



Cite this: *Nat. Prod. Rep.*, 2021, 38, 301

The year 2020 in natural product bioinformatics: an overview of the latest tools and databases†

Marnix H. Medema 

Covering: 2020

Received 27th November 2020

DOI: 10.1039/d0np00090f

rsc.li/npr

Bioinformatic approaches to document and analyse chemical structures, biosynthetic gene clusters and analytical data play an important role in the study of natural products. Every year, such a large number of new algorithms, tools and databases are released, that it is difficult to keep track of all the latest developments. The aim of this short article is to provide a concise overview of and reference to the major tools, methods and databases that have been released in the past year.

Introduction

The study of natural products involves various different types of data, including structural, genomic, metabolomic and spectroscopic data. All these types of data require computational algorithms and resources to effectively process, analyze and contextualize them. The past decade has seen an acceleration in the development of new tools and databases that are relevant to natural product researchers. Here, we provide a concise overview of the latest tools and databases for the analysis of natural product chemical structures, the identification and annotation of biosynthetic gene clusters, and the analysis of natural product diversity in metabolomic datasets. The ESI† includes a table listing all tools and databases discussed here.

Chemical structure databases

Following the release of the Natural Product Atlas in late 2019,¹ several specialized databases for natural products from specific organisms or compound classes were released. These included a new version 3.0 of the Streptome-DB,² which included ~2500 new natural product structures from streptomycetes, as well as CyanoMetDB,³ a database covering ~2000 natural product structures from cyanobacteria. From a compound class-guided perspective, MacrolactoneDB⁴ appeared, which includes ~14 000 macrolactone structures and their bioactivity information. The NORINE database for nonribosomal peptides also saw a new release,⁵ which included integration of the recently published retrobiosynthetic algorithm rBAN⁶ to automatically identify the constituent monomers and other building blocks of these important natural products. While natural product

structures remain distributed across many different databases, the COLleCtion of Open NatUral producTs (COCONUT)⁷ combines structures from a wide range of open-access databases into a single resource.

Cheminformatic tools

To utilize and leverage such structural data, a number of relevant new cheminformatic tools have appeared. NPClassifier⁸ is a deep-learning-based algorithm that can help automatically classify sets of structures (*e.g.*, taken from a database or obtained from a set of library matches in a mass-spectrometric dataset) into classes and superclasses; thus, it can automatically identify whether molecules are, *e.g.*, terpenoids, polyketides or peptides. To map chemical space in more detail, and to identify structural similarities between molecules, molecular fingerprints are often used. In this area, two new fingerprint technologies, NC-MFP⁹ and MAP4,¹⁰ were presented that showed promising performance in explaining biological activities or differentiating closely related metabolites, respectively. Finally, to help seed compound structure databases, a new method, DECIMER,¹¹ was developed to recognise chemical structures from images in journal papers.

Identifying biosynthetic gene clusters

Genome mining is playing a more and more important role in natural product discovery. A range of well-known methods is available to identify biosynthetic gene clusters (BGCs) in genomes. Several of these were updated this year, such as PRISM4 (ref. 12) (see discussion under 'Predicting chemical structures'), as well as SEMPI version 2.0,¹³ which includes matching of predicted BGC products to natural product databases. Several new approaches were added to this set of tools: EvoMining¹⁴ is able to look for bacterial biosynthetic pathways

Bioinformatics Group, Wageningen University, Wageningen, The Netherlands. E-mail: marnix.medema@wur.nl

† Electronic supplementary information (ESI) available. See DOI: 10.1039/d0np00090f



that show no or only limited sequence similarity to known biosynthetic systems, by identifying paralogues of primary metabolic enzymes that have undergone accelerated evolution towards a secondary metabolic function. Aimed at fungi, CO-OCCUR¹⁵ provides a new way of identifying BGCs based on shared syntenic relationships between biosynthetic genes. Another fungal BGC identification tool, TOUCAN,¹⁶ was also released. A particularly challenging type of BGCs to computationally identify are those encoding the biosynthesis of Ribosomally synthesized and Posttranslationally modified Peptides (RiPPs), because of the apparent large diversity of unknown RiPP classes for which rule-based detection is not possible (as the required knowledge to design such rules is not yet available).¹⁷ Several new algorithms were released this year that utilize machine learning and pattern-recognition approaches to this end, including DeepRiPP,¹⁸ RRE-finder¹⁹ and decRiPPter,²⁰ on top of other approaches like RiPPER²¹ that had been published last year, some of these aiming at the identification of novel RiPP classes. Another tool to identify RiPP biosynthetic pathways, RODEO, was extended with capabilities to explicitly identify linaridins.²²

