid queries[30]. The resulting output can be used to map the MS/MS spectra that contain diagnostic ions ranging from non-hydroxylated to penta-hydroxylated bile acids. Additional predefined MS/MS patterns from the MassQL compendium[10] or user-specific queries can also be used for customized analyses.

Code and web applications availability

The codes for all the MetaboApps described here are available in the following repositories:

- PostMN MassQL: https://github.com/helenamrusso/massql_post_MN
- Multi-step MassQL: https://github.com/wilhan-nunes/streamlit-Multi-Step-MassQL
- Reference Data-Driven: https://github.com/bittremieux-lab/gnps-rdd
- Drug Readout: https://github.com/wilhan-nunes/streamlit_drug_readouts
- Food Readout: https://github.com/wilhan-nunes/streamlit-food-readouts
- CMMC dashboard: https://github.com/wilhan-nunes/streamlit_CMMC_analysis-dashboard
- FBMNStats: https://github.com/Functional-Metabolomics-Lab/FBMN-STATS
- Chemical Proportionality: https://github.com/Functional-Metabolomics-Lab/ChemProp-Web-App
- CorrOmics: https://github.com/Functional-Metabolomics-Lab/Corromics
- Conjugated Metabolome Explorer: https://github.com/Philipbear/conjugated_metabolome
- Reverse Metabolomics: https://github.com/wilhan-nunes/streamlit-Reverse-Metabolomics

Documentations can be accessed at https://wang-bioinformatics-lab.github.io/GNPS2_Documentation/metaboapps_overview/

Acknowledgments

We thank the NIDDK (NIH) for supporting the Collaborative Microbial Metabolite Center (U24DK133658). W.B. acknowledges support by the Research Foundation – Flanders (FWO G0AGQ24N, G0AHY25N). D.P. was supported by the German Research Foundation (EXC 2124) and the Simons Foundation (SFI-LS-ECIAMEE-00013858). A.M.C.-R. and P.C.D. were supported by the Gordon and Betty Moore Foundation grant GBMF12120. YE acknowledges the Chen Zuckerberg Initiative (CZI) for funding. P.C.F. acknowledges the São Paulo Research Foundation (FAPESP) (grant 2024/17170-0) for funding.

Author contributions

H.M-R, W.D.G.N, H.N.Z, K.E.K., S.X., H.G., J.A., V.C-L., P.R., I.M., A.K.P.S., A.W., R.K., Y.E.A, W.B, D.P., M.W., P.C.D. developed the MetaboApps. P.C.F. created documentation. H.M.-R. and P.C.D. wrote the manuscript. All authors tested the web applications. All authors reviewed and edited the manuscript.

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. Nothias, L.-F. *et al.* Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).
3. Watrous, J. *et al.* Mass spectral molecular networking of living microbial colonies. *Proc. Natl. Acad. Sci. U. S. A.* **109**, E1743-52 (2012).
4. Yurekten, O. *et al.* MetaboLights: open data repository for metabolomics. *Nucleic Acids Res.* **52**, D640–D646 (2024).
5. Sud, M. *et al.* Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**, D463-70 (2016).
6. Alygizakis, N. A. *et al.* NORMAN digital sample freezing platform: A European virtual platform to exchange liquid chromatography high resolution-mass spectrometry data and screen suspects in “digitally frozen” environmental samples. *Trends Analyt. Chem.* **115**, 129–137 (2019).
7. Zenodo. https://zenodo.org.
8. Deutsch, E. W. *et al.* Universal Spectrum Identifier for mass spectra. *Nat. Methods* **18**, 768–770 (2021).
9. Bittremieux, W. *et al.* Universal MS/MS Visualization and Retrieval with the Metabolomics Spectrum Resolver Web Service. *bioRxiv* 2020.05.09.086066 (2020) doi:10.1101/2020.05.09.086066.
10. Damiani, T. *et al.* A universal language for finding mass spectrometry data patterns. *Nat. Methods* **22**, 1247–1254 (2025).
11. Bolyen, E. *et al.* Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* **37**, 852–857 (2019).
12. Gonzalez, A. *et al.* Qiita: rapid, web-enabled microbiome meta-analysis. *Nat. Methods* **15**, 796–798 (2018).
13. Pakkir Shah, A. K. *et al.* Statistical analysis of feature-based molecular networking results from non-targeted metabolomics data. *Nat. Protoc.* **20**, 92–162 (2025).
14. Pang, Z. *et al.* MetaboAnalyst 6.0: towards a unified platform for metabolomics data processing, analysis and interpretation. *Nucleic Acids Res.* **52**, W398–W406 (2024).
15. Morton, J. T. *et al.* Learning representations of microbe-metabolite interactions. *Nat. Methods* **16**, 1306–1314 (2019).
16. Singh, A. *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* **35**, 3055–3062 (2019).
17. Singh, K. S. *et al.* MEANtools: multi-omics integration towards metabolite anticipation and

biosynthetic pathway prediction. *bioRxiv* (2024) doi:10.1101/2024.12.22.629970.
18. Nantasenamat, C., Biswas, A., Nápoles-Duarte, J. M., Parker, M. I. & Dunbrack, R. L., Jr. Building bioinformatics web applications with Streamlit. in *Cheminformatics, QSAR and Machine Learning Applications for Novel Drug Development* 679–699 (Elsevier, 2023).
19. Müller, T. D. *et al.* OpenMS WebApps: Building user-friendly solutions for MS analysis. *J. Proteome Res.* **24**, 940–948 (2025).
20. Kontou, E. E. *et al.* UmetaFlow: an untargeted metabolomics workflow for high-throughput data processing and analysis. *J. Cheminform.* **15**, 52 (2023).
21. McKinney, W. Data Structures for Statistical Computing in Python. in *Proceedings of the Python in Science Conference* 56–61 (SciPy, 2010).
22. McDonald, D. *et al.* American gut: An open platform for citizen science microbiome research. *mSystems* **3**, (2018).
23. Mohanty, I. *et al.* MS/MS mass spectrometry filtering tree for bile acid isomer annotation. *bioRxiv* (2025) doi:10.1101/2025.03.04.641505.
24. Gauglitz, J. M. *et al.* Enhancing untargeted metabolomics using metadata-based source annotation. *Nat. Biotechnol.* **40**, 1774–1779 (2022).
25. Zhao, H. N. *et al.* Empirically establishing drug exposure records directly from untargeted metabolomics data. *bioRxiv* (2024) doi:10.1101/2024.10.07.617109.
26. Collaborative Microbial Metabolite Center Knowledgebase. https://cmmc-kb.gnps2.org/.
27. Zuffa, S. *et al.* microbeMASST: a taxonomically informed mass spectrometry search tool for microbial metabolomics data. *Nat Microbiol* **9**, 336–345 (2024).
28. Petras, D. *et al.* Chemical Proportionality within Molecular Networks. *Anal. Chem.* **93**, 12833–12839 (2021).
29. Wang, M. *et al.* Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).
30. Mohanty, I. *et al.* The underappreciated diversity of bile acid modifications. *Cell* **187**, 1801-1818.e20 (2024).