

Showcasing research from Professor Lihan Zhang's laboratory, Department of Chemistry, Westlake University, Hangzhou, China.

A metabologenomics strategy for rapid discovery of polyketides derived from modular polyketide synthases

Polyketides, such as the antibiotic erythromycin, are a prominent class of natural products with diverse bioactivities. Current genome information of microbes has revealed vast untapped polyketides from nature. To rapidly discover these compounds, we developed a metabologenomics strategy that combines mass defect filtering with bioinformatic analysis. This approach was successfully applied to characterize novel polyketides from a well-studied *Streptomyces* strain, highlighting the potential for uncovering hidden metabolic diversity.

### As featured in:



See Lihan Zhang *et al.*,  
*Chem. Sci.*, 2025, **16**, 1696.

Cite this: *Chem. Sci.*, 2025, **16**, 1696 All publication charges for this article have been paid for by the Royal Society of ChemistryReceived 24th June 2024  
Accepted 1st November 2024DOI: 10.1039/d4sc04174g  
[rsc.li/chemical-science](http://rsc.li/chemical-science)

## Introduction

Natural products are the primary source of clinically important therapeutic agents, agrochemicals, and various functional molecules.<sup>1,2</sup> Numerous bioactive natural products from plants, fungi, and bacteria have been discovered by traditional activity-guided isolation.<sup>3</sup> While chemical extraction-based investigations, often targeted at a few talented genera such as *Streptomyces* and *Penicillium*, have faced challenges of frequent rediscovery of known compounds,<sup>4</sup> recent advances in genome sequencing technology have uncovered a significant number of uncharacterized biosynthetic gene clusters (BGCs) not only from previously understudied taxa but also from well-characterized strains.<sup>5–7</sup> Hence, genome mining has become a promising approach to fully exploit microbial natural products for the discovery of new functional molecules.<sup>8–12</sup> However, characterization of the products of a target gene cluster often requires time-consuming genetic manipulations,<sup>10,13</sup> hindering

the rapid identification of natural products. As a result, the pace of natural product discovery has struggled to keep up with the ever-growing wealth of genome sequence information.

To achieve high-throughput identification of natural products, metabolomics has emerged as a powerful tool, supported by the advances in increased resolution and sensitivity of mass spectrometry (MS) instruments.<sup>5,14–16</sup> Metabolomic analyses by liquid chromatography-mass spectrometry (LC-MS) coupled with tandem mass spectrometry (MS/MS) have proven effective in the rapid dereplication of known molecules and their derivatives, as exemplified by several tools such as GNPS<sup>17</sup> and SIRIUS.<sup>18</sup> In addition, the growing knowledge of natural product biosynthesis has facilitated the integration of genomics with metabolomics, known as metabologenomics,<sup>19,20</sup> which aims to characterize metabolites based on the BGC information or conversely connect metabolites with their BGCs. Metabologenomics is particularly effective for natural products with well-characterized biosynthetic pathways and has demonstrated its application in the discovery of peptides<sup>21–23</sup> and glycosides,<sup>24</sup> which exhibit clear MS/MS fragmentation patterns that correlate mass data with structural information.

Polyketides are a major class of natural products with remarkable structural diversity, including clinical medicines such as erythromycin and rapamycin. Biosynthetically, these polyketides are derived from modular type I polyketide synthases (PKSs), which operate in an assembly line-like mechanism.<sup>25,26</sup> This modular, ordered biosynthetic mechanism allows the fine prediction of polyketide scaffold structures by

<sup>a</sup>Department of Chemistry, Fudan University, Shanghai 200433, China

<sup>b</sup>Key Laboratory of Precise Synthesis of Functional Molecules of Zhejiang Province, Department of Chemistry, School of Science and Research Center for Industries of the Future, Westlake University, Hangzhou 310030, China

<sup>c</sup>Institute of Natural Sciences, Westlake Institute for Advanced Study, Hangzhou 310024, China

<sup>d</sup>Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou 310030, China.  
E-mail: [zhanglihan@westlake.edu.cn](mailto:zhanglihan@westlake.edu.cn)

† Electronic supplementary information (ESI) available. See DOI: <https://doi.org/10.1039/d4sc04174g>



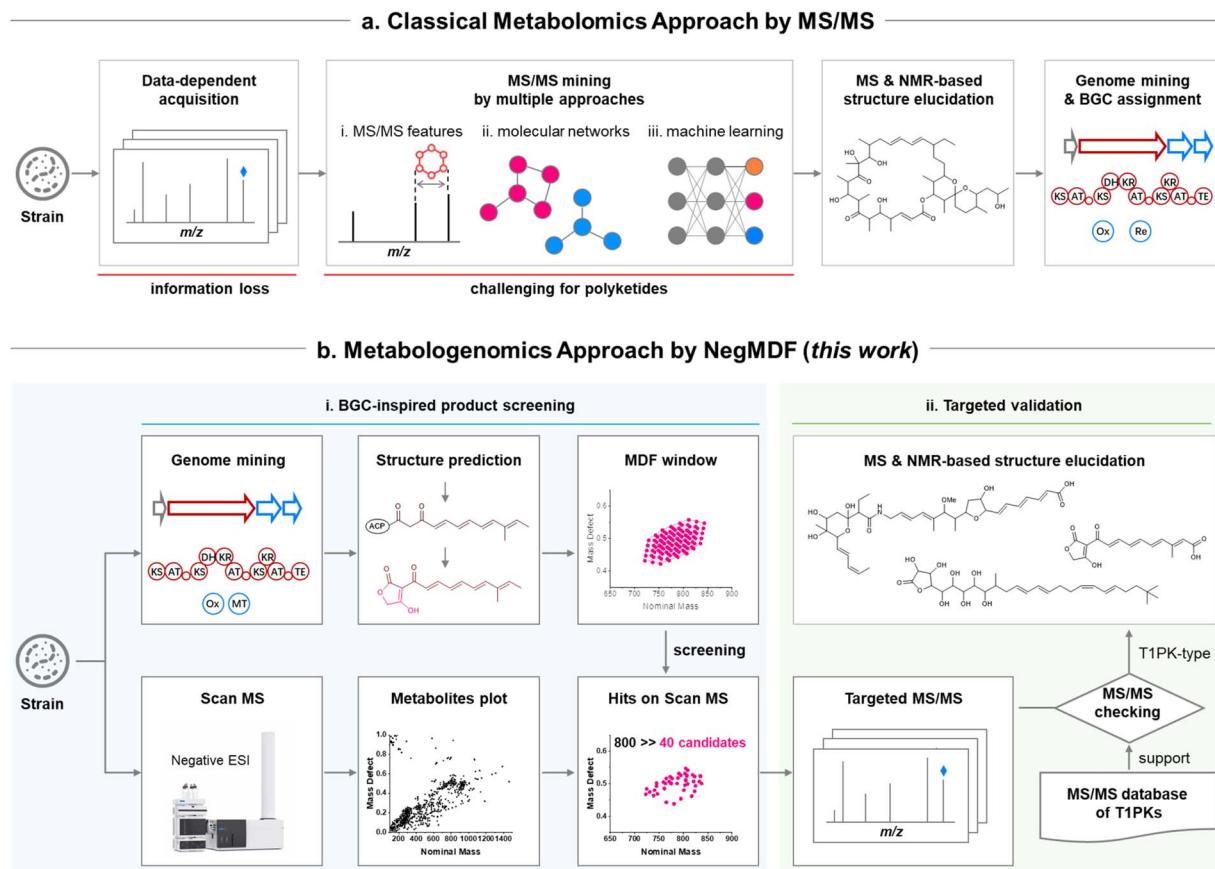


Fig. 1 The schemes for discovery of bacterial T1PKs guided by MS. (a) Classic metabolomics approach by MS/MS-based screening. (b) The metabogenomics approach developed in this work, which employs BGC-guided mass defect filtering (MDF) under negative scan MS as the initial screening phase, followed by targeted validation by MS/MS-based polyketide identification.

bioinformatic analysis, making assembly line polyketides an attractive target for genome mining.<sup>27,28</sup> However, despite their well-characterized biosynthesis, metabogenomics has not been widely applied to modular PKSs due to the absence of regular MS/MS fragmentation patterns exhibited by the product polyketides.<sup>29,30</sup> As such, the lack of general principles for MS/MS analysis of polyketides has hampered the rapid discovery of these compounds through metabolomics.

