Natural Product Reports

ROYAL SOCIETY OF CHEMISTRY

VIEWPOINT

View Article Online
View Journal | View Issue

Cite this: *Nat. Prod. Rep.*, 2021, **38**, 1947

Benefiting from big data in natural products: importance of preserving foundational skills and prioritizing data quality

Nadja B. Cech, *[a] Marnix H. Medema *[b] and Jon Clardy *[c]

Systematic, large-scale, studies at the genomic, metabolomic, and functional level have transformed the natural product sciences. Improvements in technology and reduction in cost for obtaining spectroscopic, chromatographic, and genomic data coupled with the creation of readily accessible curated and functionally annotated data sets have altered the practices of virtually all natural product research laboratories. Gone are the days when the natural products researchers were expected to devote themselves exclusively to the isolation, purification, and structure elucidation of small molecules. We now also engage with big data in taxonomic, genomic, proteomic, and/or metabolomic collections, and use these data to generate and test hypotheses. While the oft stated aim for the use of large-scale -omics data in the natural products sciences is to achieve a rapid increase in the rate of discovery of new drugs, this has not yet come to pass. At the same time, new technologies have provided unexpected opportunities for natural products chemists to ask and answer new and different questions. With this viewpoint, we discuss the evolution of big data as a part of natural products research and provide a few examples of how discoveries have been enabled by access to big data. We also draw attention to some of the limitations in our existing engagement with large datasets and consider what would be necessary to overcome them.

Received 16th September 2021

DOI: 10.1039/d1np00061f

rsc.li/npr

1. Introduction

Natural products research today is increasingly dependent on so called "big data" – the systematically curated data from large-scale genomic, metabolomic, and functional studies. The transformation of our field from what was once a prototypical small science took place over about a 100 year period, starting with recognition of the therapeutic value of microbial natural products. This recognition was sparked by Fleming's discovery of penicillin from *Penicillium chrysogenum*,[1] followed by penicillin's development during WWII, and its introduction into widespread clinical use. Postwar discoveries from the Waksman laboratory at Rutgers University illustrated the generality of mining microbes for antibiotics. The importance of these discoveries was recognized through the 1952 Nobel Prize for the discovery of streptomycin, the first antibiotic that was effective in treating tuberculosis. Soon, researchers around the world were finding microbially-derived molecules that led to what we now know as the Golden Age of Antibiotics. Recognition of the importance of antibiotic (and other drug) discovery inspired intense and systematic searches for therapeutically useful small molecules. Initial research towards this goal was largely conducted by performing phenotypic screening to identify active lead extracts, and following up on these leads by bioactivity-guided fractionation, isolation, and structure elucidation. Here we refer to these approaches as the "foundational skills" of natural products chemistry because they have played such an important role in drug discovery. It is notable that all of the important classes of natural product derived antibiotics used clinically, as well as multiple transformative drugs for diseases like cancer (taxol and camptothecin, discovered by Wall and Wani[2]) and malaria (artemisinin, discovered by Tu Youyou[3]), were discovered without the recent advances in NMR and mass spectrometry that enable rapid structure elucidation and without access to big data as we define it today.

Another pivotal moment in the evolution of the natural products field came when the Hopwood lab at the John Innes Centre reported that the genes encoding the enzymes responsible for the production of a natural product by *Streptomyces coelicolor* were clustered on a stretch of DNA.[4] This recognition led to our current ability to detect biosynthetic gene clusters (BGCs) in microbial (and other) genomes, and to parse these clusters to describe the molecules they produce. Most importantly, the development of tools to detect gene clusters opened up the possibility of interrogating pre-existing genomic data to

[a]*Chemistry, University of North Carolina Greensboro, USA. E-mail: nadja_cech@uncg.edu*

[b]*Bioinformatics, Wageningen University, The Netherlands. E-mail: marnix.medema@wur.nl*

[c]*Biological Chemistry and Molecular Pharmacology, Harvard Medical School, USA. E-mail: jon_clardy@hms.harvard.edu*

probe for compounds of interest. The ability to harness genomics for natural products research was further fueled by the large-scale sequencing of genomes and later metagenomes, advances that were possible thanks to drastic reductions in DNA sequencing costs and increased capacity for computer storage. Technological innovations led to sociological adjustments as well. What used to be isolated natural products research efforts began to involve team and community contributions. Natural products scientists no longer devoted themselves exclusively to isolation and structure elucidation. They also undertook efforts to curate and maintain databases and to improve the tools for analyzing them.

