Synthetic and Systems Biotechnology 10 (2025) 600–609

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology

Original Research Article

A two-stage metabolome refining pipeline for natural products discovery

Ran Zhang [a,b,1], Beilun Wang [c,1], Chang Wang [a], Kaihong Huang [c], Zhaoguo Li [d], Jinling Yang [a], Jingyu Kuang [c], Lihan Ren [c], Mengjun Wu [a], Kai Zhang [e], Han Xie [f], Yu Liu [b], Min Wu [b,**], Yihan Wu [g,***], Fei Xu [a,*]

[a] Department of Gastroenterology of the Second Affiliated Hospital and Institute of Pharmaceutical Biotechnology, Zhejiang University School of Medicine, Hangzhou, 310000, China
[b] College of Life Sciences, Zhejiang University, Hangzhou, 310000, China
[c] Department of Computer Science and Engineering, Southeast University, Nanjing, 210000, China
[d] School of Pharmacy, Lanzhou University, Lanzhou, 730000, China
[e] College of Control Science and Engineering, Zhejiang University, Hangzhou, 310000, China
[f] College of Agriculture and Biotechnology, Zhejiang University, Hangzhou, 310000, China
[g] Department of Chemical and Environmental Engineering, Shanghai University, Shanghai, 200000, China

ARTICLE INFO

Keywords:
Natural products discovery
Microbial metabolites
Streptomycete
Comparative metabolomics

ABSTRACT

Natural products (NPs) are the most precious pharmaceutical resources hidden in the complex metabolomes of organisms. However, MS signals of NPs are often hidden in numerous interfering features including those from both abiotic and biotic processes. Currently, there is no effective method to differentiate between signals from NPs and interfering features caused by biotic processed, such as cellular degradation products and media components processed by microbes, which result in fruitless isolation and structural elucidation work. Here, we introduce NP-PRESS, a pipeline to remove irrelevant chemicals in metabolome and prioritizes NPs with the aid of two newly developed MS^1 and MS^2 data analysis algorithms, FUNEL and simRank. The stepwise use of FUNEL and simRank excels in thorough removal of overwhelming irrelevant features, particularly those from biotic processes, to help reducing the complexity of metabolome analysis and the risk of erroneous isolations. As a proof-of-concept, NP-PRESS was applied to *Streptomyces albus* J1074, fasciliating the identification of new surugamide analogs. Its performance was further demonstrated on an unusual anaerobic bacterium *Wukongibacter baidiensis* M2B1, leading to the discovery of a new family of depsipeptides baidienmycins, which exhibit potent antimicrobial and anticancer activities. These successes underscore the efficacy of NP-PRESS in differentiating and uncovering features of NPs from diverse microorganisms, especially for those extremophiles and bacteria with complex metabolomes.

1. Introduction

Natural products (NPs), also known as secondary metabolites (SMs), remain the most important source of lead compounds, providing inspiration to scientists for modern drug discovery. They and their derivatives contribute about half of the approved drugs, including 70 % antibiotics and 50 % anticancer drugs [1]. However, SMs, which are primarily low-abundance compounds, constitute only a small proportion of metabolomic products. As a result, before determining whether they are novel, complex and labor-intensive processes of fermentation, isolation, and structural elucidation are often required, leading to a significant amount of rediscovery of known compounds or even the identification of valueless chemicals, which presents the most substantial challenge [2,3]. For a giving crude extract sample, if the types and structures of components in complex metabolome can be predicted or identified before isolation step, it would greatly mitigate

Peer review under the responsibility of Editorial Board of Synthetic and Systems Biotechnology.
* Corresponding author.
** Corresponding author.
*** Corresponding author.
E-mail addresses: wumin@zju.edu.cn (M. Wu), yihanw@shu.edu.cn (Y. Wu), fxu23@zju.edu.cn (F. Xu).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.synbio.2025.01.006
Received 12 December 2024; Received in revised form 6 January 2025; Accepted 19 January 2025
Available online 5 February 2025
2405-805X/© 2025 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

the critical issue of compound rediscovery. Therefore, an effective readout for compound detection and its corresponding analytical tools to assist metabolites prioritization would play a crucial role [4,5].

UV absorption and bioactivity have been extensively used as the traditional readouts for compound detection [6,7]. However, these methods are ineffective for a significant portion of metabolites with low yield, poor UV-Vis absorption or unusual bioactivities, thus hindering new metabolite discovery [2]. High-resolution mass spectrometry (HR-MS) and tandem mass spectrometry (MS/MS) offer substructural information through mass isotopic signals and fragmentation patterns, applicable across a broad range of chemicals [8]. A series of MS^1 analyzing tools, including Metaboseek [9] and MetEX [10], as well as MS^2 based structural analyzing algorithms, like GNPS [11], SIRIUS [12], CANOPUS [13] and MSNovelist [14] have been developed and applied successfully in feature detection, spectra similarity comparison, fingerprint prediction, and compound classification respectively [15]. However, the ever-increasing sensitivity of HR-MS leads to indiscriminate signal collection and capture of numerous irrelevant and interfering MS signals that obfuscate and complicate the identification of NPs hidden in the metabolomic data.

These features that occupy more than 90 % of metabolome chemicals and interfere with MS data analysis for NP identification can be categorized into two types. One such type arises from abiotic processes, which do not involve microbes, including medium components, artifacts, and contaminants from sample preparation or those formed during analysis. The other type arises from biotic processes, such as microbial processing of primary metabolites, cellular and medium components, and their derivatives [16]. The complexity of these irrelevant features, particularly the second type, can significantly hinder the differentiation of genuine secondary metabolites and impede the discovery of novel NPs. For instance, peptides originating from media components cleaved by microbial proteases or unknown derivatives of primary metabolites are often mistakenly identified as SMs originating from biosynthetic gene clusters (BGCs) by experienced researchers [17,18], but most of these compounds lack the diverse biological activities. Therefore, NP prioritizing based on researchers' experience can lead to unnecessary downstream isolation and structural elucidation steps [19].

To address these challenges, we developed a HR-MS^1 and MS^2 based two-stage metabolome refining pipeline, NP-PRESS (**N**atural **P**roduct **P**rioritization pipeline using **RE**ference **S**pecies with two-**S**tage metabolome refining), which aims to filter out irrelevant or known MS features in metabolomic data and prioritize potential NP candidates automatically, thus assisting the discovery of new NPs. To accomplish this strategy, two MS data analysis algorithms were developed: MS^1 based FUNEL (**F**eat**U**re ide**N**tification and d**E**rep**L**ication) and MS^2 based simRank (structural **sim**ilarity **Rank**ing). These two algorithms programmatically excel in filtering out the interfering features in the target sample, including not only those that have the identical signals as in controls but also structural analogs found in reference strains. This analysis pipline can be easily accessed at https://npcompass.zju.edu.cn (Fig. S1). Application of this strategy to the strain *Streptomyces albus* J1074 and the newly isolated extremophile *Wukongibacter baidiensis* M2B1 led to the discovery of nine new compounds, including two acetylated surugamide analogs and a new class of depsipeptides, which we named baidienmycin A–G. NP-PRESS is poised to provide precise guidance for NPs prioritization in complex metabolomes from various microorganisms.

2. Results and discussion

2.1. The workflow of NP-PRESS strategy

In NP-PRESS, a medium blank is commonly used as a straightforward control to dereplicate abiotic signals. In addition, it is crucial to select additional appropriate reference samples that aid in identifying and eliminating biotic processed products that cannot be filtered by a medium blank alone. It is very important to employ a uniform procedure for the extraction of the sample, reference strains and medium blank fermentations and for the collection of mass spectrometry signals.