In this study, we present a novel metabogenomics approach, coined as NegMDF strategy, for the efficient characterization of assembly line polyketides (Fig. 1). We demonstrate that mass defect filtering (MDF)<sup>31–33</sup> combined with bioinformatic structural prediction is an effective strategy as a MS/MS-independent screening step for the discovery of polyketides. In addition, we reveal that negative mode of electrospray ionization (ESI) is more suitable for polyketide detection. As a proof-of-concept, we characterized 22 type I PKS-derived polyketides from a single microbe, *Streptomyces cattleya* NRRL 8057, from which we discovered cattleyatetronates, a new class of rare tetrone-containing polyketides. This study establishes an MDF-based metabogenomics workflow for the rapid screening of microbial natural products and will accelerate the genome mining of assembly line polyketides.

## Results

### The mass defect filtering of bacterial natural products

To establish a metabolomics strategy for the discovery of type I PKS-derived polyketides (T1PKs; containing polyenes, polyols, macrolides, macrolactams, polyethers, etc.), a method called mass defect filtering (MDF)<sup>31</sup> caught our attention due to its ease of use and high flexibility. MDF uses mass defect, the difference between the exact mass of a target molecule and its nominal integer mass, to analyze product ions by taking advantage of a molecule's unique atomic mass.<sup>32,33</sup> Typically, the mass defect of analyte ions can be plotted over their nominal integer mass to give a distribution pattern of product ions in a metabolomics sample (Fig. 2), which can further be filtered by specifying a mass defect range on the plot distribution. This method has multiple advantages for the analysis of T1PKs. Firstly, it does not rely on MS/MS fragmentation, and secondly, it can utilize the full information of primary mass scan results. This is in contrast to general MS/MS analysis that is commonly performed under data-dependent acquisition, which only picks up abundant ions for fragmentation.

Although MDF has been applied in the analyses of drug metabolism<sup>34</sup> and several specific classes of plant natural products,<sup>35,36</sup> its application in bacterial natural products



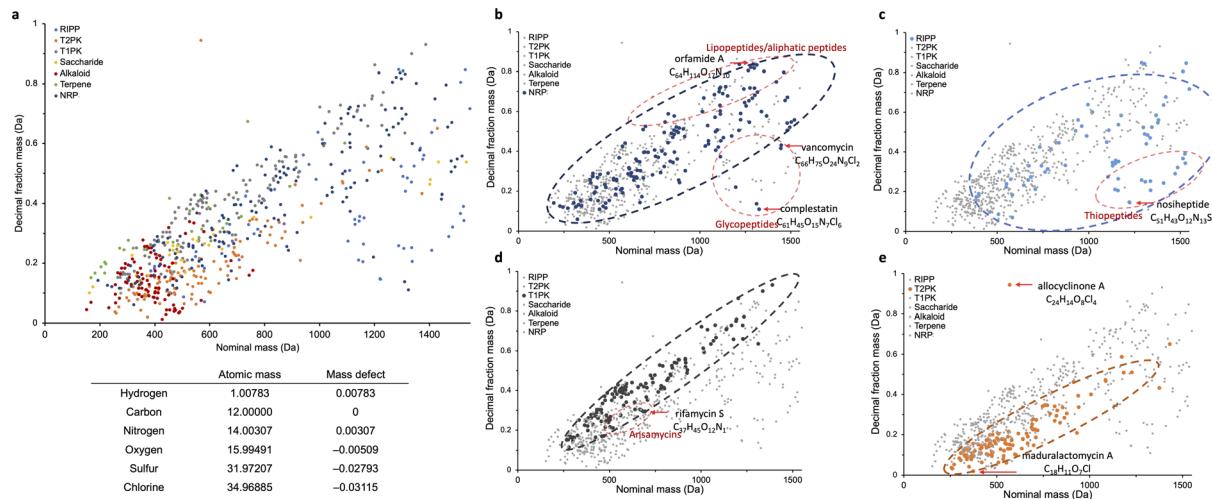


Fig. 2 Mass defect plotting of bacterial natural products. RIPP: ribosomally synthesized and posttranslationally modified peptides; T2PK: type II PKS-derived polyketides; T1PK: type I PKS-derived polyketides; NRP: nonribosomal peptides. (a) Summary of bacterial natural products based on biosynthetic classes. Representative atoms and their mass defect contributions were provided. (b) The plot of NRPs. (c) The plot of RPPs. (d) The plot of T1PKs. (e) The plot of T2PKs.

remains unexplored. To demonstrate its applicability, we first analyzed the mass defect distribution of over 800 diverse bacterial natural products from MIBiG database<sup>37</sup> according to their biosynthetic classes. The mass defect plot showed that each class of natural products has its own distribution pattern (Fig. 2). For example, T1PKs has higher mass defect than polyketides produced by type II PKS (T2PKs, mostly consisting of aromatic polyketides), highlighting their differences in unsaturation degree and atom composition. In addition, the MDF plot can further indicate structural features of compounds within the same class. For instance, in nonribosomal peptides (NRPs), compounds rich in aliphatic carbons such as lipopeptides appear at higher end of the distribution due to having more hydrogen atoms, but compounds rich in aromatic rings and chlorines such as glycopeptides appear at the lower end due to higher unsaturation degree (*i.e.* less hydrogens) and having atoms with large negative defect. Although MDF is logically simple, our analysis illustrates its utility in rapidly visualizing complex metabolomics data with chemical composition information, which otherwise requires highly technical knowledge and skills to realize.

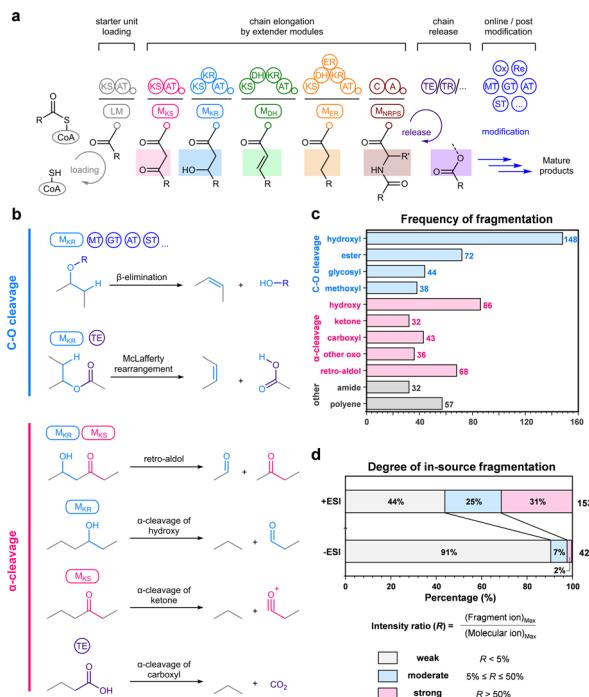
Determining an appropriate window size for filtering is critical for MDF analysis, and bioinformatic prediction of the products gives an ideal starting point for MDF screening. We developed a workflow for the structural prediction of T1PKs based on bioinformatic analyses (ESI Fig. 1†) and here summarize the process as follows. First, PKS gene clusters in the genome of a target strain are identified by BGC prediction tools such as antiSMASH<sup>38</sup> and PRISM.<sup>39</sup> If a BGC has high similarity to a reported BGC, then the exact mass information of the known compound can be directly used for filtering. Otherwise, the predicted structure is established by three steps: (i) the “head” and “tail” of the polyketide scaffold can be predicted based on the phylogenetic clades of the first elongating KS domains,<sup>40</sup> and by the type of off-loading domains,<sup>41</sup>