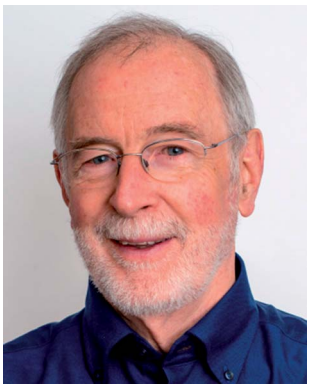
Changes in the magnitude and type of data available to researchers in the natural products field are reflected in what we view as natural products research today. In search of relevant natural products, we routinely interrogate entire genomes or metagenomes and complex mixtures of proteins (proteomes) or small molecule metabolites (metabolomes). Increasingly, the data that support these projects do not reside entirely in a single laboratory but are shared in public and community-supported databases. The consequences of this shift, which we refer to here as the big data revolution in natural products, forms the basis of this themed issue of Natural Product Reports.

Reading the articles included in this themed issue, it is possible to imagine a future for natural products that is increasingly collaborative, leveraging the collective intelligence, skillsets, perspectives, and, importantly, *data* of scientists around the world. In idealized natural products research projects of the future, organisms of interest would be selected not only based on serendipity and accessibility of a given organism, but by comparing genetic, spectroscopic, or functional data from curated databases (Chevrette *et al.*, DOI: 10.1039/D1NP00013F; Bauman *et al.*, DOI: 10.1039/D1NP00032B; Chevrette and Handelsman, DOI: 10.1039/D1NP00044F). Computational algorithms trained on data from such databases would be used to predict the structures of the secondary metabolites produced by the organisms of interest (Caesar *et al.*, DOI: 10.1039/D1NP00036E),[5] the families to which these compounds belong (van Santen *et al.*, DOI: 10.1039/D0NP00053A), and even their biological activity (Jeon *et al.*, DOI: 10.1039/D1NP00016K). Larger scale, better integrated and higher quality datasets of the gene sequences, protein sequences and small molecule structures associated with living organisms (Bauman *et al.*, DOI: 10.1039/D1NP00032B) would empower future research using artificial intelligence as a powerful new discovery tool (Jeon *et al.*, DOI: 10.1039/D1NP00016K). An important caveat to these optimistic scenarios is the eventual necessity of actually producing the molecules whose existence and properties can be inferred so that their structures and functions can be experimentally verified. Thus, it is critical that in our pursuit of new and exciting technologies, we do not lose sight of the need to train the younger generation of natural products scientists in the skills necessary to isolate and solve natural product structures.

The reviews collected in this themed issue describe rapid progress that is being made on many fronts, all of which promise to contribute to a more integrated, collaborative, and efficient future for natural products research. These reviews also

Dr Nadja Cech is Patricia A. Sullivan Distinguished Professor of Chemistry at the University of North Carolina at Greensboro (UNCG). She leads a dynamic research group that develops metabolomics tools to study biologically relevant natural products. Dr Cech's research is funded by the National Institutes of Health through the Center of Excellence for Natural Product Drug Interaction Research (NaPDI, https://napdicenter.org/) and the Center for High Content Functional Annotation of Natural Products (HiFAN, https://hifan2.sites.ucsc.edu/).

Dr Marnix Medema is an Assistant Professor of Bioinformatics at Wageningen University, The Netherlands. His group develops computational algorithms and databases to unravel natural product biosynthesis using omics data, and applies these methods to study molecular interactions in microbiomes as well as to accelerate natural product drug discovery.



Dr Jon Clardy is the Hsien Wu & Daisy Yen Wu Professor of Biological Chemistry at Harvard Medical School. These days he leads a small but focused team that explores the relations of metabolites from gut microbes, the immune system, and health and disease in humans. His research is largely funded by the National Institutes of Health through the National Center for Complementary and Integrative Health (NCCIH), Harvard Medical School (HMS), and the Center for the Study of Inflammatory Bowel Disease (CSIBD).

highlight many critical barriers that still exist to fully leveraging big data for successful natural products research. Spectroscopic data, structural data, and genomic data are distributed across many databases, none of which are fully comprehensive. Many databases are not curated and may include erroneous information, such as incorrectly annotated gene clusters, incorrect structures, or errors in annotation of spectral data. Also, databases containing gene and protein sequences and chemical structures are, as of yet, not optimally integrated with each other. Meta-data is lacking. Much of the data that has been collected thus far is either proprietary or buried in the scientific literature in a format that is not easily searchable. Significant resources in terms of time and money are needed to address all of these issues, and the work required is, to quote van Santen *et al.* (DOI: 10.1039/D0NP00053A), "unglamorous." Finally, the understandable desire on the part of many researchers to protect the intellectual property associated with their data hinders efforts to make those data more freely available.