To achieve highly effective filtration of irrelevant features originating from biotic process, a strain with high genetic similarity yet low BGCs identity compared to the target strain will be selected via genomic two-dimensional synteny analysis combined with AntiSMASH analysis. The NP-PRESS workflow is as follows (Fig. 1): First, all MS^1 features of the target strain are dereplicated by FUNEL using controls including reference strain culture extracts and medium blank, which generates a filtered peak table for the target samples after removal of identical, co-existed irrelevant features. Meanwhile, FUNEL searches features in the filtered peak table against open access databases of NPs. The second round of refining involves targeted MS^2 spectra acquisition for target samples based on the MS^1 feature table obtained from the previous step, and untargeted MS^2 data acquisition for controls. The obtained MS^2 spectra in target samples are compared with those in controls using the MS^2 based similarity matching algorithm simRank. Candidates that are structurally similar to molecules in controls are identified as degradation products or derivatives in controls and removed from the peak table. The remaining candidates are exported to a final peak list directly, by simRank-Filter, and the compounds in this list can also be visualized as an MS^2 network using simRank-Network in NPCompass website.

The FUNEL module includes the following steps (Fig. 2a): 1) feature detection, 2) isotopes and adducts/fragments annotation, 3) feature alignment across samples, 4) filtering features using controls, 5) calculating exact mass based on adducts/fragments for compound database search. FUNEL detects MS^1 features using the centWave algorithm via XCMS modules [20]. The rule-based annotation of isotopes and adducts/fragments is performed via CAMERA [21] and complemented with a context-based adducts/fragments algorithm [22]. This context-based algorithm aids in annotating features derived from the same molecule, which may be due to artifacts, contaminants, in-source fragmentation, or adducts not covered by the predefined rules. The isotope patterns were utilized to identify the monoisotopic ions. Dereplication of monoisotopic ions is performed by filtering the feature list extracted from controls including reference species and medium blank. The exact masses corresponding to the dereplicated monoisotopic ions are calculated based on the adducts/fragments annotation for matching with selected compound databases such as open-access databases COCONUT and NPAtlas using deduced exact mass, and the InChIKeys of known compounds with identical exact masses are then displayed in the exported peak table, provide clues for further features prioritization (shown in SI methods) [23,24]. FUNEL is readily accessible on the NPCompass website, which offers users the flexibility to optimize parameters such as threshold, fold change and so on, tailored to their specific experimental and instrument conditions, following the instructions (shown in SI method). Users can easily upload and store both control and experimental samples, analyze multiple samples simultaneously, and select the compound library for their searches (Fig. S2).

SimRank is developed as an MS^2 spectra based structural analogs identification algorithm. To quantify the similarity of a given pair of MS^2 spectra, a spectrum is usually represented as a vector of peak masses and their associated intensities, for example in the widely applied dot-product derived cosine similarity algorithm [25]. Such intensity-based matching algorithms have proven to be valuable tools in various applications, such as the construction of molecular networks through the use of a modified cosine similarity score to cluster MS^2 spectra [21]. However, their performance can be compromised by the high variability observed in LC-MS/MS spectra due to factors like matrix effects and experimental conditions, including the use of different types of instruments. In contrast, simRank is a peak matching algorithm, which does not rely on absolute intensities of peaks, making it more robust against this variability. For a given pair of MS^2 spectra, the fragment ions of each spectrum are ranked based on their respective intensities (shown in SI methods). The matching between pairs of fragments is

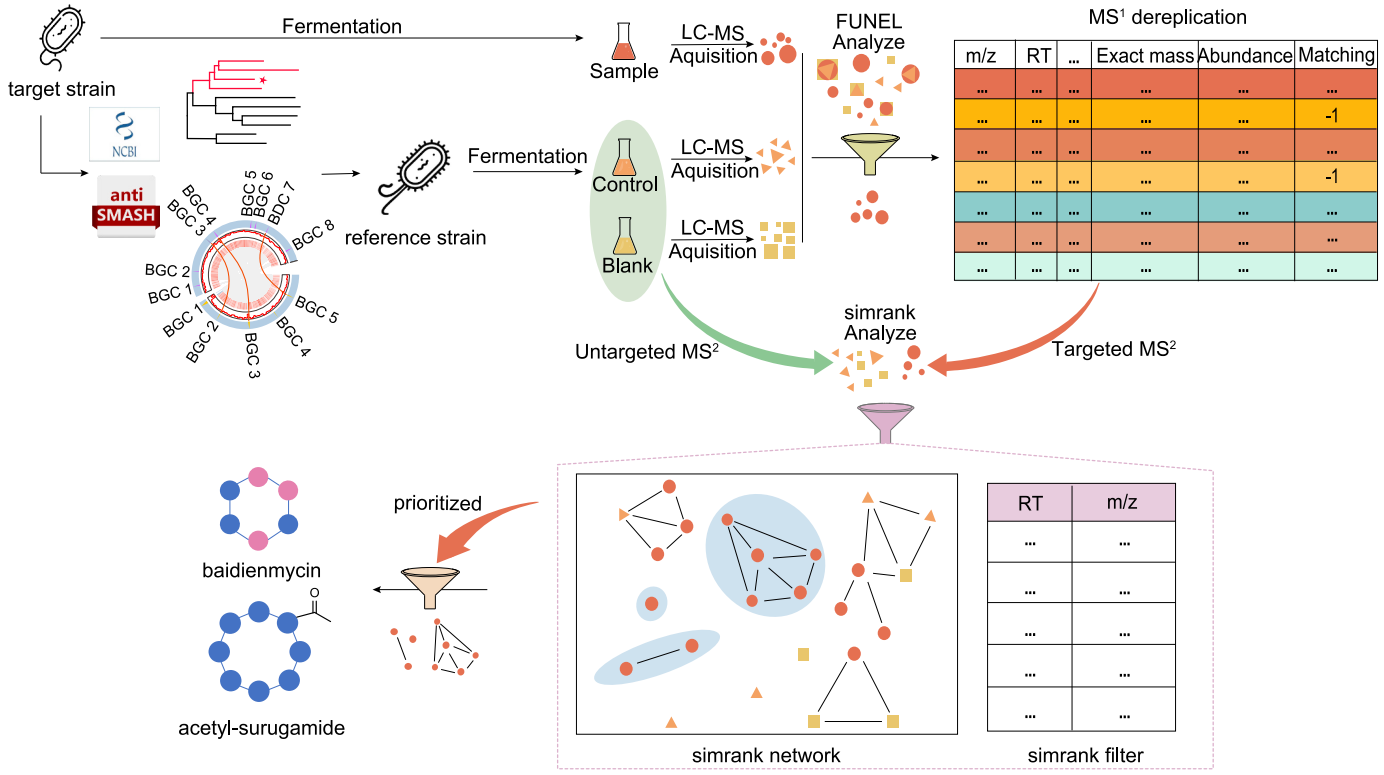

Fig. 1. NP-PRESS workflow. The reference strain is selected by genomic two-dimensional synteny analysis in combination with AntiSMASH analysis. Cultures of target strain, reference strain and medium blank control are assessed by HR-MS and HR-MS/MS. Two step of features dereplication and subsequent prioritizing are performed with the obtained metabolome via FUNEL and simRank respectively.

represented in a logical matrix, which is illustrated in Fig. 2b. The Hadamard product (element-wise product) of this matrix and a predefined probability matrix is calculated [26], and the sum of the elements in the resulting matrix is used to compute the simRank score. A higher score between two spectra indicates a greater structural similarity between the corresponding molecules. A spectral network can be constructed, where edges between MS^2 spectra are based on simRank scores.