respectively; (ii) the extender unit used in polyketide biosynthesis, commonly being malonyl-CoA or methylmalonyl-CoA, can further be predicted by analyzing the types of acyltransferase (AT);<sup>42</sup> (iii) post-PKS modifications are analyzed by taking their tailoring enzymes into account, such as methyltransferase (MT), glycosyltransferase (GT), sulfotransferase (ST), and redox enzymes, to predict functional group modifications. After predicting the core structure of a target BGC, we introduce potential variants to proof deviation from prediction and to enable discovery of compounds with unknown modifications. This variation can include, for example, the chain release by macrocyclization or by hydrolysis, and the presence and absence of further oxidations and dehydrations. As a result, a series of molecular formulas can be obtained for a BGC of interest, serving as the MDF window for the product ion screening (Fig. 1b).

### MS/MS features of bacterial T1PKs

Although the MDF-based metabolomics screening enables rapid prioritization of product ions, validation of the filtered ions by a targeted MS/MS analysis is necessary, because different classes of natural products may overlap on the mass defect plot. However, a general guideline for MS/MS interpretation of polyketides on ESI-MS has not been well-established. To understand common features of the fragmentation of T1PKs, we collected MS/MS spectra of 222 polyketide compounds mostly under ESI-MS from references (r1–r222, ESI Dataset 1†). Importantly, these polyketides also have their biosynthetic gene cluster information available, allowing us to summarize key fragmentation patterns of T1PKs linked to the genomic features (Fig. 3).

The most frequent fragmentation in T1PKs is the cleavage of carbon–oxygen bond (C–O cleavage), commonly achieved through a  $\beta$ -elimination mechanism (Fig. 3b and c). PKSs are



**Fig. 3** The typical fragmentation reactions for bacterial T1PKs. (a) The representative modules, domains and tailoring enzymes observed in the biosynthesis of T1PKs. (b) The common fragmentation patterns of C–O cleavage and  $\alpha$ -cleavage by collision-induced dissociation (CID). (c) The frequency of representative CID-reactions observed in the MS/MS spectra of 230 T1PKs (other oxo-groups include epoxy, ether and methoxy). (d) The frequency of in-source fragmentation observed in 195 scan-MS spectra. The degree of fragmentation was calculated by comparing the intensity of the most abundant assignable fragment ion (e.g.  $[M-H_2O + H]^+$ ,  $[M-MeOH + H]^+$ , or  $[M-glycosyl + H]^+$ ) to that of the most abundant molecular ion (e.g.  $[M + H]^+$ ,  $[M + Na]^+$ , or  $[M - H]^-$ ).

often rich in secondary alcohols biosynthesized by the ketoreductase-containing modules ( $M_{KR}$ ), and sequential dehydration reactions are frequently observed in their MS/MS analyses.<sup>43</sup> Besides, these hydroxyl groups can further be modified by tailoring enzymes such as MT, GT, and ST, leading to the corresponding diagnostic mass shifts (ESI Fig. 2†). For ester bond-containing polyketides, cleavage of C–O bond can proceed via McLafferty rearrangement to yield the carboxylic acid fragment.

In addition, fragments derived from  $\alpha$ -cleavage of a carbonyl or a hydroxyl group are frequently observed (Fig. 3 and ESI Fig. 3†). Retro-aldol cleavage is particularly common in compounds with a 3-hydroxyl ketone moiety (ESI Fig. 4†), and these  $\alpha$ -cleavages can be informative to elucidate the main carbon scaffold of T1PKs. Other types of fragments are also observable in the MS/MS patterns of T1PKs, such as the allyl cleavages of polyene moieties derived from dehydratase-containing PKS modules ( $M_{DH}$ ) (ESI Fig. 5†), though the ion intensity of them is usually low. The amide bond cleavage can also be observed when a BGC contains nonribosomal peptide synthase module ( $M_{NRPS}$ ). Overall, the C–O cleavage and the  $\alpha$ -cleavage of a C–C bond next to a C–O bond account for most fragmentations in the MS/MS analyses of typical T1PKs (Fig. 3c).

To validate these observations, we tested MS/MS fragmentation pattern of polyketide standards (s1–s8) by self-acquisition. The tested polyketides showed clear fragments which can be assigned by C–O cleavages and  $\alpha$ -cleavages (ESI Dataset 1†). Besides, we found that fragment ions of  $m/z = 59.0139$  and  $73.0295$  were commonly observed with varied intensity under negative ESI (ESI Fig. 6†), serving as diagnostic fragments for T1PKs with an acetyl/propionyl esters or a 3-hydroxyl acid/ester moiety. Overall, these analyses reveal general tendencies of the fragmentation of T1PKs, which can aid identification of T1PKs by MS/MS analysis.

### The in-source fragmentation of bacterial T1PKs in ESI-MS

In LC-ESI-MS analyses, polyketides tend to demonstrate sufficient ionization response in both positive and negative modes. However, despite ESI being considered as a soft ionization source, we noticed that the occurrence of in-source fragmentation is still prevalent, particularly under positive ionization mode, resulting in the depletion of the target ion during the primary mass scan. To quantify the tendency of in-source fragmentation, we assessed the primary MS scan spectra of 173 T1PKs from the references we curated (ESI Dataset 2†). As depicted in Fig. 3d, a substantial proportion of T1PKs exhibited strong (fragment ion shows  $>50\%$  intensity than molecular ion) to moderate (50–5%) in-source fragmentation in the positive ESI mode, while this ratio was less than 10% in the negative ESI mode. This tendency was also confirmed by our in-house experimental analysis of the standards (ESI Fig. 7†), which can be attributed to the difference between charge migration fragmentation and charge retention fragmentation of C–O cleavage (ESI Fig. 8†).<sup>44,45</sup>

Since MDF analysis uses molecular ion information from primary MS scan, it is important to suppress in-source fragmentation, which not only reduces the ion response of the targets, but also introduces disturbing fragment ions. Besides, positive ionization mode has higher background noises, as peptides and alkaloids tend to exhibit high ionization response due to the existence of basic nitrogen atoms. Thus, we propose the use of negative mode scan as the standard MS condition for the MDF analysis of T1PKs.

### Establishment of the NegMDF workflow

Based on the above analyses, we establish a metabologenomics workflow, coined as NegMDF, for rapid discovery of T1PKs (Fig. 1b). The NegMDF workflow consists of two phases. In the initial screening phase, the metabolomics data obtained under ESI (–) full scan MS is visualized by mass defect plotting, which is then filtered by bioinformatics-guided MDF window to obtain candidate ions. Next, these candidate ions are validated by targeted MS/MS analyses to identify T1PKs of a target BGC.

To illustrate the NegMDF workflow, we first applied it for genome mining of oligomycin from *S. avermitilis*, a reported oligomycin producer<sup>46</sup> for validation. First, we collected metabolomic data under negative ESI mode and obtained 877 total metabolite ions after subtracting background noises and media components. Next, the structure prediction was

performed based on the oligomycin BGC information, resulting in the predicted molecular formula of  $C_{44}H_{72}O_{10}$  with potential variations (Fig. 4a and ESI Fig. 9†). This MDF window included 61 ions, of which 42 were confirmed as oligomycin-type compounds upon MS/MS analyses that showed characteristic dehydration fragments and retro-aldol fragments predicted by the bioinformatic analysis (Fig. 4 and ESI Table 1†). We also analyzed avermectins, another class of T1PKs produced by *S. avermitilis* by the NegMDF workflow. A distinct cluster of avermectins was observed in the mass defect plot (Fig. 4d), showing no overlap with the oligomycin region. Other detectable compounds in ESI negative mode, including desferrioxamins, also exhibited significant differentiation from the polyketide regions. These results highlight that the NegMDF workflow can effectively identify T1PKs from a complex bacterial metabolomics sample.