What incentives exist to overcome these barriers? The big data revolution has not just changed how we do natural products research, it has also changed what we define as natural products research, creating opportunities to explore new questions and to interface in new ways with scientists in related fields. In the coming years, we expect that the impact of big data will continue to be felt across the community of scientists doing natural products research, and that creative solutions will be developed to address the most pressing hurdles currently hampering such progress. Here we discuss some of these specific hurdles and provide a few examples of how big data has been effectively leveraged despite them.

2. Developments in storing and accessing natural product chemical structures

The review included in this themed issue by van Santen *et al.* (DOI: 10.1039/D0NP00053A) describes existing databases for depositing structures of natural products, their promise, and their limitations. While much of the structural data on natural products has historically remained buried in the primary literature, trapped behind paywalls and/or in a format that is not easily searchable, there is currently great momentum towards making spectroscopic and structural data on natural products freely accessible. Two notable open access databases that have recently been developed to store natural products structural data are Natural Products Atlas[6] and COCONUT.[7] Already, databases such as these are enabling improvements in the way we do natural products research. For example, the COCONUT database has recently been leveraged to create a resource (called LOTUS) directly linking natural product chemical structures to freely available articles describing the characterization and biological evaluation of the compounds.[8] In another recent report, the Natural Products Atlas was queried by Robey *et al.*[9] to collect data from 15 213 fungal metabolites. These compounds were organized into molecular families and used to annotate the fungal gene clusters from 1000 fungal genomes. These examples are a portent of the future potential of linking different types of data relevant to natural products research, a topic discussed in more detail in several reviews in this themed issue (Bauman *et al.*, DOI: 10.1039/D1NP00032B; Caesar *et al.*, DOI: 10.1039/D1NP00036E; van Santen *et al.*, DOI: 10.1039/D0NP00053A).[5]

Our field is likely nearing a tipping point where researchers will begin to rely more on open-access databases and less on historical subscription-only compendia of natural product structures such as MarinLit, AntiBase, and the Dictionary of Natural Products. At the present time, however, the open access databases are still not comprehensive enough to entirely replace the subscription-based sources, and the existence of multiple different platforms for storing chemical structures, each with different coverage and linked to different types of information, creates a great deal of confusion among researchers. We are still some way off from the future vision of a single, comprehensive database or systematic cross-linking and integration of existing databases. There is no question, however, that the existence of such resources would be of tremendous benefit to the continued success of natural products research endeavors.

3. Developments in compiling, accessing and using spectroscopic data

It is becoming increasingly popular to make the spectroscopic data that accompanies natural products research studies freely available. This practice has the potential to enable more rapid and efficient structure elucidation than is currently possible, and to facilitate large-scale studies that leverage datasets across laboratories. A recent review by McAlpine *et al.* provides excellent perspective on the challenges and opportunities associated with sharing NMR data.[10] The sharing of liquid chromatography-mass spectrometry (LC-MS) datasets comes with its own set of challenges, foremost among them the inherent variability in results across platforms. For example, it should be possible to putatively identify unknown molecules by comparing their fragmentation spectra (MS-MS data) against databases created for known molecules. This practice is becoming more routine, particularly as access to fragmentation data is enhanced due to the expansion of databases such as GNPS (reviewed in Jarmusch *et al.*, DOI: 10.1039/D1NP00040C) and the availability of structure elucidation tools.[11] However, the quality of data in MS-MS databases is somewhat variable and searchable MS-MS data is still lacking for many known natural products. As the size and quality of MS-MS databases increases, we predict that the use of mass spectrometry data as a first step towards structure elucidation will become a more routine practice in natural products research. Nonetheless, because of differences in fragmentation behavior across platforms and similarity in fragmentation of isomers, MS-MS spectral matching will never be the definitive answer for compound identification. Orthogonal data (from NMR, MicroED, and/or in-house analysis of standards) will always be

required to confirm putative structural assignments made with MS.