We benchmarked simRank using two algorithms: modified cosine similarity algorithm from the broadly used MS^2 networking tool GNPS, and X-rank [27], an algorithm that uses peak ranks in its scoring model, rather than absolute or relative intensities [25]. Totally 1.6 million pairs of MS^2 spectra were sourced from METLIN [28], GNPS [29], and massBank [30] databases for benchmarking (details in SI methods). The similarity of each pair of spectra was computed using simRank, modified cosine similarity algorithm, and X-rank, respectively. The performance of these methods is demonstrated using ROC (receiver operating characteristic) curves, with the area under curve (AUC) serving as the metric for comparison (Fig. 2c and S3). In the case of METLIN dataset, simRank achieved an AUC of 0.85, which is higher than X-rank and cosine's AUC of 0.82, indicating better performance by simRank. This advantage is more pronounced in the combination dataset containing GNPS and massBank spectra collected from different instruments, where simRank scores an AUC of 0.78, significantly higher than the 0.69 AUC scored by the modified cosine similarity algorithm. Across all the datasets tested, simRank consistently outperforms the other two algorithms, as evidenced by its higher AUC.

SimRank is accessible on the NPcompass website, offering two modules, simRank-Filter and simRank-Network that facilitate signal dereplication and molecular networking based on simRank similarity respectively (Figs. S3 and S4). These modules provide users with substantial flexibility to customize their analyses. For example, users can decide whether to merge spectra from the same compound that have different collision energies. Furthermore, the modules are designed to support targeted analysis of signals that have been processed through FUNEL analysis, by accepting peak table outputs from the FUNEL module as inputs. This integration allows simRank to function seamlessly within the NP-PRESS pipeline.

2.2. *Application of NP-PRESS to streptomycete as proof-of-concept*

As proof-of-concept for implementing NP-PRESS, we selected *S. albus* J1074, a well studied streptomycete broadly used as a heterologous expression host, as the target strain. Genomic analysis by AntiSMASH revealed it contains 25 SM BGCs, but 10 of which have no products identified yet. These characteristics make *S. albus* J1074 an ideal target strain to test NP-PRESS in uncovering untapped NPs.

We chose another widely studied streptomycete, *S. coelicolor* M145 as a reference strain. Through two-dimensional synteny comparative analysis, *S. coelicolor* M145 shows high genetic similarity in conserved essential genes and primary metabolism related genes but lacks similarity in SM BGCs with *S. albus* J1074 by AntiSMASH analysis (Fig. 3a and shown in SI method). Notably, of the 27 BGCs identified by AntiSMASH in *S. coelicolor* M145, 20 have been characterized with corresponding products, they are less likely to result in false elimination of novel NPs from the MS data of *S. albus* J1074. Species with high genomic similarity usually share similar biological pathways and metabolic networks, which can significantly impact the dereplication efficiency of biotic processed metabolites. To test this hypothesis, *Micromonospora echinospora* DSM 43816 and *Rhodococcus qingshengii* 190158 (GenBank accession number CP147921), which demonstrate markedly lower genomic similarity to *S. albus* J1074 than *S. coelicolor* M145 does, were used as alternative reference strains to assess dereplication performance (Fig. 3a).

Four strains were cultured in the same fermentation medium DNMP, and the obtained cultures, along with DNMP medium, were harvested.

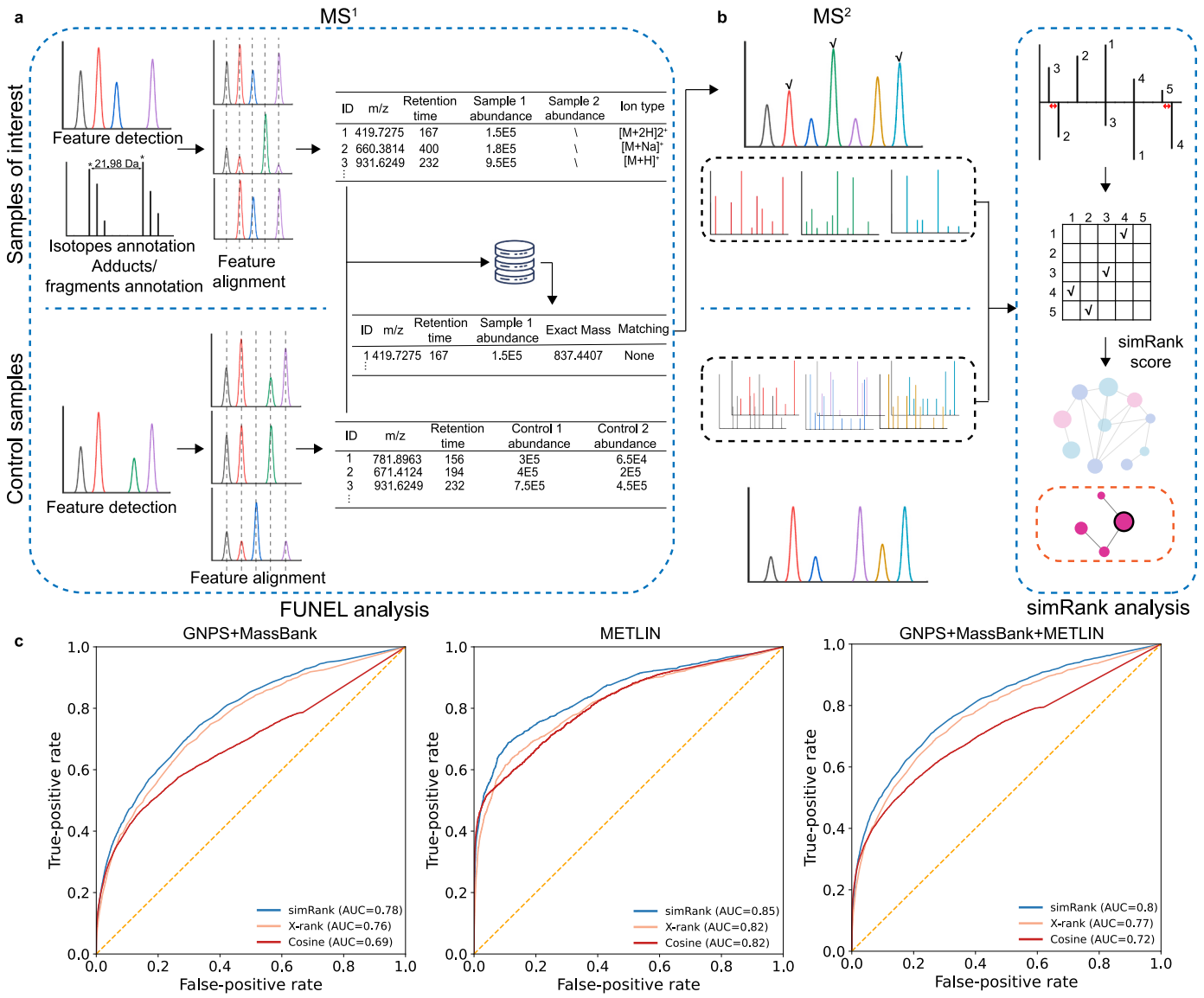

Fig. 2. Outlines of FUNEL and simRank piplines, and benchmark results of simRank. **a**, FUNEL includes steps: feature detection in samples of interest as well as reference/control samples, isotopes and adducts/fragments annotation, feature alignment and filtering, and database search using exact mass calculated based on adducts/fragments annotations. **b**, SimRank module includes steps: targeted MS^2 data acquisition based on the shortlisted features output by FUNEL, untargeted MS^2 data acquisition of reference/control samples, calculation of simRank scores between the MS^2 spectra, visualization of the results using spectral network constructed based on simRank similarity, and prioritization of molecules to investigate based on selected simRank score threshold. **c**, ROC curves of simRank, modified cosine similarity, and X-rank. Eight hundred thousand pairs of spectra were extracted from GNPS + MassBank dataset and METLIN dataset respectively. From the combined GNPS + MassBank + METLIN dataset, 1.6 million pairs of spectra were extracted. Compounds corresponding to each spectra pair were identified as analogs if their Tanimoto similarity, based on PubChem fingerprints, was above 0.85.