Charting biosynthetic gene cluster diversity

To be able to cope with datasets covering thousands or even hundreds of thousands of genomes, new algorithms were released to chart the diversity of BGCs in genomic data. Bi-GSCAPE and CORASON²³ enable automated sequence similarity networking and reconstruction of BGC phylogenies to facilitate the exploration of thousands of BGCs from diverse organisms. More recently, BiG-SLICE²⁴ was released, which scales up this principle by allowing the grouping of millions of BGCs into gene cluster families; the BiG-FAM database²⁵ makes these gene cluster families easily searchable for the scientific communities, and allows assignment of BGCs to such families directly

from antiSMASH results. The new cblaster tool²⁶ provides a quick way to perform similarity searches of BGCs by remote querying the NCBI web services, and to enable visual gene cluster comparisons between selected BGCs, the related clinker²⁷ tool provides a highly user-friendly method.

Biosynthetic gene cluster databases

Several databases of biosynthetic gene clusters were also updated or released this year. The MIBiG repository for experimentally characterized biosynthetic gene clusters saw a second release,²⁸ in which 851 new BGCs were added and the database was made searchable online. Two databases for computationally predicted BGCs, antiSMASH-DB²⁹ and IMG-ABC,³⁰ were also updated with new features, including extension with fungal data and fully refreshed contents, respectively. A new atlas of fungal BGCs from ~1000 fungal genomes, called Prospect, was also released, which includes gene cluster family assignments for these gene clusters.³¹ Finally, databases with curated sets of high-quality genomes, such as the ActDES database for actinomyces³² released this year, will make it easier to navigate high-quality data when navigating biosynthetic potential of various taxa.

Target-based genome mining

Finding the needle in the haystack within these giant datasets is not trivial. Target-based genome mining approaches make it possible to identify BGCs encoding the production of natural products with a biological activity of interest, such as antibiotics. An updated version of the ARTS pipeline³³ now enables identification of potential self-resistance genes in BGCs from across the tree of life, including metagenomic data. A similar approach, specifically dedicated to polyketide BGCs, was also released by others.³⁴ Additionally, a new study shows that transporter-encoding genes can also be used as functional markers for target-based genome mining.³⁵

Predicting chemical structures

The ability to (partially) predict chemical structures of the products of BGCs is key for identifying potential chemical novelty during the genome mining process, as well as for matching BGCs to metabolites from analytical data. Several new tools have been developed that can aid in such efforts. The new version 4 of PRISM¹² has improved chemical structure prediction capabilities, which made it possible to train machine-learning models to predict the biological activity of BGC products based on these structure predictions. Two new algorithms, DDAP³⁶ and PKSpop,³⁷ provide improved prediction of docking domain interactions between polyketide synthases, which determine the order of these enzymes in the assembly lines, and thus also the order of the incorporated monomers in their final products. To go from monomers towards final products, another group published a machine-learning method that predicts macrocyclization patterns for both polyketides and nonribosomal peptides.³⁸ Extending beyond the scope of



Marnix Medema is an Assistant Professor of Bioinformatics at Wageningen University, The Netherlands. He obtained a Biology BSc (Radboud University Nijmegen, 2006) and a Biomolecular Sciences MSc (University of Groningen, 2008). In 2013, he completed his PhD with Eriko Takano and Rainer Breitling in Groningen; during this period, he was also a visiting fellow with Michael Fischbach at

the University of California, San Francisco. Following a postdoc at the Max Planck Institute for Marine Microbiology in Bremen, Germany, he joined Wageningen University in 2015. There, his group develops computational methodologies to unravel natural product biosynthesis using omics data, and applies these methods to the study of molecular interactions in microbiomes.