Variability in MS data across platforms also occurs due to differences in the type of clusters, fragments, and adducts produced by different electrospray source designs and configurations. This variability in ionization behavior makes it difficult to compare LC-MS datasets between laboratories.[12] Furthermore, there is a tendency to overestimate the complexity of metabolomics datasets because each individual analyte gives rise to more than one signal.[13] The complexity of MS datasets is further increased due to interference from chemical species that are present as contaminates in the system, including the solvents, the column, the plumbing, and even the laboratory atmosphere.[14,15] The complexity of mass spectrometry datasets becomes particularly challenging in untargeted metabolomics experiments, where the goal is often to track or annotate all analytes present in each mixture. Thus, there is a need for effective approaches to reduce the complexity of such datasets, either by grouping the signals associated with single analytes,[16] and/or removing irrelevant signals that arise from chemical interference.[14]

It is exciting to observe that an increasing number of researchers are now uploading LC-MS datafiles to accompany their papers in servers such as GNPS-massIVE.[17] Poor annotation of data files, poor data quality (*i.e.* noise in the data or lack of appropriate blanks, QC samples, and/or replicates), uniqueness of the data to the platform on which they were collected, and lack of associated metadata often limits the value of these data to researchers in other laboratories. The establishment and adoption of best practices for collecting, processing, and sharing metabolomics data for natural products would help to address some of these limitations.

Despite the associated challenges, we are beginning to see research projects that query publicly available mass spectrometry data across laboratories to answer scientifically interesting questions (see Jarmusch *et al.*, DOI: 10.1039/D1NP00040C). As a recent example, Jarmusch *et al.* developed a tool (called ReDU)[18] that enables comparison of the shared and different chemicals between groups of samples, and makes it possible to conduct repository-scale molecular networking. Using this tool, they profiled the distribution of 12-ketodeoxycholic acid, cholic acid, and rosuvastatin by mining more than five thousand different data files for human fecal material across the life cycle. Their results provide insight into how the type of microbes in the gut microbiome change as humans age.

4. Developments in compiling, accessing, and using genomic data

With the accelerating increase in data volumes, accurate annotation of these genomic data also becomes more and more challenging (see van Santen *et al.*, DOI: 10.1039/D0NP00053A and Caesar *et al.*, DOI: 10.1039/D1NP00036E). While the functions of genes in the first sequenced genomes were largely annotated and carefully curated by hand, this has of course become completely infeasible for the many thousands of genomes being sequenced. The result is that all annotations are generated using automated pipelines and usually provide a generic clue at best. These pipelines often simply copy the annotation of the closest match in the database, and if this happens time after time, the relationship with any experimentally characterized gene often becomes very distant. Moreover, this procedure is very prone to propagation of errors. Manual curation of experimentally characterized proteins, genes and BGCs in dedicated databases therefore remains crucial, but even these can produce errors: for example, during the construction of version 2 of the MIBiG repository for experimentally characterized BGCs,[19] dozens of structures were corrected from version 1 that had been incorrectly assigned by annotators or that contained errors.

But even the primary data cannot always be trusted. For pragmatic reasons, volume is often preferred above quality, especially for targeted 'screening' approaches. Therefore, many highly fragmented draft genomes are found in the databases and many genomes contain misassembled BGCs. These may easily give false impressions of biosynthetic diversity that is not truly there and may lead to faulty hypotheses being generated. Deposition of the raw data should therefore be more strongly encouraged (or even demanded), as seemingly interesting variations can then be reassessed by, *e.g.*, repeating the assembly before investing in expensive and time-consuming experiments.

The sizes of omics datasets used as the basis for natural product discovery have increased by multiple orders of magnitude over the past decade. Whereas ten years ago, using 10–20 genome sequences as the basis for a natural product genome mining project was still revolutionary, many thousands are often used these days. A case in point is the effort by Warp Drive Bio to identify new rapamycin analogues in a collection of ~135 000 actinobacterial draft genomes.[20] The rationale for this search stemmed from the fact that rapamycin and the related metabolite FK506 were known to bind two different targets. In both cases, binding is mediated by a conserved structural moiety that binds the FKBP12 protein, which then helps binding to the target through protein–protein interactions. The authors hypothesized that, within this class of polyketides, many additional biosynthetic pathways might have evolved to bind a range of other protein targets with the aid of FKBP12. They therefore performed low-coverage sequencing of thousands of actinobacterial genomes to scan for the presence of the lysine cyclodeaminase gene, which is involved in the biosynthesis of pipecolate, a key structural component of the part of both FK506 and rapamycin that binds the FKBP12 active site. Although these genome assemblies were undoubtedly noisy and will not have contained many full-length BGCs for the production of rapamycin analogues, this allowed the effective prioritization of strains that might contain them. All strains with hits to this gene were subjected to complete genome sequencing and assembly to reveal the presence of BGCs potentially encoding the production of new rapamycin analogues. In the end, the team found five BGCs with novel architectures, and were able to identify a new natural product that targets human centrosomal protein 250 (CEP250), a protein that had been thought to be 'undruggable' due to its flat surface. While only