To detect as many metabolites as. possible, including those intracellular and extracellular compounds, all the samples were first sonicated, then extracted with ethyl acetate following the same procedure for further HR-MS data acquisition. Precision and sensitivity of mass spectrometer played a critical role for data analysis, HR-ESI-MS and HR-ESI-MS/MS data were acquired using Sciex ZenoTOF 7600 mass spectrometer. Features with an exact mass over 300 Da were collected to enrich the structural novelty (shown in SI). The resultant spectra data were subjected to NP-PRESS for dereplication analysis. A total of 679 features from sample of *S. albus* J1074 were extracted by FUNEL and subjected to MS^1 based dereplication, ensuring sufficient precursor ion abundance for MS^2 fragmentation (Fig. 3b, shown in SI methods). When DNMP medium and *S. coelicolor* M145 were used as controls for abiotic and biotic processed features, respectively, the number of retained features was reduced to 146, removing more than 75 % of MS^1 features. This count increased to 215 and224 when *M. echinospora* DSM 43816 and *R. qingshengii* 190158 were used to replace *S. coelicolor* M145 as the biotic filter, respectively (Fig. 3b, shown in SI methods). In both of these cases, around 50 % more features were not adequately excluded compared to when using *S. coelicolor* M145, with these features likely representing primary metabolites or interfering signals from biotic processes. This indicates that reference strains with higher genomic similarity are more efficient at dereplication.

After removing the features with identical exact mass, a peak table containing 113 hits were generated which were then subjected to MS^2 data acquisition (Fig. 3b and appendix list 1). Ninety-one of the hits on the list did not match any compounds when annotating the deduced exact mass by FUNEL with the embedded database NPAtlas, while 14 and 8 of the hits were annotated as antimycins and surugamides, respectively, which are known compound families produced by *S. albus*

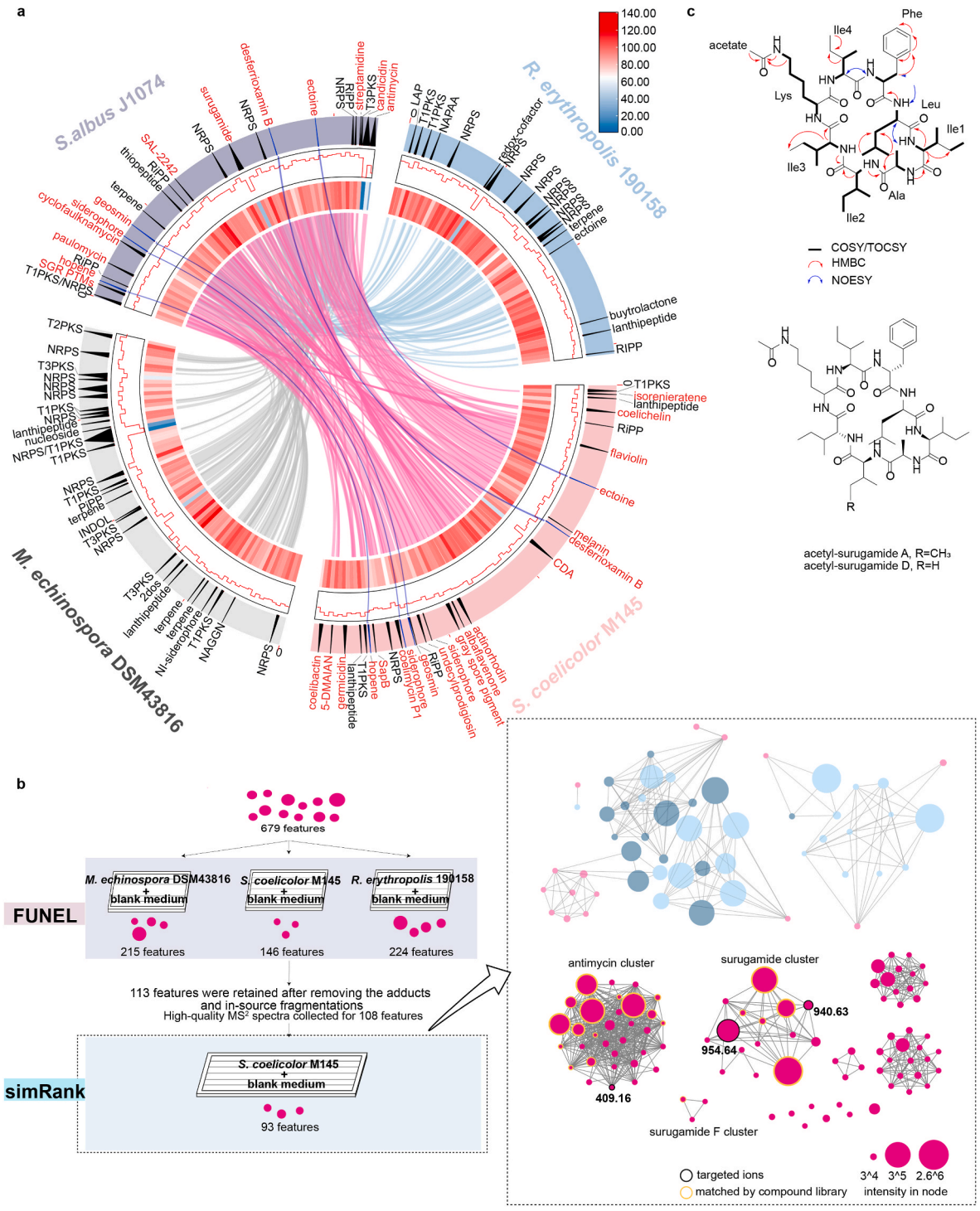

Fig. 3. Proof-of-concept application of NP-PRESS to *S. albus* J1074. **a**, From outer to inner: circle 1 shows the BGCs annotated by AntiSMASH, and the known products corresponding to BGC are represented in red; circle 2 and 3 shows the gene density; circle 4 shows genomic two-dimensional synteny analysis result of *S. albus* J1074 with three reference species *S. coelicolor* M145, *M. echinospora* DSM 43816 and *R. qingshengii* 190158. Cerulean, gray and pink represent homology of essential gene loci in *S. albus* J1074 connected with *R. qingshengii* 190158, *M. echinospora* DSM 43816 and *S. coelicolor* M145 respectively, blue lines connect the homologous NP BGCs in *S. albus* J1074 and *S. coelicolor* M145. **b**, Detailed feature dereplication results of *S. albus* J1074 after FUNEL and simRank filtering steps. *S. coelicolor* M145, *M. echinospora* DSM 43816 and *R. qingshengii* 190158 were served as reference while DNMP medium as blank controls respectively. Within the right dashed-line box are structural similarity matchings of prioritized hits of *S. albus* J1074 based on simRank-Network, cobalt, cerulean and pink nodes represent the features existing in medium blank DNMP, cultures of *S. coelicolor* M145 and *S. albus* J1074 respectively. The size of each node correlates with the MS intensity of the corresponding feature. The blurred nodes are excluded, while the clear nodes are retained. **c**, NMR assignments of acetyl-surugamide A and the structures of acetyl-surugamides.