### Investigating T1PKs from *S. cattleya* by NegMDF

Next, we tested the NegMDF workflow for polyketide mining of another *Streptomyces* strain, *S. cattleya* NRRL 8057, a famous

producer of thienamycin,<sup>47</sup> cephamycin C,<sup>48</sup> and fluorinated antibiotics.<sup>49</sup> Among 39 regions of biosynthetic gene clusters predicted by antiSMASH, four of them contained type I PKSs with over 5 modules (ESI Table 2†), but only one type of T1PK, L-681,217 (**1a**)<sup>50</sup> and demethyl L-681,217 (**1b**)<sup>51</sup> was reported from *S. cattleya* (by PKS Region 1.15) despite more than 40 years of research history for this strain. Region 2.6 (but BGC) showed highly similar organization to butyrolactol BGC (BGC *orf10*),<sup>52</sup> indicating butyrolactol-like metabolites hidden in *S. cattleya*. To demonstrate dereplication of these known polyketides and to find more congeners from the metabolomics data, we designed two MDF windows based on the molecular formula of L-681,217 and butyrolactol A (Fig. 5a, see ESI Fig. 10 and 11† for details).

In total three culture extracts of ISP2, A3M, and MSF media were analyzed by the NegMDF workflow. By MDF window for the Region 1.15, we obtained 27, 32, and 17 candidate ions from the three cultures respectively (Fig. 5b). Subsequent targeted MS/MS analyses indicated that 19, 27 and 14 hits from the three media, respectively, showed a similar fragmentation pattern with L-681,217. Accordingly, we assigned 13 structures of L-

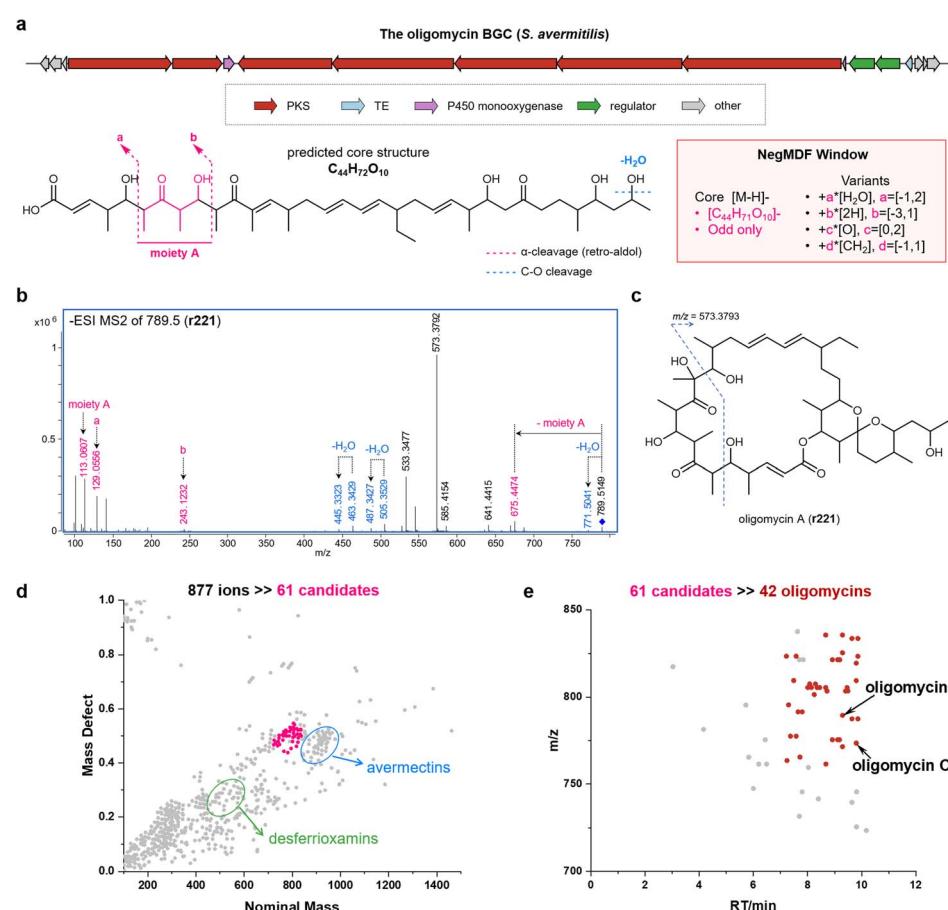
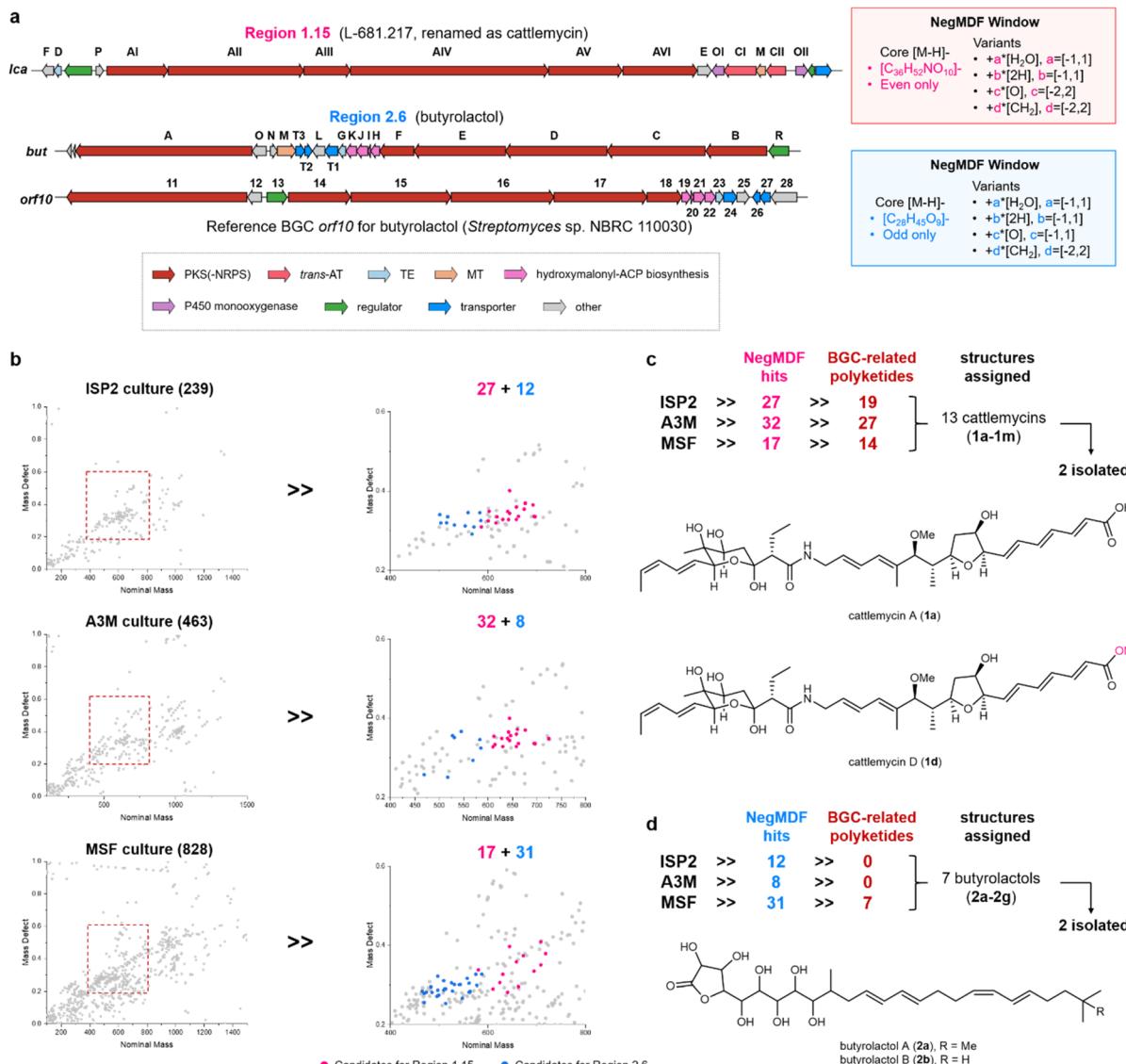