Highlight

megasynthases, the AdenylPred³⁹ algorithm presents a new method to predict catalytic functions and substrate specificities for the whole superfamily of adenylate-forming enzymes, which include not only nonribosomal peptide synthetase adenylation domains, but also *e.g.* fatty-acyl CoA-ligases and beta-lactone synthetases.

Analysing natural product NMR data

Elucidating chemical structures is arguably worth more than predicting them. NMR data play a crucial role in this, but algorithms to automate the analysis of such data have been lagging. In the past year, some exciting breakthroughs were published in this area. The SMART 2.0 algorithm⁴⁰ is a convolutional neural network-based approach that automatically generates structure hypotheses from ¹H-¹³C-HSQC spectra. Other methods that aid in interpreting NMR spectra also appeared, including a classifier that assigns molecules to a natural product class based on ¹³C spectra⁴¹ and the DP4-AI machine learning algorithm that aims to automate structure assignment from NMR spectra.⁴² For the analysis of natural product mixtures (extracts or fractions), MixONat⁴³ provides a new tool for automated dereplication.

Developments in mass spectrometry data analysis

Within the realm of analytical techniques, the analysis of tandem mass-spectrometric (MS/MS) data has been revolutionized in recent years, and a range of groundbreaking new methods have been added in 2020. The ZODIAC algorithm⁴⁴ uses Gibbs sampling and Bayesian statistics to accurately predict molecular formulas for a compound by considering joint fragments and losses in fragmentation trees, CANOPUS⁴⁵ uses fragmentation spectra to automatically classify molecules into ~2500 classes with deep learning, and Retip⁴⁶ provides a new way of predicting metabolite retention times from chemical structures. MetFID⁴⁷ provides a new neural network-based algorithm to predict compound fingerprints from MS/MS spectra, which aids the structural annotation of the underlying metabolites. With MASST,⁴⁸ a 'BLAST for molecules' was introduced that facilitates rapid similarity searches for MS/MS spectra and allows users to assess in which publicly available samples a metabolite of interest is present. Several new methods for and improvements to molecular networking technologies were also put forward, including Spec2Vec,⁴⁹ which uses natural language processing to identify similarities in a way that takes into account patterns observed across large datasets. Additionally, feature-based molecular networking (FBMN)⁵⁰ was introduced in the Global Natural Products Social Molecular Networking (GNPS) infrastructure, which incorporates information from ion mobility separation. Additionally, the GNPS framework was also improved to facilitate the analysis of gas chromatography-mass spectrometry data,⁵¹ and the ReDU system makes it possible to straightforwardly re-analyze public MS/MS datasets by identifying them through a controlled

vocabulary.⁵² As an alternative to molecular networking, Qemistree facilitates analysing chemical diversity from MS data using hierarchical clustering.⁵³ Finally, several developments in databases of mass spectra are notable: BMDMS-NP⁵⁴ provides a comprehensive library of almost 3000 ESI-MS/MS spectra for plant natural products, while METLIN provides molecular standards for ~850 000 metabolites, including many natural products.⁵⁵

Linking MS data to structures and gene clusters

Improvements have also been made for the analysis of specific types of molecules, such as peptides: with CycloNovo,⁵⁶ a new software was released that enables high-throughput *de novo* sequencing of peptides from MS/MS data. In addition, NRPro⁵⁷ automatically annotates and dereplicates peptidic natural products based on their tandem mass spectra. Such peptides can be linked to BGCs with increasing effectiveness, with methods such as MetaMiner,⁵⁸ which matches genomically predicted peptides with their possible modifications to the monomers inferred from MS data. Connecting MS data to BGCs can also be done based on absence/presence correlations of molecules and gene clusters across strains, and the NPLinker framework provides the first full-fledged software that automates this, and also introduces a new scoring function.⁵⁹

Conclusions

Computational methods are becoming more and more ingrained in the day-to-day science of natural product researchers, and the speed with which new methods are introduced reflects this. Even outside the familiar realms of natural product bioinformatics outlined above, exciting new approaches are being introduced, including a new deep learning approach to predict antibiotic activities from chemical structures⁶⁰ and computational approach to automatically plan efficient routes toward the total synthesis of natural products.⁶¹ The year 2021 is likely to again provide a similar range of new approaches, and navigating the diversity of available algorithms will become an increasingly important skill for those who are trained in natural product science.