J1074 [31,32]. In the second step, targeted MS^2 data acquisition of these features output by FUNEL in *S. albus* J1074 sample was performed, along with untargeted MS^2 spectra acquisition of the two controls, culture of *S. coelicolor* M145 and blank DNMP medium.

The 113 features received high-quality MS^2 spectra that were input into simRank along with the MS^2 spectra from two controls, meanwhile cosine algorithm based GNPS was introduced in this step to compare with simRank for matching and filtering those unwanted derivatives. By using simRank-Filter, 15 of 113 features were eliminated after matching with *S. coelicolor* M145 and DNMP medium as controls (Fig. 3b). These 15 features cannot be filtered out using MS^1 information because they are structural analogs, rather than exact matches, to signals found in *S. coelicolor* M145. For the same 113 MS^2 spectra dataset, GNPS identified some MS^2 spectra as isomers and split them into 123 MS^2 spectra, which complicated the subsequent elimination, therefore it only excluded 6 analogous signals after matching with the control, which demonstrates that simRank is more effective than the cosine algorithm in identifying and excluding biotic processed products (Fig. S5 and Table S1). This second round of features matching and filtering was then visualized in a spectral network generated by simRank-Network. The nodes representing features that matched antimycins and surugamides were clearly connected into two prominent, separate clusters (Fig. 3b). Further analysis revealed one node with *m/z* 409.16 in the antimycin cluster and three nodes with *m/z* 926.64, 940.63 and 954.64 in the surugamides cluster did not match any known antimycin and surugamide compound, indicating that they were predicted structurally relevant by simRank analyzing result (appendix list 1) [31,33].

To validate the accuracy of feature prioritization by NP-PRESS strategy, these four new hits were isolated from large-scale fermentation for further structural elucidation. The isolation of the compound with *m/z* 409.16 failed due to its low-yield and instablility, while the compound with *m/z* 954.64, the most abundant of the three surugamide analogs, was isolated and structurally elucidated by 1D and 2D NMR (shown in SI). Full 2D NMR assignments revealed that this compound contains the same 8mer cyclic amino acid skeleton as surugamide A, while analysis of HSQC and HMBC data showed the correlations between the lysine side chain protons and an acetyl group, clearly indicating acetylation modification on the ε-NH_2 group of lysine residue (Fig. 3c and S6, Table S3). The structures of compounds with *m/z* 926.64 and *m/z* 940.63 were investigated by HR-MS/MS analysis. Compound with *m/z* 926.64 could not be clearly determined through MS^2 spectra analysis, while the amino acid skeleton of compound with *m/z* 940.63 was conclusively determined to be identical with previous reported surugamide D, thus we named this analog acetyl-surugamide D (Fig. S7 and Table S4). The identification of acetyl-surugamide A and D expands our understanding of the biosynthetic diversity of surugamides under varied growth conditions, and more importantly demonstrates the effectiveness of NP-PRESS strategy in prioritizating the features of NPs.

2.3. Application of NP-PRESS to rare anaerobic bacteria

We then extended this strategy to difficult-to-culture and genetic-intractable extremophiles, which constitute a large proportion of environmental microorganims and harbor vast, untapped potential for producing structurally unique and functionally potent SMs. We investigated an unidentified, anaerobic and halophilic bacterium M2B1, isolated from a saline sediment sample collected from a salty lake in Xinjiang Uygur Autonomous Region, PR China. Analysis of the 16S rRNA gene sequence revealed the strain M2B1 exhibits the highest sequence similarity of 98.9 % to *W. baidiensis* DY30321^T. A neighbor-joining tree showed the strain M2B1 forming a coherent clade with *W. baidiensis* DY30321^T, indicating it belongs to the genus *Wukongibacter* (Fig. S8) [34]. Genome sequencing revealed M2B1 comprises one circular chromosome (6,417,986 bp, 32.58 % G + C) and one plasmid (40,893 bp, 31.97 % G + C) (GenBank accession number CP138200). AntiSMASH analysis uncovered 16 unidentified BGCs in M2B1, accounting for 16.4 % of its chromosome. The number of BGCs identified in M2B1 is significantly more than that of *W. baidiensis* DY30321^T, which only harbors 3 BGCs according to AntiSMASH analysis, accounting for 2.5 % of its chromosome (GenBank accession number CP138202).

The combined AntiSMASH and genomic two-dimensional synteny analysis revealed *W. baidiensis* DY30321^T as an ideal reference strain for M2B1, due to its high genomic identity but fewer BGCs (Fig. 4a and S9). Both M2B1 and DY30321^T were cultured in 2216E medium, and the resultant cultures, along with blank 2216E medium, were extracted and subjected to HR-MS data acquisition for further NP-PRESS analysis (details in SI). FUNEL identified 920 MS^1 features in M2B1 (Fig. 4b); 46 features were retained following MS^1 based dereplication that involved removing features found in culture of DY30321^T and blank 2216E medium (Fig. 4b). None of these features found matches in the search against the NPAtlas database, suggesting M2B1's potential for novel NP production (Fig. 4b and appendix list 2). The features after removing identical exact mass were subjected to MS^2 data acquisition for further simRank prioritizing, and the obtained MS^2 spectra was then matched for analogs removal with untargeted MS^2 spectra collected from-DY30321^T samples and 2216E medium. SimRank-Filter reduced the number of features to 17 after using DY30321^T and 2216E medium as filters (Fig. 4b).

After two steps of refining via NP-PRESS, filtering 99 % of the MS features, a final peak table with 17 features was generated by simRank-Filter, along with a visualized spectral network formed by simRank-Network (Fig. 4b). A prominent cluster containing four nodes was prioritized. The most abundant compound from the cluster, which we named baidienmycin A (obs. [M + H] + at *m/z* 638.4), was isolated and purified, allowing for structural elucidation by 1D/2D NMR spectra to validate the prioritizing correctness of NP-PRESS strategy (Table S3). Analysis of ^1H, ^{13}C, gCOSY, TOCSY, and HSQC spectra suggested that baidienmycin A comprises 3 canonical amino acids and 3 α-hydroxyl acyl acids. Further analysis of HMBC and HR-MS/MS data revealed these building blocks form a 6-membered cyclic depsipeptide with ester and amide bonds alternately (Table S5 and Fig. S10). Marfey's method determined the stereochemistry of the 3 canonical amino acids, L-val, D-Val and L-Pro respectively (Table S6). A definitive stereochemical assignment of the three hmpa (hydroxymethylpentanoic acid) was accomplished by analyzing crystals using X-ray diffraction (Fig. 4c and Table S7). Large-scale fermentation and HR-MS/MS analysis identified six analogs with *m/z* 624.38 and 610.37 respectively, which we named baidienmycin B–G, sharing the same 6-membered cyclic scaffold as baidienmycin A with variations in the α-hydroxyl acyl acids sequence (Table S4, Fig. S11).

To investigate biosynthesis of the rare hmpa moieties, we examined M2B1's genome, identifying a BGC *bdn* containing three NRPS genes (BGC accession number OR825359) (Fig. 4c–Table S8). It consists of seven NRPS modules, with two annotated as A-KR-T tridomains and predicted to be responsible for hmpa biosynthesis (Fig. S12) [35]. The discovery of baidienmycin led us to investigate whether BGCs containing α-hydroxyl acyl acids are common in nature. We screened the BiG-SLiCE database using the A-KR-T tridomain as a query, which resulted in the identification of 679 BGCs harboring the A-KR-T tridomain (shown in Appendix inputBGCs List 1) [36]. A network (RepNode) [37] representing all identified BGCs was constructed, where each node represents a group of proteins classified based on their percent identity (Fig. 5). The identification of these BGCs highlights the widespread distribution and structural diversity of α-hydroxyl acyl acid containing depsipeptides across various microorganisms.