Fig. 4 The NegMDF workflow established for genome mining of type I polyketide synthetase with oligomycin as an example. (a) Genetic organization of oligomycin BGC from *S. avermitilis*. NegMDF parameters were designed based on the predicted polyketide and potential variants. (b) The MS/MS pattern of oligomycin A. Fragments assignable based on the bioinformatic prediction were labeled. (c) The structure of oligomycin A. (d) The metabolomics of *S. avermitilis* on negative ESI was obtained and pretreated by MZmine3. In total 877 ions were obtained from MSF media cultivation extract after subtracting background noise and media components, of which 61 ions located in the oligomycin window (pink dots). (e) Target MS/MS analysis confirmed 42 ions as oligomycins, including known products oligomycin A and C.





**Fig. 5** Dereplication and discovery of new polyketide congeners by NegMDF. (a) BGC organizations of Region 1.15 (*lca* BGC) for L-681.217 (cattlemycins) and Region 2.6 (*but* BGC) from *S. cattleya* as well as the known BGC (*orf10*) for butyrolactols (MIBiG ID: BGC0001537). NegMDF windows were designed based on the reported structure of L-681.217 and butyrolactols. (b) The MDF analyses of *S. cattleya* culture extracts from ISP2, A3M and MSF media. NegMDF hits are shown in pink for the *lca* BGC and in blue for the *but* BGC. (c) Targeted validation outcome and the confirmed structures of cattlemycins. Out of 76 potential hit ions from three culture conditions, 60 ions showed cattlemycin-like MS/MS patterns, of which 13 structures were proposed (**1a–m**), and two (**1a** and **1d**) were confirmed by NMR. (d) Targeted validation outcome and the confirmed structures of butyrolactols. Out of 51 potential hit ions from three culture conditions, 7 ions showed butyrolactol-like MS/MS patterns, whose structures were proposed (**2a–g**), and two (**2a** and **2b**) were confirmed by NMR.

681.217 and its derivatives (here renamed as cattlemycins A–M), including two knowns **1a** (cattlemycin A) and **1b** (cattlemycin B) and 11 new congeners (Fig. 5c and ESI Table 7†). The structure of a new congener, cattlemycin D (**1d**), was further determined to be a methylester derivative of **1a** by NMR analyses. The stereoconfiguration of **1** was proposed based on NOESY spectrum, bioinformatic analysis, and the comparison to kirromycin,<sup>53</sup> considering the high similarity between the two PKSs (ESI Table 10 and Fig. 18†). The biosynthesis of **1d** might involve an uncharacterized methyltransferase outside this BGC.<sup>54</sup>

By MDF window for the Region 2.6 (*but* BGC), we obtained 12, 8, and 31 candidate ions from the three cultures respectively (ESI Table 8). Targeted MS/MS validation found 7 true hits only from the MSF culture. Two major products, butyrolactols A and B (**2a** and **2b**) were verified by isolation and NMR analyses (Fig. 5d). We also identified three new shunt products, which are likely derived from the use of malonyl-ACP instead of hydroxymalonyl-ACP by the last module of *but* PKS (**2c**) and from the skipping of the last KR domain (**2d** and **2e**). Overall, these two examples illustrate the application of NegMDF

workflow in rapidly identifying T1PKs of a target gene cluster and in the discovery of minor congeners.

### Discovery of new polyketides, cattleyatetronates, by NegMDF

Encouraged by the above results, we explored the remaining two novel PKS gene clusters in *S. cattleya* by the NegMDF workflow. Although we were not able to characterize the product of the Region 1.3, which likely to be silent under our cultivation conditions, we successfully discovered the T1PK product of the Region 2.15, namely *ctt* BGC (Fig. 6a). The antiSMASH analysis found high similarity of *ctt*EFGHI to the chain release genes in tetroxasin BGC,<sup>55</sup> indicating a potential tetroxate moiety in our target product. Further analysis found adjacent tailoring enzymes including a cytochrome P450 (CttL), its redox partners (CttNK), a FAD-dependent monooxygenase (CttO), an oxidoreductase (CttQ), an aldehyde dehydrogenase (CttJ), and a methyltransferase (CttM). Based on the PKS module composition and potential modification by these enzymes, we designed

a NegMDF window (ESI Fig. 12†), and applied this filter to find target products in metabolomics of *S. cattleya*. From the MDF plot, we found two tetroxate-like ions, *m/z* 289.0717 and *m/z* 275.0923, out of 15 ions in the window range (Fig. 6b), whose predicted chemical formulas ( $C_{15}H_{14}O_6$  and  $C_{15}H_{16}O_5$ ) match the oxidized products of the predicted polyketide backbone. The MS/MS validation of these ions further revealed a major fragment with a mass shift of 43.9898 corresponding to decarboxylation and the fragments of *m/z* 99.0085 and *m/z* 125.9955 corresponding to tetroxate moiety from  $\alpha$ -cleavage<sup>44</sup> (Fig. 6c). Similar fragmentation ions were also observed in other reported tetroxate compounds r100–r103.

Notably, the two tetroxates were only expressed at low yields in the A3M medium and absent in other media (ESI Table 9†). The intensity of their ions was too low to be caught by data-dependent acquisition, and no significant ions can be identified in the scan MS of positive ESI mode (ESI Fig. 13†), highlighting the high sensitivity of our NegMDF workflow for T1PK discovery. By co-culture with *Tsukamurella pulmonis*, a reported