Conflicts of interest

M. H. M. is a co-founder of Design Pharmaceuticals and a member of the scientific advisory board of Hexagon Bio.

References

- 1 J. A. van Santen, G. Jacob, A. L. Singh, V. Aniebok, M. J. Balunas, D. Bunsko, F. C. Neto, L. Castaño-Espriu, C. Chang, T. N. Clark, J. L. Cleary Little, D. A. Delgadillo, P. C. Dorrestein, K. R. Duncan, J. M. Egan, M. M. Galey, F. P. J. Haeckl, A. Hua, A. H. Hughes, D. Iskakova, A. Khadilkar, J.-H. Lee, S. Lee, N. LeGrow, D. Y. Liu, J. M. Macho, C. S. McCaughey, M. H. Medema,



- R. P. Neupane, T. J. O'Donnell, J. S. Paula, L. M. Sanchez, A. F. Shaikh, S. Soldatou, B. R. Terlouw, T. A. Tran, M. Valentine, J. J. J. van der Hooft, D. A. Vo, M. Wang, D. Wilson, K. E. Zink and R. G. Linington, *ACS Cent. Sci.*, 2019, 5, 1824–1833.
- 2 A. F. A. Moumbock, M. Gao, A. Qaseem, J. Li, P. A. Kirchner, B. Ndingkokhar, B. D. Bekono, C. V. Simoben, S. B. Babiaka, Y. I. Malange, F. Sauter, P. Zierep, F. Ntie-Kang and S. Günther, *Nucleic Acids Res.*, 2020, 49, D600–D604.
- 3 M. R. Jones, E. Pinto, M. A. Torres, F. Dörr, H. Mazur-Marzec, K. Szubert, L. Tartaglione, C. Dell'Aversano, C. O. Miles, D. G. Beach, P. McCarron, K. Sivonen, D. P. Fewer, J. Jokela and E. M.-L. Janssen, *bioRxiv*, 2020, DOI: 10.1101/2020.04.16.038703.
- 4 P. P. K. Zin, G. J. Williams and S. Ekins, *Sci. Rep.*, 2020, 10, 6284.
- 5 A. Flissi, E. Ricart, C. Campart, M. Chevalier, Y. Dufresne, J. Michalik, P. Jacques, C. Flahaut, F. Lisacek, V. Leclère and M. Pupin, *Nucleic Acids Res.*, 2020, 48, D465–D469.
- 6 E. Ricart, V. Leclère, A. Flissi, M. Mueller, M. Pupin and F. Lisacek, *J. Cheminf.*, 2019, 11, 13.
- 7 M. Sorokina and C. Steinbeck, *J. Cheminf.*, 2020, 12, 20.
- 8 H. Kim, M. Wang, C. Leber, L.-F. Nothias, R. Reher, K. B. Kang, J. J. J. van der Hooft, P. Dorrestein, W. Gerwick and G. Cottrell, *ChemRxiv*, 2020, DOI: 10.26434/chemrxiv.12885494.v1.
- 9 M. Seo, H. K. Shin, Y. Myung, S. Hwang and K. T. No, *J. Cheminf.*, 2020, 12, 6.
- 10 A. Capecchi, D. Probst and J.-L. Reymond, *J. Cheminf.*, 2020, 12, 43.
- 11 K. Rajan, A. Zielesny and C. Steinbeck, *J. Cheminf.*, 2020, 12, 65.
- 12 M. A. Skinnider, C. W. Johnston, M. Gunabalasingam, N. J. Merwin, A. M. Kieliszek, R. J. MacLellan, H. Li, M. R. M. Ranieri, A. L. H. Webster, M. P. T. Cao, A. Pfeifle, N. Spencer, Q. H. To, D. P. Wallace, C. A. Dejong and N. A. Magarvey, *Nat. Commun.*, 2020, 11, 6058.
- 13 P. F. Zierep, A. T. Ceci, I. Dobrusin, S. C. Rockwell-Kollmann and S. Günther, *Metabolites*, 2021, 11, 13.
- 14 N. Sélem-Mojica, C. Aguilar, K. Gutiérrez-García, C. E. Martínez-Guerrero and F. Barona-Gómez, *Microb. Genomes*, 2020, 5, e000260.
- 15 E. Gluck-Thaler, S. Haridas, M. Binder, I. V. Grigoriev, P. W. Crous, J. W. Spatafora, K. Bushley and J. C. Slot, *Mol. Biol. Evol.*, 2020, 37, 2838–2856.
- 16 H. Almeida, S. Palys, A. Tsang and A. B. Diallo, *NAR: Genomics Bioinf.*, 2020, 2, lqaa098.
- 17 A. M. Kloosterman, M. H. Medema and G. P. van Wezel, *Curr. Opin. Biotechnol.*, 2020, 69, 60–67.