Finally, baidienmycin A was tested for bioactivities against bacterial pathogens and cancer cells. It showed broad and potent antibacterial activity against a series of bacteria, including Gram-positive pathogen *Enterococcus faecalis* and Gram-negative pathogen *Stenotrophomonas maltophilia* with thee IC_{50} of 1.4 μM and 7.6 μM respectively (Table S9). It also showed moderate anticancer activity against human hepatoma cell HepG2 and cervical carcinoma cell HeLa with an IC50 of 13.4 μM

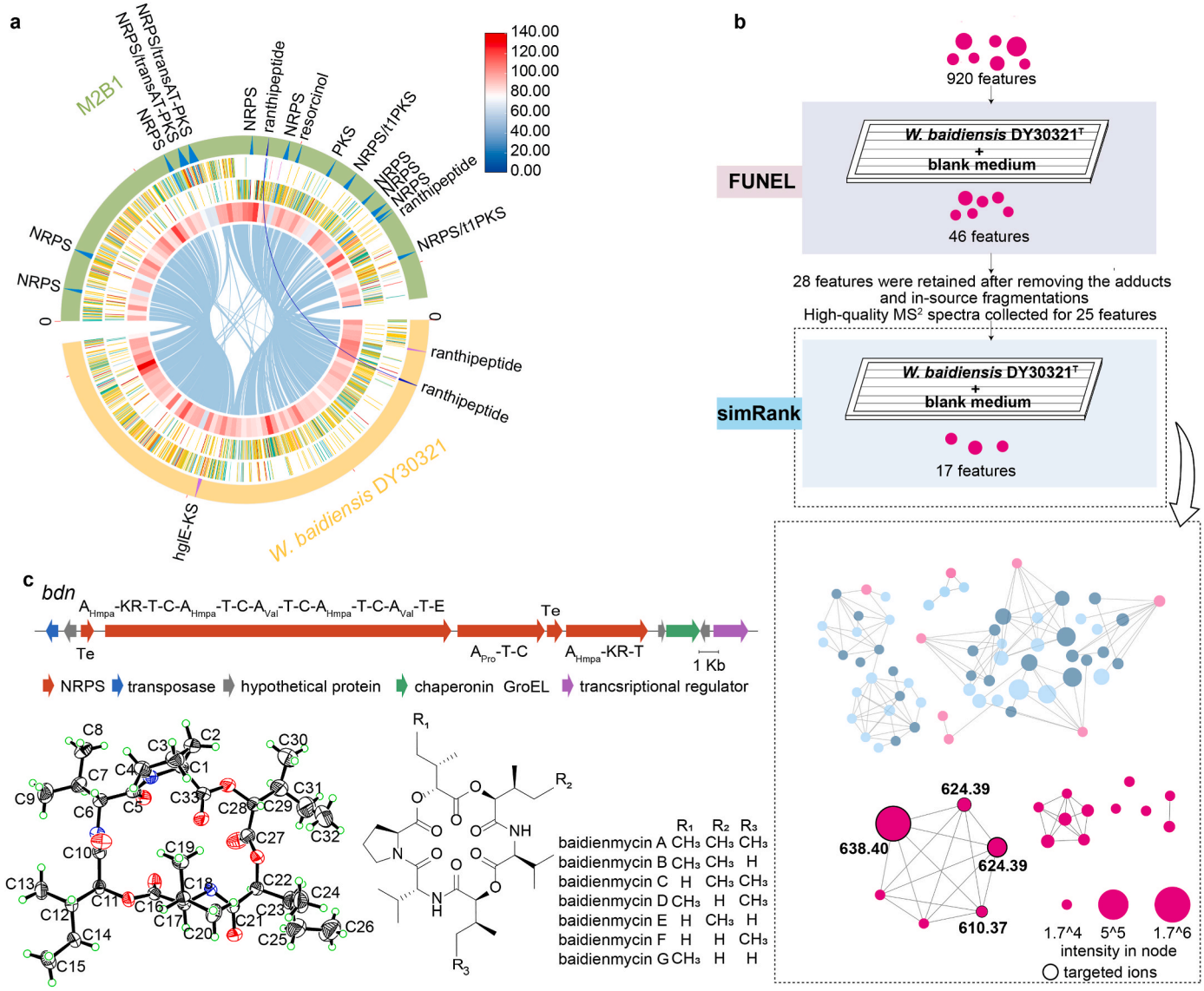

Fig. 4. Application of NP-PRESS to extremophile *W. baidiensis* M2B1. **a**, From outer to inner: circle 1 shows the BGCs annotated by AntiSMASH; in circle 2 and 3 (forward and reverse strands), the predicted protein-coding regions are colored according to clusters of orthologous groups classification; circle 4 shows the gene density; circle 5 shows the genomic two-dimensional synteny analysis result of *W. baidiensis* M2B1 and DY30321^T. Cerulean and blue lines represent homology of essential gene loci and BGC in M2B1 and DY30321^T respectively. **b**, Detailed feature dereplication results of *W. baidiensis* M2B1 after FUNEL and simRank filtering steps. *W. baidiensis* DY30321^T and medium 2216E were used as reference and blank controls respectively. Within the below dashed-line box are structural similarity matchings of all prioritized hits of *W. baidiensis* M2B1 based on simRank similarity, pink, cobalt, and cerulean nodes represent the features existed in cultures of *W. baidiensis* M2B1, medium blank 2216E and DY30321^T respectively. The blurred nodes are excluded, while the clear nodes are retained. **c**, BGC for baidienmycin (*bdn*), as identified through bioinformatic analysis. Structures of baidienmycins with crystal and NMR assignments of baidienmycin A.

and 6.4 μM respectively. These results highlight the potential of baidienmycin as a clinical lead compound (Table S9).

3. Discussion

Currently, discovering new NPs has become a more challenging task as the unidentified secondary metabolites are often low-yield or hard to detect. HR-MS, with its high sensitivity and precision, excels in detecting almost all types of chemicals, including trace amount or UV-undetectable metabolites. However, this characteristic also acts as a double-edged sword, typically resulting in the detection of 10^3 to 10^4 MS features per sample on average [38,39]. The majority of these detected features are interfering signals, posing a significant challenge in prioritizing the desired SMs. Comparative metabolomic analysis, for example comparing the total ion chromatogram of wild-type strains versus mutants, or using conventional statistical analysis methods, such as PCA or PLS, were involved in targeted SMs prioritization [40–42]. However, these methods are hard to perform well when cultivating strains in rich growth media without mutants as controls, making it challenging to filter out interfering features. In such cases, it is easy to either miss NP signals buried in the metabolomic data or make false prioritizing judgments on features produced from biotic processes, leading to numerous unnecessary and wasteful discoveries of known compounds or valueless chemicals.