finding one new BGC of this type in every ~30 000 genomes may be perceived as disappointing, the study did show that targeted screening can be used to find needles in a big data haystack that may be useful starting points for drug discovery. In a similar manner, metagenomic screening efforts (discussed in-depth by Robinson *et al.*, DOI: 10.1039/D1NP00006C) have been used to unearth new calcium-dependent lipopeptide antibiotics.[21,22]

Untargeted approaches constitute another way in which genomic big data can be utilized, *e.g.* to chart extant natural product diversity and guide discovery efforts to the most promising taxa and BGCs to avoid rediscovery and target natural products with relevant activities (also discussed in Chevrette and Handelsman, DOI: 10.1039/D1NP00044F). For example, an algorithm called BiG-SLiCE was recently used to analyze global biosynthetic diversity of ~1.2 million BGCs across >200 000 microbial genomes.[19] Converting BGC sequences into vectors of numerical features made it computationally feasible to identify relationships between gene clusters at a global scale, and to rapidly assign any given query BGC to a gene cluster family containing its known and unknown relatives. Moreover, this technology also enabled a study to quantitatively assess natural product biosynthetic diversity across the tree of life,[23] which suggests that only ~3% of genomically encoded natural product classes have been discovered thus far, and that highly studied taxa like *Streptomyces* still harbor many yet-unknown natural products. Notably, this is not necessarily a guarantee that mining the data for these unknown natural products will also yield many new drugs like urgently needed antibiotics, as these might already have been oversampled by genome-independent discovery approaches in previous decades that primarily screened for biological activities of interest in culture extracts.[24]

5. The future of big data in the natural product sciences

What does the future hold for natural products research? The articles contained in this themed issue point to many tantalizing possibilities. Among these is that as the quantity and quality of data generated about natural products expand, so does the potential of applying artificial intelligence analytical approaches (such as machine learning) to advance our field. Recent reports suggest the promise of such approaches (see also Jeon *et al.*, DOI: 10.1039/D1NP00016K).[25,26] Successful application of machine learning methodologies requires high-quality training data, and we are currently held back by a lack of curated data sets of all types – chemical structures, spectroscopic data, protein and gene sequence data, and various kinds of biological assessments. Also lacking are comprehensive databases linking structure or gene cluster to function in a useful way, although this is an area of rapid development (Bauman *et al.*, DOI: 10.1039/D1NP00032B; Caesar *et al.*, DOI: 10.1039/D1NP00036E). There are still significant hurdles that must be overcome in terms of the quality and accessibility of all types of big data. Accumulation of low-quality data and pervasive errors in data annotation may severely hamper efforts to benefit from these big data with artificial intelligence and machine learning approaches. Hence, investing in carefully vetted and well-standardized datasets should receive priority in the future. As the required data sizes are beyond the capabilities of individual laboratories, this will require cross-laboratory and ideally coordinated international efforts to generate datasets in standardized ways and curate them according to standardized protocols.

As we look forward to the advances that may be enabled by the big data revolution, it is tempting to discuss them in opposition to the technologies that characterized our past. Indeed, it is common to hear the leading researchers in our field dismiss projects that rely on bioassay-guided fractionation out of hand, speaking in disparaging terms about 'grind and find' science. We contend that it is worthwhile to view the foundational skills of natural products discovery – isolation and purification – not in opposition to big data approaches, but as complementary to them. Chemists skilled at isolation and structure elucidation will always be a valuable part of the natural products research team, because all predictions need to be validated by ground truth. Any use of machine learning algorithms to utilize large datasets depends strongly on reliable training data on chemical structures, enzyme functions and biological activities, which, regardless of exciting technological developments, still must be produced the hard way. Moreover, big data driven approaches are by nature hypothesis generating, and studies with purified material are critical to validate predictions about structure and activity. It is humbling to note that while this themed issue is filled with numerous tantalizing vignettes about what is currently possible thanks to big data or what may be possible in the future, we cannot yet point to a single drug that has been discovered using exclusively big data approaches. Perhaps this is simply because the contributions of big data to the field of natural products drug discovery are too young to pan out in concrete ways. It is also possible that finding a clinically useful new drug is too high a bar to set. If that is indeed the case, the question remains, what should be the litmus test for a truly successful natural products research project? If the goal is not drug discovery, we may need to rethink the narrative we use to sell the value of our research endeavors.