- 18 N. J. Merwin, W. K. Mousa, C. A. Dejong, M. A. Skinnider, M. J. Cannon, H. Li, K. Dial, M. Gunabalasingam, C. Johnston and N. A. Magarvey, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, 117, 371–380.
- 19 A. M. Kloosterman, K. E. Shelton, G. P. van Wezel, M. H. Medema and D. A. Mitchell, *mSystems*, 2020, 5, e00267–20.
- 20 A. M. Kloosterman, P. Cimerancic, S. S. Elsayed, C. Du, M. Hadjithomas, M. S. Donia, M. A. Fischbach, G. P. van Wezel and M. H. Medema, *PLoS Biol.*, 2020, 18, e3001026.
- 21 J. Santos-Aberturas, G. Chandra, L. Frattaruolo, R. Lacroix, T. H. Pham, N. M. Vior, T. H. Eyles and A. W. Truman, *Nucleic Acids Res.*, 2019, 47, 4624–4637.
- 22 M. A. Georgiou, S. R. Dommaraju, X. Guo and D. A. Mitchell, *ACS Chem. Biol.*, 2020, 15, 2976–2985.
- 23 J. C. Navarro-Muñoz, N. Selem-Mojica, M. W. Muldowney, S. Kautsar, J. H. Tryon, E. I. Parkinson, E. L. C. De Los Santos, M. Yeong, P. Cruz-Morales, S. Abubucker, A. Roeters, W. Lokhorst, A. Fernandez-Guerra, L. T. D. Cappelini, R. J. Thomson, W. W. Metcalf, N. L. Kelleher, F. Barona-Gomez and M. H. Medema, *Nat. Chem. Biol.*, 2020, 16, 60–68.
- 24 S. A. Kautsar, J. J. J. van der Hooft, D. de Ridder and M. H. Medema, *GigaScience*, 2021, 10, g1aa154.
- 25 S. A. Kautsar, K. Blin, S. Shaw, T. Weber and M. H. Medema, *Nucleic Acids Res.*, 2020, 49, D490–D497.
- 26 C. L. M. Gilchrist, T. J. Booth and Y.-H. Chooi, *bioRxiv*, 2020, DOI: 10.1101/2020.11.08.370601.
- 27 C. L. M. Gilchrist and Y.-H. H. Chooi, *Bioinformatics*, 2021, DOI: 10.1093/bioinformatics/btab007.
- 28 S. A. Kautsar, K. Blin, S. Shaw, J. C. Navarro-Muñoz, B. R. Terlouw, J. J. J. van der Hooft, J. A. van Santen, V. Tracanna, H. G. Suarez Duran, V. Pascal Andreu, N. Selem-Mojica, M. Alanjary, S. L. Robinson, G. Lund, S. C. Epstein, A. C. Sisto, L. K. Charkoudian, J. Collemare, R. G. Linington, T. Weber and M. H. Medema, *Nucleic Acids Res.*, 2020, 48, D454–D458.
- 29 K. Blin, S. Shaw, S. A. Kautsar, M. H. Medema and T. Weber, *Nucleic Acids Res.*, 2020, 49, D639–D643.
- 30 K. Palaniappan, I.-M. A. Chen, K. Chu, A. Ratner, R. Seshadri, N. C. Kyrpides, N. N. Ivanova and N. J. Mouncey, *Nucleic Acids Res.*, 2020, 48, D422–D430.
- 31 M. T. Robey, L. K. Caesar, M. T. Drott, N. P. Keller and N. L. Kelleher, *bioRxiv*, 2020, DOI: 10.1101/2020.09.21.307157.
- 32 J. K. Schniete, N. Selem-Mojica, A. S. Birke, P. Cruz-Morales, I. S. Hunter, F. Barona-Gómez and P. A. Hoskisson, *Microb. Genomes*, 2021, DOI: 10.1099/mgen.0.000498.
- 33 M. D. Mungan, M. Alanjary, K. Blin, T. Weber, M. H. Medema and N. Ziemert, *Nucleic Acids Res.*, 2020, 48, W546–W552.
- 34 G. A. Vandova, A. Nivina, C. Khosla, R. W. Davis, C. R. Fisher and M. E. Hillenmeyer, *bioRxiv*, 2020, DOI: 10.1101/2020.06.01.128595.
- 35 A. Crits-Christoph, N. Bhattacharya, M. R. Olm, Y. S. Song and J. F. Banfield, *Genome Res.*, 2020, DOI: 10.1101/gr.268169.120.
- 36 T. Li, A. Tripathi, F. Yu, D. H. Sherman and A. Rao, *Bioinformatics*, 2020, 36, 942–944.
- 37 Y. Wang, M. C. Marrero, M. H. Medema and A. D. J. van Dijk, *Bioinformatics*, 2020, 19, 4846–4853.
- 38 P. Agrawal and D. Mohanty, *Bioinformatics*, 2020, DOI: 10.1093/bioinformatics/btaa851.