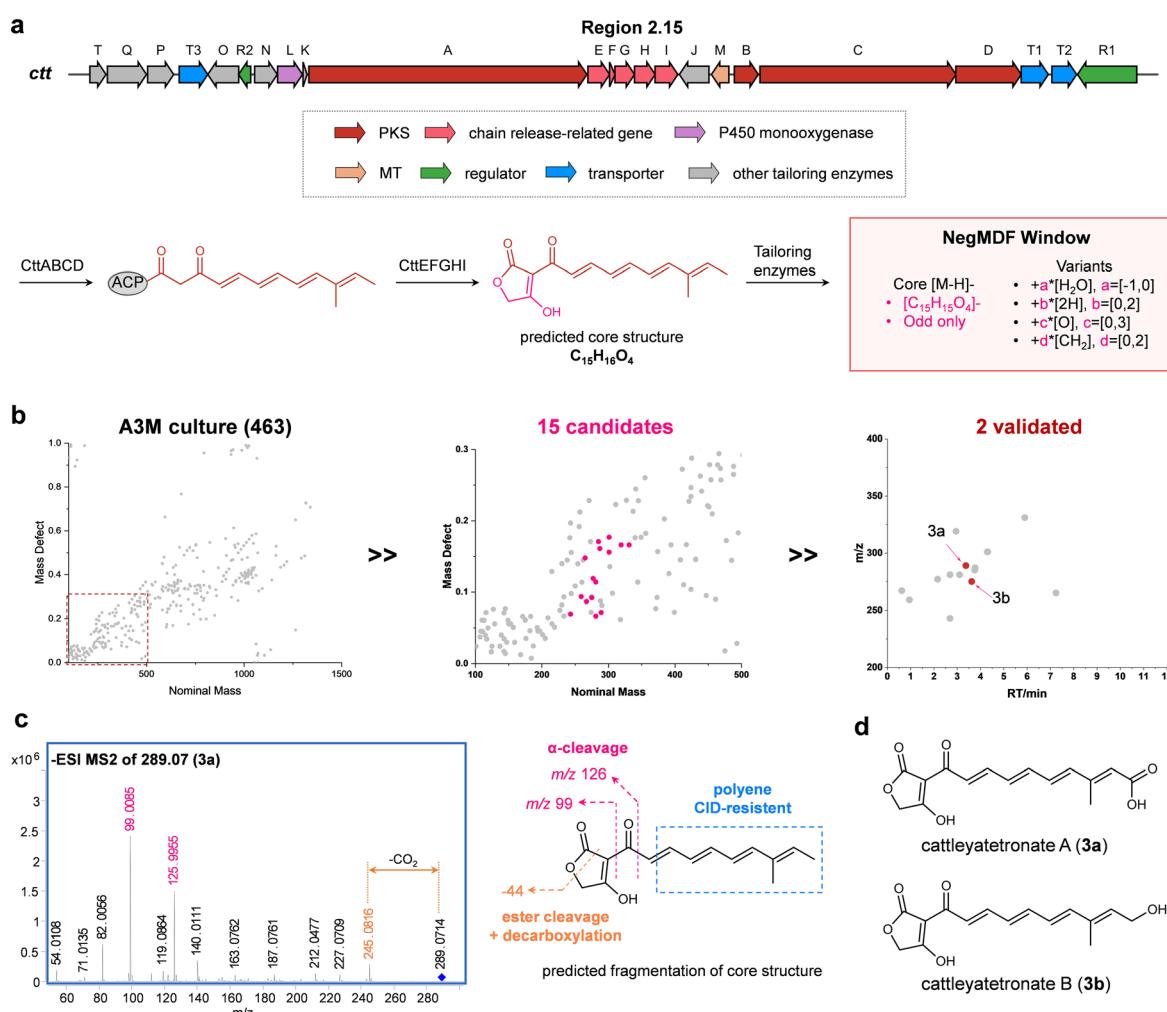


Fig. 6 Discovery of new polyketides, cattleyatetronates. (a) Genetic organization of Region 2.15 (*ctt* BGC) from *S. cattleya*. NegMDF window was designed based on the predicted polyketide core structure and tailoring enzymes. (b) The NegMDF screening of A3M culture extract. 15 candidates were obtained by the MDF window, of which 2 were validated to be *ctt* BGC products. (c) The MS/MS spectrum of **3a** and the assignable fragments based on the bioinformatic prediction. (d) The structures of cattleyatetronates A (**3a**) and B (**3b**).



inducer for *Streptomyces* metabolites,<sup>56</sup> we were able to improve their productivity, which enabled structure elucidation of cattleyatetronate A (**3a**) by NMR analysis (Fig. 6d). The P450 enzyme CttL may be responsible for the multistep oxidation of the terminal methyl group to generate the carboxyl group<sup>57</sup> *via* the hydroxyl intermediate cattleyatetronate B (**3b**).

Tetronate is an important moiety found in many polyketides, such as agglomerins, abyssomicins and chlorothricin. BGCs for this unit contain three conserved enzymes, FabH, ACP and FkbH, for loading of a C<sub>3</sub> glyceryl unit at the terminal of polyketides.<sup>58</sup> The exception is tetronasin BGC, which lacks the FkbH-like enzyme for glyceryl unit biosynthesis and instead loads a C<sub>2</sub> glycolyl unit for tetronate construction, though the exact mechanism for the formation of its tetronate ring remains elusive. Thus, the *ctt* BGC in *S. cattleya* represents, to our knowledge, the second example for C<sub>2</sub> glycolyl-containing tetronate, here categorized as type-B tetronate (ESI Fig. 14 and 15†).

## Conclusions

In recent years, genome sequencing has revealed a large number of BGCs harboring type I polyketide synthases, yet the vast majority of their products remain uncharacterized.<sup>59</sup> In this work, we developed a metabologenomics workflow of NegMDF that combines bioinformatic structural prediction, negative ESI-MS-based MDF analysis, and targeted MS/MS validation for rapid screening and identification of T1PKs. In total, 22 T1PKs were identified from a well-studied strain *S. cattleya*, of which 18 are unreported previously. Because the NegMDF workflow does not rely on MS/MS analyses at the initial screening phase and just provides one 2D plot per metabolomics sample, it allows sensitive and convenient screening of candidate ions of a BGC; this is in contrast to MS/MS-based metabolomics, which relies on data-dependent acquisition, leading to the loss of low-abundance ions, and often produces much complicated datasets when dealing with large number of samples.

The accuracy of our workflow depends on correct structure prediction of a PKS gene cluster. Although the carbon scaffold of a polyketide can be well-predicted based on the collinearity rule, it remains challenging to precisely predict the outcomes of post-PKS modification reactions and PKSs containing uncanonical or iterative domains.<sup>60</sup> To alleviate these shortcomings, the MDF window size can be adjusted to include more ions, though it also increases false positives and does not help analysing compounds with large mass deviation by unexpected modification. In such cases, other techniques such as isotope labelling<sup>61</sup> can be employed for further characterization. Nonetheless, MDF analysis requires neither expensive isotopes nor specific MS instrument, and can be implemented without technical difficulty. Because MS/MS spectra is often dependent with instruments under different ionization, dissociation, and detection methods, we envision that the mass defect plots may serve as a more compatible and economic format to store complex metabolomics data for large-scale analyses. It is tempting to assume that even unknown microbial extracts can be characterized by using mass defect plots as a fingerprint for chemotaxonomic classification.

Another key aspect of this study involves the thorough analysis of MS spectra from over 200 bacterial T1PKs, leading to a systematic understanding of their MS fragmentation. Although MS/MS fragmentation rules has been extensively studied under electron ionization (EI) conditions, the radical species produced by EI can exhibit different fragmentation behaviors than ionic species produced by ESI.<sup>45</sup> Thus, our systematic analysis of the fragmentation behavior of T1PKs under ESI conditions not only provides the basis for targeted product validation, but also facilitates the development of MS/MS analysis tools for the metabolomics of T1PKs.

In summary, we present the NegMDF metabologenomics pipeline for the characterization of polyketides derived from modular PKSs. We envision that MDF-based screening would be an efficient and convenient strategy to complement MS/MS-based metabolomics for the discovery of hidden microbial natural products.

## Data availability

All data supporting the results in this study are available within the paper and its ESI and ESI Datasets.†

## Author contributions

R.-Z. L. conducted all MS analyses and dataset collection. R.-Z. L. and Z. Z. performed product isolation and structural elucidation. M. L. prepared the Python script. R.-Z. L. performed the bioinformatic analyses. R.-Z. L. and L. Z. conceived the research, designed the experiments and wrote the paper.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

We thank the Instrumentation and Service Center for Molecular Sciences at Westlake University for the assistance in MS and NMR analyses. This work was supported by the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (2023SDXHDX0007) and National Natural Science Foundation of China General Program (22177092) to L.Z.

## Notes and references

- 1 A. G. Atanasov, S. B. Zotchev, V. M. Dirsch and C. T. Supuran, Natural products in drug discovery: advances and opportunities, *Nat. Rev. Drug Discovery*, 2021, **20**, 200–216.
- 2 D. J. Newman and G. M. Cragg, Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019, *J. Nat. Prod.*, 2020, **83**, 770–803.
- 3 L. Katz and R. H. Baltz, Natural product discovery: past, present, and future, *J. Ind. Microbiol. Biotechnol.*, 2016, **43**, 155–176.
- 4 J. Hubert, J.-M. Nuzillard and J.-H. Renault, Dereplication strategies in natural product research: How many tools and



methodologies behind the same concept?, *Phytochem. Rev.*, 2017, **16**, 55–95.