The optimists among us believe that by engaging with big data we are developing the tools today that will enable discovery of the drugs of tomorrow. It may also be true that these tools do not live up to our hopes for drug discovery, but by adopting them that we are shifting the focus of our field in new (and perhaps even more exciting) directions. Some would say that this is already happening. Regardless, it is obvious that big data approaches have irrevocably altered the landscape for natural products researchers, and that we will continue to engage with big data in the future. We expect that such engagements will deepen our understanding of life on our planet, and we hope that the fruits of these labors will increasingly be shared equitably to improve the quality of life for those who inhabit it.

6. Funding

NBC receives funding from the Center for High Content Functional Annotation of Natural Products (HiFAN) under grant number U41AT008718 from the National Center for

Complementary and Integrative Health (NCCIH) and the Office of Dietary Supplements (ODS), components of the U.S. National Institutes of Health. MHM is supported through an ERC Starting Grant [948 770-DECIPHER].

7. Conflicts of interest

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

8. Acknowledgments

We thank Ashley Scott for graphical art assistance with the table of contents graphic.

9. References

1 A. Fleming, *Br. J. Exp. Pathol.*, 1929, **10**, 226–236.
2 M. C. Wani, H. L. Taylor, M. E. Wall, P. Coggon and A. T. McPhail, *J. Am. Chem. Soc.*, 1971, **93**, 2325–2327.
3 Y. Tu, *Angew. Chem., Int. Ed.*, 2016, **55**, 10210–10226.
4 F. Malpartida and D. A. Hopwood, *Nature*, 1984, **309**, 462–464.
5 J. J. J. van der Hooft, H. Mohimani, A. Bauermeister, P. C. Dorrestein, K. R. Duncan and M. H. Medema, *Chem. Soc. Rev.*, 2020, **49**, 3297–3314.
6 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.
7 M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik and C. Steinbeck, *J. Cheminf.*, 2021, **13**, 2.
8 A. Rutz, M. Sorokina, J. Galgonek, D. Mietchen, E. Willighagen, J. Graham, R. Stephan, R. Page, J. Vondrášek, C. Steinbeck, G. F. Pauli, J.-L. Wolfender, J. Bisson and P.-M. Allard, *bioRxiv*, 2021, DOI: 10.1101/2021.02.28.433265.
9 M. T. Robey, L. K. Caesar, M. T. Drott, N. P. Keller and N. L. Kelleher, *Proc. Natl. Acad. Sci. U. S. A.*, 2021, **118**, e2020230118.
10 J. B. McAlpine, S.-N. Chen, A. Kutateladze, J. B. MacMillan, G. Appendino, A. Barison, M. A. Beniddir, M. W. Biavatti, S. Bluml, A. Boufridi, M. S. Butler, R. J. Capon, Y. H. Choi, D. Coppage, P. Crews, M. T. Crimmins, M. Csete, P. Dewapriya, J. M. Egan, M. J. Garson, G. Genta-Jouve, W. H. Gerwick, H. Gross, M. K. Harper, P. Hermanto, J. M. Hook, L. Hunter, D. Jeannerat, N.-Y. Ji, T. A. Johnson, D. G. I. Kingston, H. Koshino, H.-W. Lee, G. Lewin, J. Li, R. G. Linington, M. Liu, K. L. McPhail, T. F. Molinski, B. S. Moore, J.-W. Nam, R. P. Neupane, M. Niemitz, J.-M. Nuzillard, N. H. Oberlies, F. M. M. Ocampos, G. Pan, R. J. Quinn, D. S. Reddy, J.-H. Renault, J. Rivera-Chávez, W. Robien, C. M. Saunders, T. J. Schmidt, C. Seger, B. Shen, C. Steinbeck, H. Stuppner, S. Sturm, O. Taglialatela-Scafati, D. J. Tantillo, R. Verpoorte, B.-G. Wang, C. M. Williams, P. G. Williams, J. Wist, J.-M. Yue, C. Zhang, Z. Xu, C. Simmler, D. C. Lankin, J. Bisson and G. F. Pauli, *Nat. Prod. Rep.*, 2019, **36**, 35–107.