- 39 S. L. Robinson, B. R. Terlouw, M. D. Smith, S. J. Pidot, T. P. Stinear, M. H. Medema and L. P. Wackett, *J. Biol. Chem.*, 2020, **295**, 14826–14839.
- 40 R. Reher, H. W. Kim, C. Zhang, H. H. Mao, M. Wang, L.-F. Nothias, A. M. Caraballo-Rodriguez, E. Glukhov, B. Teke, T. Leao, K. L. Alexander, B. M. Duggan, E. L. Van Everbroeck, P. C. Dorrestein, G. W. Cottrell and W. H. Gerwick, *J. Am. Chem. Soc.*, 2020, **142**, 4114–4120.
- 41 S. H. Martínez-Treviño, V. Uc-Cetina, M. A. Fernández-Herrera and G. Merino, *J. Chem. Inf. Model.*, 2020, **60**, 3376–3386.
- 42 A. Howarth, K. Ermanis and J. Goodman, *Chem. Sci.*, 2020, **11**, 4351–4359.
- 43 A. Bruguière, S. Derbré, J. Dietsch, J. Leguy, V. Rahier, Q. Pottier, D. Bréard, S. Suor-Cherer, G. Vialat, A.-M. Le Ray, F. Saubion and P. Richomme, *Anal. Chem.*, 2020, **92**, 8793–8801.
- 44 M. Ludwig, L.-F. Nothias, K. Dührkop, I. Koester, M. Fleischauer, M. A. Hoffmann, D. Petras, F. Vargas, M. Morsy, L. Aluwihare, P. C. Dorrestein and S. Böcker, *Nat. Mach. Intell.*, 2020, **2**, 629–641.
- 45 K. Dührkop, L.-F. Nothias, M. Fleischauer, R. Reher, M. Ludwig, M. A. Hoffmann, D. Petras, W. H. Gerwick, J. Rousu, P. C. Dorrestein and S. Böcker, *Nat. Biotechnol.*, 2020, DOI: 10.1038/s41587-020-0740-8.
- 46 P. Bonini, T. Kind, H. Tsugawa, D. K. Barupal and O. Fiehn, *Anal. Chem.*, 2020, **92**, 7515–7522.
- 47 Z. Fan, A. Alley, K. Ghaffari and H. W. Ransom, *Metabolomics*, 2020, **16**, 104.
- 48 M. Wang, A. K. Jarmusch, F. Vargas, A. A. Aksenov, J. M. Gauglitz, K. Weldon, D. Petras, R. da Silva, R. Quinn, A. V. Melnik, J. J. J. van der Hooft, A. M. Caraballo-Rodriguez, L. F. Nothias, C. M. Aceves, M. Panitchpakdi, E. Brown, F. Di Ottavio, N. Sikora, E. O. Elijah, L. Labarta-Bajo, E. C. Gentry, S. Shalapour, K. E. Kyle, S. P. Puckett, J. D. Watrous, C. S. Carpenter, A. Bouslimani, M. Ernst, A. D. Swafford, E. I. Zúñiga, M. J. Balunas, J. L. Klassen, R. Loomba, R. Knight, N. Bandeira and P. C. Dorrestein, *Nat. Biotechnol.*, 2020, **38**, 23–26.
- 49 F. Huber, L. Ridder, S. Verhoeven, J. H. Spaaks, F. Diblen, S. Rogers and J. J. J. van der Hooft, *bioRxiv*, 2020, DOI: 10.1101/2020.08.11.245928.