5 J. R. Doroghazi, J. C. Albright, A. W. Goering, K.-S. Ju, R. R. Haines, K. A. Tchalukov, D. P. Labeda, N. L. Kelleher and W. W. Metcalf, A roadmap for natural product discovery based on large-scale genomics and metabolomics, *Nat. Chem. Biol.*, 2014, **10**, 963–968.

6 P. Cimermancic, M. H. Medema, J. Claesen, K. Kurita, L. C. W. Brown, K. Mavrommatis, A. Pati, P. A. Godfrey, M. Koehrsen and J. Clardy, Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters, *Cell*, 2014, **158**, 412–421.

7 A. Gavrilidou, S. A. Kautsar, N. Zaburannyi, D. Krug, R. Müller, M. H. Medema and N. Ziemert, Compendium of specialized metabolite biosynthetic diversity encoded in bacterial genomes, *Nat. Microbiol.*, 2022, **7**, 726–735.

8 E. Kenshole, M. Herisse, M. Michael and S. J. Pidot, Natural product discovery through microbial genome mining, *Curr. Opin. Chem. Biol.*, 2021, **60**, 47–54.

9 E. Zazopoulos, K. Huang, A. Staffa, W. Liu, B. O. Bachmann, K. Nonaka, J. Ahlert, J. S. Thorson, B. Shen and C. M. Farnet, A genomics-guided approach for discovering and expressing cryptic metabolic pathways, *Nat. Biotechnol.*, 2003, **21**, 187–190.

10 K. Scherlach and C. Hertweck, Mining and unearthing hidden biosynthetic potential, *Nat. Commun.*, 2021, **12**, 3864.

11 M. H. Medema, T. de Rond and B. S. Moore, Mining genomes to illuminate the specialized chemistry of life, *Nat. Rev. Genet.*, 2021, **22**, 553–571.

12 N. Ziemert, M. Alanjary and T. Weber, The evolution of genome mining in microbes—a review, *Nat. Prod. Rep.*, 2016, **33**, 988–1005.

13 M. Zerikly and G. L. Challis, Strategies for the discovery of new natural products by genome mining, *ChemBioChem*, 2009, **10**, 625–633.

14 F. Xu, Y. Wu, C. Zhang, K. M. Davis, K. Moon, L. B. Bushin and M. R. Seyedsayamdst, A genetics-free method for high-throughput discovery of cryptic microbial metabolites, *Nat. Chem. Biol.*, 2019, **15**, 161–168.

15 B. C. Covington, J. A. McLean and B. O. Bachmann, Comparative mass spectrometry-based metabolomics strategies for the investigation of microbial secondary metabolites, *Nat. Prod. Rep.*, 2017, **34**, 6–24.

16 D. Krug and R. Müller, Secondary metabolomics: the impact of mass spectrometry-based approaches on the discovery and characterization of microbial natural products, *Nat. Prod. Rep.*, 2014, **31**, 768–783.

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 and F. Vargas, Reproducible molecular networking of untargeted mass spectrometry data using GNPS, *Nat. Protoc.*, 2020, **15**, 1954–1991.

18 K. Dührkop, M. Fleischauer, M. Ludwig, A. A. Aksенов, A. V. Melnik, M. Meusel, P. C. Dorrestein, J. Rousu and S. Böcker, SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information, *Nat. Methods*, 2019, **16**, 299–302.

19 A. W. Goering, R. A. McClure, J. R. Doroghazi, J. C. Albright, N. A. Haverland, Y. Zhang, K.-S. Ju, R. J. Thomson, W. W. Metcalf and N. L. Kelleher, Metabologenomics: Correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer, *ACS Cent. Sci.*, 2016, **2**, 99–108.

20 L. K. Caesar, R. Montaser, N. P. Keller and N. L. Kelleher, Metabolomics and genomics in natural products research: complementary tools for targeting new chemical entities, *Nat. Prod. Rep.*, 2021, **38**, 2041–2065.

21 R. D. Kersten, Y.-L. Yang, Y. Xu, P. Cimermancic, S.-J. Nam, W. Fenical, M. A. Fischbach, B. S. Moore and P. C. Dorrestein, A mass spectrometry-guided genome mining approach for natural product peptidogenomics, *Nat. Chem. Biol.*, 2011, **7**, 794–802.

22 D. N. Chigumba, L. S. Mydy, F. de Waal, W. Li, K. Shafiq, J. W. Wotring, O. G. Mohamed, T. Mladenovic, A. Tripathi and J. Z. Sexton, Discovery and biosynthesis of cyclic plant peptides via autocatalytic cyclases, *Nat. Chem. Biol.*, 2022, **18**, 18–28.

23 B. Behsaz, E. Bode, A. Gurevich, Y.-N. Shi, F. Grundmann, D. Acharya, A. M. Caraballo-Rodríguez, A. Bouslimani, M. Panitchpakdi and A. Linck, Integrating genomics and metabolomics for scalable non-ribosomal peptide discovery, *Nat. Commun.*, 2021, **12**, 3225.

24 R. D. Kersten, N. Ziemert, D. J. Gonzalez, B. M. Duggan, V. Nizet, P. C. Dorrestein and B. S. Moore, Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, E4407–E4416.

25 M. Grininger, Enzymology of assembly line synthesis by modular polyketide synthases, *Nat. Chem. Biol.*, 2023, **19**, 401–415.

26 C. Hertweck, The biosynthetic logic of polyketide diversity, *Angew. Chem., Int. Ed.*, 2009, **48**, 4688–4716.

27 E. J. Helfrich, S. Reiter and J. Piel, Recent advances in genome-based polyketide discovery, *Curr. Opin. Biotechnol.*, 2014, **29**, 107–115.

28 J. B. McAlpine, B. O. Bachmann, M. Pirae, S. Tremblay, A.-M. Alarco, E. Zazopoulos and C. M. Farnet, Microbial genomics as a guide to drug discovery and structural elucidation: ECO-02301, a novel antifungal agent, as an example, *J. Nat. Prod.*, 2005, **68**, 493–496.

29 A. R. Johnson and E. E. Carlson, Collision-induced dissociation mass spectrometry: a powerful tool for natural product structure elucidation, *Anal. Chem.*, 2015, **87**, 10668–10678.

30 R. H. Wills, M. Tosin and P. B. O'Connor, Structural characterization of polyketides using high mass accuracy tandem mass spectrometry, *Anal. Chem.*, 2012, **84**, 8863–8870.

31 H. Zhang, D. Zhang and K. Ray, A software filter to remove interference ions from drug metabolites in accurate mass liquid chromatography/mass spectrometric analyses, *J. Mass Spectrom.*, 2003, **38**, 1110–1112.



32 H. Zhang, D. Zhang, K. Ray and M. Zhu, Mass defect filter technique and its applications to drug metabolite identification by high-resolution mass spectrometry, *J. Mass Spectrom.*, 2009, **44**, 999–1016.

33 L. Sleno, The use of mass defect in modern mass spectrometry, *J. Mass Spectrom.*, 2012, **47**, 226–236.

34 H. Zhang, M. Zhu, K. L. Ray, L. Ma and D. Zhang, Mass defect profiles of biological matrices and the general applicability of mass defect filtering for metabolite detection, *Rapid Commun. Mass Spectrom.*, 2008, **22**, 2082–2088.

35 Y.-L. Fan, R.-Z. Liu, Q. Tan, H.-L. Zhao, M. Song, R. Wang, P. Li and H. Yang, A database-guided integrated strategy for comprehensive chemical profiling of traditional Chinese medicine, *J. Chromatogr. A*, 2022, **1674**, 463145.