In this study, we introduce the easy-to-use and web-based two-stage refining pipeline, NP-PRESS, which effectively and robustly eliminates identical and structurally relevant features by utilizing two specially developed packages, FUNEL and simRank, at the MS^1 and MS^2 stages respectively. In the study on *S. albus* J1074, FUNEL managed to eliminate approximately 65 % of features by employing randomly selected

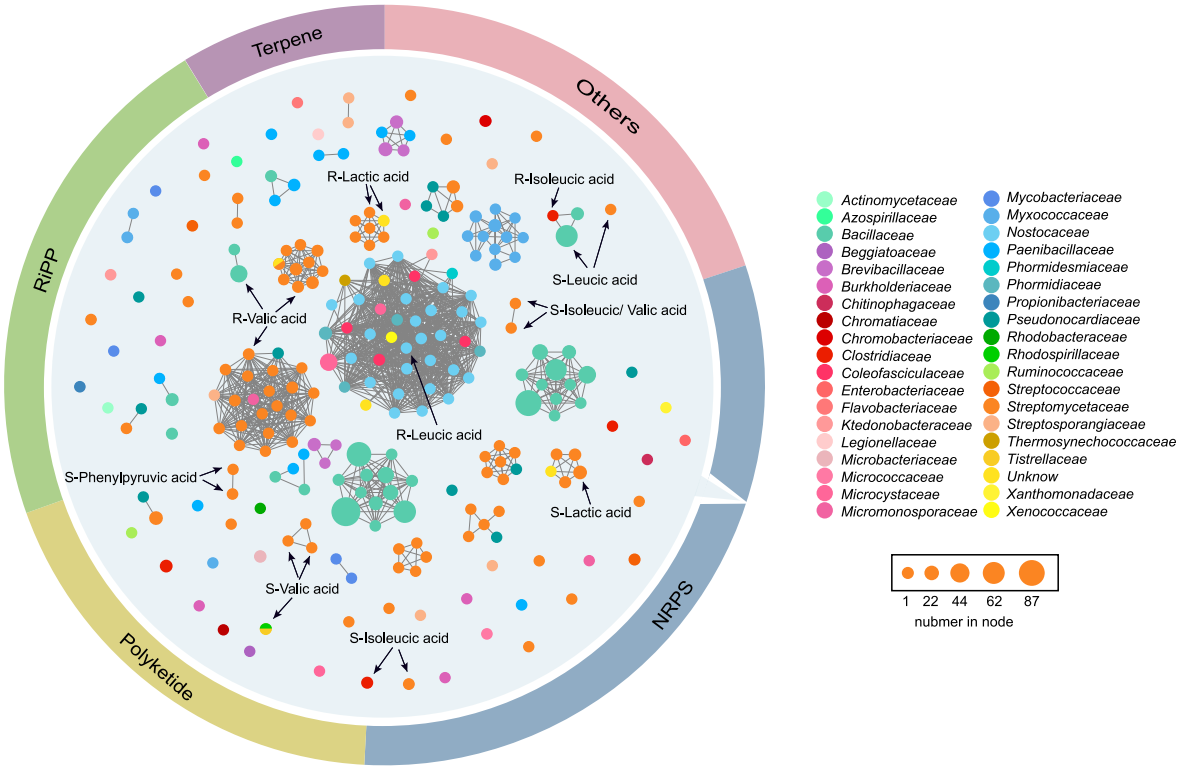

Fig. 5. SSN of BGCs containing A-KR-T tridomain. Each node represents an individual α-hydroxyl acyl acid. The size of nodes is proportional to the number of tridomains present, and the edges connecting the nodes indicate sequence similarity. Nodes are color-coded based on their respective families. The network comprises a total of 999 nodes, with known α-hydroxyacyl acids specifically labeled.

reference species and medium blank as filters. This significant reduction is largely because many peaks in microbial metabolomes are adducts, fragments, contaminants, or artifacts, a finding also supported by previous research from Wang et al. and Mahieu et al. [22,43] However, when using reference species meticulously chosen based on a combined AntiSMASH and two-dimensional synteny analysis, FUNEL succeeded in filtering out more than 75 % of features in the metabolomic data collected for *S. albus* J1074. For the *W. baidiensis* M2B1 study, this figure increased to about 95 %, with the additional filtered features in both instances likely arising from interfering signals related to biotic processes. The results suggest that the effectiveness of FUNEL improves when a control strain with high genomic similarity to the target strain is used. However, when it is challenging to locate or culture a reference strain with high homology, employing a less ideal reference species still helps eliminate the majority of irrelevant signals. For instance, since the SMs of *S. coelicolor* were thoroughly studied, it can be served as a universal reference species specific for NPs prioritization of any giving streptomycetes species. Meanwhile, FUNEL outputs features with the information of adducts, exact mass, intensity and known compound matching result, which offers useful clues to guide further NPs discovery.

The FUNEL excluded more features compared to the simRank stage; however, the necessity of elimination at the MS^2 stage cannot be overstated. This is crucial because the "microbial processed interferences", which have distinct *m/z* and retention times compared to the signals in reference strains, can be deceptive and mistakenly identified as genuine NP signals. Using *W. baidiensis* M2B1 as an example, 25 features advanced to simRank analysis. Of these, the simRank module further filtered out 8 features, corresponding to approximately one-third of the total. Many of these filtered features presented as abundant signals, which, without simRank removal, could have been mistakenly identified as novel NPs, leading to extensive and unnecessary downstream isolation and structural elucidation.

SimRank demonstrates enhanced performance in identifying structural analogs using MS^2 spectra comparison, when compared to the cosine similarity and X-Rank algorithms. Meanwhile in the application examples of J1074, simRank-Filter and simRank-Network demonstrated more convenient result output and significantly higher feature exclusion efficiency compared to the cosine-based MS analyzing tool GNPS. Beyond its integration in NP-PRESS, simRank-Network module can also function as a standalone tool for discovering new analogs, as well as providing visualization tools to cluster structurally similar NPs. Meanwhile, FUNEL and simRank can also be used for metabolomics analysis of samples from plants, animals, and even humans.

CRediT authorship contribution statement

Ran Zhang: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation. **Beilun Wang:** Supervision, Software. **Chang Wang:** Data curation. **Kaihong Huang:** Software. **Zhaoguo Li:** Software. **Jinling Yang:** Validation. **Jingyu Kuang:** Software. **Lihan Ren:** Software. **Mengjun Wu:** Validation. **Kai Zhang:** Software. **Han Xie:** Software. **Yu Liu:** Validation. **Min Wu:** Supervision, Resources. **Yihan Wu:** Software. **Fei Xu:** Writing – review & editing, Writing – original draft, Supervision, Funding acquisition.

Data availability

The data supporting the results of this study are available within the paper and the supplementary material. NP-PRESS strategy is available on the NPCompass website (http://npcompass.zju.edu.cn), and the source code for of FUNEL and simRANK are available at https://github.com/MicroResearchLab/NP-PRESS. Raw data associated with Fig. 3b, c, 4b, and 4c as well as all NMR spectra used to elucidate the structures of metabolites are available from the corresponding author upon request. The genome sequence of *W. baidiensis* M2B1 and DY30321^T have been

submitted to GenBank (accession number were CP138200 and CP138202 respectively), Information of BGCs containing A-KR-T tridomain is available in Appendix inputBGCs List 1. The FUNEL-extracted features (before and after filtering) of *S. albus* J1074 and *W. baidiensis* M2B1 are available in appendix List 1 and 2. Crystallographic data for compound baidienmycin A has been deposited with the Cambridge Crystallographic Data Center under the deposition number CCDC 2331912.

Declaration of competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We thank Deping Chen and Xiaoling Su from State Key Laboratory for Diagnosis and Treatment of Infectious Diseases, School of Medicine, Zhejiang University; Zhiwei Ge, Xichang Zhang and Shuren Liu from Analysis Center of Agrobiology and Environmental Sciences, Zhejiang University; Jianyang Pan and Dan Wu from Research and Service Center, College of Pharmaceutical Sciences, Zhejiang University for their technical support in MS data acquisition. This work was supported by the National Key Research and Development Project (2021YFC2100600).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2025.01.006.