- 50 L.-F. Nothias, D. Petras, R. Schmid, K. Dührkop, J. Rainer, A. Sarvepalli, I. Protsyuk, M. Ernst, H. Tsugawa, M. Fleischauer, F. Aicheler, A. A. Aksenov, O. Alka, P.-M. Allard, A. Barsch, X. Cachet, A. M. Caraballo-Rodriguez, R. R. Da Silva, T. Dang, N. Garg, J. M. Gauglitz, A. Gurevich, G. Isaac, A. K. Jarmusch, Z. Kamenik, K. B. Kang, N. Kessler, I. Koester, A. Korf, A. Le Gouellec, M. Ludwig, C. Martin H, L.-I. McCall, J. McSayles, S. W. Meyer, H. Mohimani, M. Morsy, O. Moyne, S. Neumann, H. Neuweiger, N. H. Nguyen, M. Nothias-Esposito, J. Paolini, V. V. Phelan, T. Pluskal, R. A. Quinn, S. Rogers, B. Shrestha, A. Tripathi, J. J. J. van der Hooft, F. Vargas, K. C. Weldon, M. Witting, H. Yang, Z. Zhang, F. Zubeil, O. Kohlbacher, S. Böcker, T. Alexandrov, N. Bandeira, M. Wang and P. C. Dorrestein, *Nat. Methods*, 2020, **17**, 905–908.
- 51 A. A. Aksenov, I. Laponogov, Z. Zhang, S. L. F. Doran, I. Belluomo, D. Veselkov, W. Bittremieux, L. F. Nothias, M. Nothias-Esposito, K. N. Maloney, B. B. Misra, A. V. Melnik, A. Smirnov, X. Du, K. L. Jones, K. Dorrestein, M. Panitchpakdi, M. Ernst, J. J. J. van der Hooft, M. Gonzalez, C. Carazzone, A. Amézquita, C. Callewaert, J. T. Morton, R. A. Quinn, A. Bouslimani, A. A. Orio, D. Petras, A. M. Smania, S. P. Couvillion, M. C. Burnet, C. D. Nicora, E. Zink, T. O. Metz, V. Artaev, E. Humston-Fulmer, R. Gregor, M. M. Meijler, I. Mizrahi, S. Eyal, B. Anderson, R. Dutton, R. Lugan, P. Le Boulch, Y. Guitton, S. Prevost, A. Poirier, G. Dervilly, B. Le Bizec, A. Fait, N. S. Persi, C. Song, K. Gashu, R. Coras, M. Guma, J. Manasson, J. U. Scher, D. K. Barupal, S. Alseekh, A. R. Fernie, R. Mirnezami, V. Vasiliou, R. Schmid, R. S. Borisov, L. N. Kulikova, R. Knight, M. Wang, G. B. Hanna, P. C. Dorrestein and K. Veselkov, *Nat. Biotechnol.*, 2020, DOI: 10.1038/s41587-020-0700-3.
- 52 A. K. Jarmusch, M. Wang, C. M. Aceves, R. S. Advani, S. Aguirre, A. A. Aksenov, G. Aleti, A. T. Aron, A. Bauermeister, S. Bolleddu, A. Bouslimani, A. M. C. Rodriguez, R. Chaar, R. Coras, E. O. Elijah, M. Ernst, J. M. Gauglitz, E. C. Gentry, M. Husband, S. A. Jarmusch, K. L. Jones, Z. Kamenik, A. Le Gouellec, A. Lu, L.-I. McCall, K. L. McPhail, M. J. Meehan, A. V. Melnik, R. C. Menezes, Y. A. M. Giraldo, N. H. Nguyen, L. F. Nothias, M. Nothias-Esposito, M. Panitchpakdi, D. Petras, R. A. Quinn, N. Sikora, J. J. J. van der Hooft, F. Vargas, A. Vrbanac, K. C. Weldon, R. Knight, N. Bandeira and P. C. Dorrestein, *Nat. Methods*, 2020, **17**, 901–904.