11 I. Blaženović, T. Kind, J. Ji and O. Fiehn, *Metabolites*, 2018, **8**, 31.
12 T. N. Clark, J. Houriet, W. S. Vidar, J. J. Kellogg, D. A. Todd, N. B. Cech and R. G. Linington, *J. Nat. Prod.*, 2021, **84**, 824–835.
13 N. G. Mahieu and G. J. Patti, *Anal. Chem.*, 2017, **89**, 10397–10406.
14 L. K. Caesar, O. M. Kvalheim and N. B. Cech, *Anal. Chim. Acta*, 2018, **1021**, 69–77.
15 B. O. Keller, J. Sui, A. B. Young and R. M. Whittal, *Anal. Chim. Acta*, 2008, **627**, 71–81.
16 C. D. Broeckling, F. A. Afsar, S. Neumann, A. Ben-Hur and J. E. Prenni, *Anal. Chem.*, 2014, **86**, 6812–6817.
17 A. T. Aron, E. C. Gentry, K. L. McPhail, L. F. Nothias, M. Nothias-Esposito, A. Bouslimani, D. Petras, J. M. Gauglitz, N. Sikora, F. Vargas, J. J. J. van der Hooft, M. Ernst, K. B. Kang, C. M. Aceves, A. M. Caraballo-Rodríguez, I. Koester, K. C. Weldon, S. Bertrand, C. Roullier, K. Sun, R. M. Tehan, P. C. Boya, M. H. Christian, M. Gutiérrez, A. M. Ulloa, J. A. Tejeda Mora, R. Mojica-Flores, J. Lakey-Beitia, V. Vásquez-Chaves, Y. Zhang, A. I. Calderón, N. Tayler, R. A. Keyzers, F. Tugizimana, N. Ndlovu, A. A. Aksenov, A. K. Jarmusch, R. Schmid, A. W. Truman, N. Bandeira, M. Wang and P. C. Dorrestein, *Nat. Protoc.*, 2020, **15**, 1954–1991.
18 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. Caraballo 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. Montoya 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.
19 S. A. Kautsar, J. J. J. van der Hooft, D. de Ridder and M. H. Medema, *GigaScience*, 2021, **10**, giaa154.
20 U. K. Shigdel, S.-J. Lee, M. E. Sowa, B. R. Bowman, K. Robison, M. Zhou, K. H. Pua, D. T. Stiles, J. A. V. Blodgett, D. W. Udwary, A. T. Rajczewski, A. S. Mann, S. Mostafavi, T. Hardy, S. Arya, Z. Weng, M. Stewart, K. Kenyon, J. P. Morgenstern, E. Pan, D. C. Gray, R. M. Pollock, A. M. Fry, R. D. Klausner, S. A. Townson and G. L. Verdine, *Proc. Natl. Acad. Sci. U. S. A.*, 2020, **117**, 17195.

21 B. M. Hover, S.-H. Kim, M. Katz, Z. Charlop-Powers, J. G. Owen, M. A. Ternei, J. Maniko, A. B. Estrela, H. Molina, S. Park, D. S. Perlin and S. F. Brady, *Nat. Microbiol.*, 2018, **3**, 415–422.
22 C. Wu, Z. Shang, C. Lemetre, M. A. Ternei and S. F. Brady, *J. Am. Chem. Soc.*, 2019, **141**, 3910–3919.
23 A. Gavriilidou, S. Kautsar, N. Zaburannyi, D. Krug, R. Müller, M. Medema and N. Ziemert, *bioRxiv*, 2021, DOI: 10.1101/2021.08.11.455920.
24 K. Lewis, *Cell*, 2020, **181**, 29–45.
25 S. H. Martínez-Treviño, V. Uc-Cetina, M. A. Fernández-Herrera and G. Merino, *J. Chem. Inf. Model.*, 2020, **60**, 3376–3386.
26 A. S. Walker and J. Clardy, *J. Chem. Inf. Model.*, 2021, **61**, 2560–2571.