36 M.-N. Li, C.-R. Li, W. Gao, P. Li and H. Yang, Highly sensitive strategy for identification of trace chemicals in complex matrix: Application to analysis of monacolin analogues in monascus-fermented rice product, *Anal. Chim. Acta*, 2017, **982**, 156–167.

37 B. R. Terlouw, K. Blin, J. C. Navarro-Munoz, N. E. Avalon, M. G. Chevrette, S. Egbert, S. Lee, D. Meijer, M. J. Recchia and Z. L. Reitz, MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters, *Nucleic Acids Res.*, 2023, **51**, D603–D610.

38 K. Blin, S. Shaw, H. E. Augustijn, Z. L. Reitz, F. Biermann, M. Alanjary, A. Fetter, B. R. Terlouw, W. W. Metcalf and E. J. Helfrich, antiSMASH 7.0: New and improved predictions for detection, regulation, chemical structures and visualisation, *Nucleic Acids Res.*, 2023, **51**, W46–W50.

39 M. A. Skinnider, C. W. Johnston, M. Gunabalasingam, N. J. Merwin, A. M. Kieliszek, R. J. MacLellan, H. Li, M. R. Ranieri, A. L. Webster and M. P. Cao, Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences, *Nat. Commun.*, 2020, **11**, 6058.

40 Z. Huang, S. Xie, R.-Z. Liu, C. Xiang, S. Yao and L. Zhang, Plug-and-Play Engineering of Modular Polyketide Synthases, *Research Square preprint*, 2024, DOI: <https://www.researchsquare.com/article/rs-3980064/v1>.

41 R. F. Little and C. Hertweck, Chain release mechanisms in polyketide and non-ribosomal peptide biosynthesis, *Nat. Prod. Rep.*, 2022, **39**, 163–205.

42 G. Yadav, R. S. Gokhale and D. Mohanty, Computational approach for prediction of domain organization and substrate specificity of modular polyketide synthases, *J. Mol. Biol.*, 2003, **328**, 335–363.

43 A. Morales-Amador, M. L. Souto, C. Hertweck, J. J. Fernández and M. García-Altares, Rapid Screening of Polyol Polyketides from Marine Dinoflagellates, *Anal. Chem.*, 2022, **94**, 14205–14213.

44 T. Fonseca, N. P. Lopes, P. J. Gates and J. Staunton, Fragmentation studies on tetronasin by accurate-mass electrospray tandem mass spectrometry, *J. Am. Soc. Mass Spectrom.*, 2004, **15**, 325–335.

45 D. P. Demarque, A. E. Crotti, R. Vessecchi, J. L. Lopes and N. P. Lopes, Fragmentation reactions using electrospray ionization mass spectrometry: an important tool for the structural elucidation and characterization of synthetic and natural products, *Nat. Prod. Rep.*, 2016, **33**, 432–455.

46 H. Ikeda, K. Shin-ya and S. Omura, Genome mining of the *Streptomyces avermitilis* genome and development of genome-minimized hosts for heterologous expression of biosynthetic gene clusters, *J. Ind. Microbiol. Biotechnol.*, 2014, **41**, 233–250.

47 J. Kahan, F. Kahan, R. Goegelman, S. Currie, M. Jackson, E. Stapley, T. Miller, A. Miller, D. Hendlin and S. Mochales, Thienamycin, a new  $\beta$ -lactam antibiotic I. Discovery, taxonomy, isolation and physical properties, *J. Antibiot.*, 1979, **32**, 1–12.

48 S. Khaoua, A. Lebrihi, P. Germain and G. Lefebvre, Cephamycin C biosynthesis in *Streptomyces cattleya*: nitrogen source regulation, *Appl. Microbiol. Biotechnol.*, 1991, **35**, 253–257.

49 M. Sanada, T. Miyano, S. Iwadare, J. M. Williamson, B. H. Arison, J. L. Smith, A. W. Douglas, J. M. Liesch and E. Inamine, Biosynthesis of fluorothreonine and fluoroacetic acid by the thienamycin producer, *Streptomyces cattleya*, *J. Antibiot.*, 1986, **39**, 259–265.

50 A. J. Kempf, K. E. Wilson, O. D. Hensens, R. L. Monaghan, S. B. Zimmerman and E. L. Dulaney, L-681, 217, a new and novel member of the efrotomycin family of antibiotics, *J. Antibiot.*, 1986, **39**, 1361–1367.

51 S. Sugai, H. Komaki, H. Hemmi and S. Kodani, Isolation and structural determination of a new antibacterial compound demethyl-L-681,217 from *Streptomyces cattleya*, *J. Antibiot.*, 2016, **69**, 839–842.

52 E. Harunari, H. Komaki and Y. Igarashi, Biosynthetic origin of butyrolactol A, an antifungal polyketide produced by a marine-derived *Streptomyces*, *Beilstein J. Org. Chem.*, 2017, **13**, 441–450.

53 T. Weber, K. J. Laiple, E. K. Pross, A. Textor, S. Grond, K. Welzel, S. Pelzer, A. Vente and W. Wohlleben, Molecular analysis of the kirromycin biosynthetic gene cluster revealed  $\beta$ -alanine as precursor of the pyridone moiety, *Chem. Biol.*, 2008, **15**, 175–188.

54 Z. Lin, Z. Hu, L. Zhou, B. Liu, X. Huang, Z. Deng and X. Qu, A large conserved family of small-molecule carboxyl methyltransferases identified from microorganisms, *Proc. Natl. Acad. Sci. U. S. A.*, 2023, **120**, e2301389120.

55 R. Little, F. C. Paiva, R. Jenkins, H. Hong, Y. Sun, Y. Demydchuk, M. Samborskyy, M. Tosin, F. J. Leeper and M. V. Dias, Unexpected enzyme-catalysed [4+2] cycloaddition and rearrangement in polyether antibiotic biosynthesis, *Nat. Catal.*, 2019, **2**, 1045–1054.

56 S. Asamizu, A. A. C. Pramana, S.-J. Kawai, Y. Arakawa and H. Onaka, Comparative metabolomics reveals a bifunctional antibacterial conjugate from combined-culture of *Streptomyces hygroscopicus* HOK021 and *Tsukamurella pulmonis* TP-B0596, *ACS Chem. Biol.*, 2022, **17**, 2664–2672.

57 J. Li, X. Tang, T. Awakawa and B. S. Moore, Enzymatic C–H Oxidation–Amidation Cascade in the Production of Natural



and Unnatural Thiotetronate Antibiotics with Potentiated Bioactivity, *Angew. Chem., Int. Ed.*, 2017, **56**, 12234–12239.

58 Y. Sun, F. Hahn, Y. Demydchuk, J. Chettle, M. Tosin, H. Osada and P. F. Leadlay, In vitro reconstruction of tetroenate RK-682 biosynthesis, *Nat. Chem. Biol.*, 2010, **6**, 99–101.

59 S. Kishore and C. Khosla, Genomic mining and diversity of assembly line polyketide synthases, *Open Biol.*, 2023, **13**, 230096.

60 Z. P. Mai, B. Zhang, Z. X. Pang, J. Shi, Z. F. Xu, B.-B. Huang, S. Y. Ma, R. H. Jiao, Z.-J. Yao, R. X. Tan and H. M. Ge, Insight into the role of a trans-AT polyketide synthase in the biosynthesis of lankacidin-type natural products, *Nat. Synth.*, 2024, **3**, 1255–1265.

61 C. S. McCaughey, J. A. van Santen, J. J. J. van der Hooft, M. H. Medema and R. G. Linington, An isotopic labeling approach linking natural products with biosynthetic gene clusters, *Nat. Chem. Biol.*, 2022, **18**, 295–304.