- 53 A. Tripathi, Y. Vázquez-Baeza, J. M. Gauglitz, M. Wang, K. Dührkop, M. Nothias-Esposito, D. D. Acharya, M. Ernst, J. J. J. van der Hooft, Q. Zhu, D. McDonald, A. Gonzalez, J. Handelsman, M. Fleischauer, M. Ludwig, S. Böcker, L.-F. Nothias, R. Knight and P. C. Dorrestein, *Nat. Chem. Biol.*, 2021, **17**, 146–151.
- 54 S. Lee, S. Hwang, M. Seo, K. B. Shin, K. H. Kim, G. W. Park, J. Y. Kim, J. S. Yoo and K. T. No, *Phytochemistry*, 2020, **177**, 112427.
- 55 J. Xue, C. Guijas, H. Paul Benton, B. Warth and G. Siuzdak, *Nat. Methods*, 2020, **17**, 953–954.
- 56 B. Behsaz, H. Mohimani, A. Gurevich, A. Prjibelski, M. Fisher, F. Vargas, L. Smarr, P. C. Dorrestein, J. S. Mylne and P. A. Pevzner, *Cell Syst.*, 2020, **10**, 99–108.e5.
- 57 E. Ricart, M. Pupin, M. Müller and F. Lisacek, *Anal. Chem.*, 2020, **92**, 15862–15871.
- 58 L. Cao, A. Gurevich, K. L. Alexander, C. B. Naman, T. Leão, E. Glukhov, T. Luzzatto-Knaan, F. Vargas, R. Quinn, A. Bouslimani, L. F. Nothias, N. K. Singh, J. G. Sanders, R. A. S. Benitez, L. R. Thompson, M.-N. Hamid, J. T. Morton, A. Mikheenko, A. Shlemov, A. Korobeynikov, I. Friedberg, R. Knight, K. Venkateswaran, W. H. Gerwick, L. Gerwick, P. C. Dorrestein, P. A. Pevzner and H. Mohimani, *Cell Syst.*, 2019, **9**, 600–608.e4.



- 59 G. H. Eldjárn, A. Ramsay, J. J. J. van der Hooft, K. R. Duncan, S. Soldatou, J. Rousu, R. Daly, J. Wandy and S. Rogers, *bioRxiv*, 2020, DOI: 10.1101/2020.06.12.148205.
- 60 J. M. Stokes, K. Yang, K. Swanson, W. Jin, A. Cubillos-Ruiz, N. M. Donghia, C. R. MacNair, S. French, L. A. Carfrae, Z. Bloom-Ackermann, V. M. Tran, A. Chiappino-Pepe, A. H. Badran, I. W. Andrews, E. J. Chory, G. M. Church, E. D. Brown, T. S. Jaakkola, R. Barzilay and J. J. Collins, *Cell*, 2020, **181**, 475–483.
- 61 B. Mikulak-Klucznik, P. Gołębiowska, A. A. Bayly, O. Popik, T. Klucznik, S. Szymkuć, E. P. Gajewska, P. Dittwald, O. Staszewska-Krajewska, W. Beker, T. Badowski, K. A. Scheidt, K. Molga, J. Młynarski, M. Mrksich and B. A. Grzybowski, *Nature*, 2020, **588**, 83–88.