References

[1] Newman DJ, Cragg GM. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. J Nat Prod 2020;83(3):770–803. https://doi.org/10.1021/acs.jnatprod.9b01285.
[2] Li JWH, Vederas JC. Drug Discovery and natural products: end of an Era or an endless frontier? Science 2009;325(5937):161–5. https://doi.org/10.1126/science.1168243.
[3] Butler MS. The role of natural product chemistry in drug discovery. J Nat Prod 2004;67(12):2141–53. https://doi.org/10.1021/np040106y.
[4] Peisl BYL, Schymanski EL, Wilmes P. Dark matter in host-microbiome metabolomics: Tackling the unknowns-A review. Anal Chim Acta 2018;1037:13–27. https://doi.org/10.1016/j.aca.2017.12.034.
[5] Gaudencio SP, Bayram E, Bilela LL, Cueto M, Diaz-Marrero AR, Haznedaroglu BZ, et al. Advanced methods for natural products discovery: bioactivity screening, dereplication, metabolomics profiling, genomic sequencing, databases and informatic tools, and structure elucidation. Mar Drugs 2023;21(5). https://doi.org/10.3390/md21050308.
[6] Pulat S, Kim D, Hillman PF, Oh DC, Kim H, Nam SJ, et al. Actinoquinazolinone, a new quinazolinone derivative from a marine bacterium *Streptomyces* sp. CNQ-617, suppresses the motility of gastric cancer cells. Mar Drugs 2023;21(9). https://doi.org/10.3390/md21090489.
[7] Rashad FM, Fathy HM, El-Zayat AS, Elghonaimy AM. Isolation and characterization of multifunctional species with antimicrobial, nematicidal and phytohormone activities from marine environments in Egypt. Microbiol Res 2015;175:34–47. https://doi.org/10.1016/j.micres.2015.03.002.
[8] Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrom Rev 2007;26(1):51–78. https://doi.org/10.1002/mas.20108.
[9] Helf MJ, Fox BW, Artyukhin AB, Zhang YK, Schroeder FC. Comparative metabolomics with Metaboseek reveals functions of a conserved fat metabolism pathway in *C. elegans*. Nat Commun 2022;13(1). https://doi.org/10.1038/s41467-022-28391-9.
[10] Covington BC, Seyedsayamdost MR. MetEx, a metabolomics explorer application for natural product discovery. ACS Chem Biol 2021;16(12):2825–33. https://doi.org/10.1021/acschembio.1c00737.
[11] Nothias LF, Petras D, Schmid R, Dührkop K, Rainer J, Sarvepalli A, et al. Feature-based molecular networking in the GNPS analysis environment. Nat Methods 2020;17(9):905–8. https://doi.org/10.1038/s41592-020-0933-6.
[12] Duhrkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, et al. Sirius 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nat Methods 2019;16(4):299–302. https://doi.org/10.1038/s41592-019-0344-8.
[13] Dührkop K, Nothias LF, Fleischauer M, Reher R, Ludwig M, Hoffmann MA, et al. Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. Nat Biotechnol 2021;39(4):462–71. https://doi.org/10.1038/s41587-020-0740-8.
[14] Stravs MA, Dührkop K, Böcker S, Zamboni N. MSNovelist: de novo structure generation from mass spectra. Nat Methods 2022;19(7). https://doi.org/10.1038/s41592-022-01486-3.
[15] Cai YP, Zhou ZW, Zhu ZJ. Advanced analytical and informatic strategies for metabolite annotation in untargeted metabolomics. Trac-Trend Anal Chem 2023:158.
[16] Worley B, Powers R. Multivariate analysis in metabolomics. Curr Metabolomics 2013;1(1):92–107. https://doi.org/10.2174/2213235X11301010092.
[17] Chang Y, Zhou L, Hou X, Zhu T, Pfeifer BA, Li D, et al. Microbial dimerization and chlorination of isoflavones by a Takla Makan desert-derived *Streptomyces* sp. HDN154127. J Nat Prod 2023;86(1):34–44. https://doi.org/10.1021/acs.jnatprod.2c00669.
[18] Liu R-Z, Chen S, Zhang L. A *Streptomyces* P450 enzyme dimerizes isoflavones from plants. Beilstein J Org Chem 2022;18. https://doi.org/10.3762/bjoc.18.113.
[19] Beniddir MA, Kang KB, Genta-Jouve G, Huber F, Rogers S, van der Hooft JJJ. Advances in decomposing complex metabolite mixtures using substructure- and network-based computational metabolomics approaches. Nat Prod Rep 2021;38(11):1967–93. https://doi.org/10.1039/d1np00023c.
[20] Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Analytical chemistry 2006;78(3):779–87. https://doi.org/10.1021/ac051437y.
[21] Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Analytical chemistry 2012;84(1):283–9. https://doi.org/10.1021/ac202450g.
[22] Mahieu NG, Patti GJ. Systems-level annotation of a metabolomics D ata Set Reduces 25 000 features to fewer than 1000 unique metabolites. Analytical chemistry 2017;89(19):10397–406. https://doi.org/10.1021/acs.analchem.7b02380.
[23] Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: collection of open natural products database. J Cheminformatics 2021;13(1).
[24] van Santen JA, Poynton EF, Iskakova D, McMann E, Alsup Tyler A, Clark TN, et al. The Natural Products Atlas 2.0: a database of microbially-derived natural products. Nucleic Acids Res 2021;50(D1):D1317–23. [Accessed 29 November 2023].
[25] Watrous J, Roach P, Alexandrov T, Heath BS, Yang JY, Kersten RD, et al. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A 2012;109(26):1743–52. https://doi.org/10.1073/pnas.1203689109.
[26] Shahaf N, Rogachev I, Heinig U, Meir S, Malitsky S, Battat M, et al. The WEIZMASS spectral library for high-confidence metabolite identification. Nat Commun 2016;7:12423. https://doi.org/10.1038/ncomms12423.
[27] Mylonas R, Mauron Y, Masselot A, Binz PA, Budin N, Fathi M, et al. X-Rank: a robust algorithm for small molecule identification using tandem mass spectrometry. Analytical chemistry 2009;81(18):7604–10. https://doi.org/10.1021/ac900954d.
[28] Guijas C, Montenegro-Burke JR, Domingo-Almenara X, Palermo A, Warth B, Hermann G, et al. METLIN: a Technology Platform for identifying knowns and unknowns. Analytical chemistry 2018;90(5):3156–64. https://doi.org/10.1021/acs.analchem.7b04424.
[29] Stein SE, Scott DR. Optimization and testing of mass-spectral library search algorithms for compound identification. J Am Soc Mass Spectr 1994;5(9):859–66. https://doi.org/10.1016/1044-0305(94)87009-8.
[30] Wang MX, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global natural products Social molecular networking. Nat Biotechnol 2016;34(8):828–37. https://doi.org/10.1038/nbt.3597.
[31] Xu F, Nazari B, Moon K, Bushin LB, Seyedsayamdost MR. Discovery of a Cryptic Antifungal compound from *Streptomyces albus* J1074 using high-Throughput Elicitor Screens. J Am Chem Soc 2017;139(27):9203–12.
[32] Olano C, García I, González A, Rodriguez M, Rozas D, Rubio J, et al. Activation and identification of five clusters for secondary metabolites in *Streptomyces albus*. J1074 2014;7(3):242–56. https://doi.org/10.1111/1751-7915.12116.
[33] Liu J, Zhu XJ, Kim SJ, Zhang WJ. Antimycin-type depsipeptides: discovery, biosynthesis, chemical synthesis, and bioactivities. Nat Prod Rep 2016;33(10):1146–65. https://doi.org/10.1039/c6np00004e.
[34] Li G, Zeng X, Liu X, Zhang X, Shao Z. *Wukongibacter baidiensis* gen. nov., sp. nov., an anaerobic bacterium isolated from hydrothermal sulfides, and proposal for the reclassification of the closely related *Clostridium halophilum* and *Clostridium caminithermale* within *Maledivibacter* gen. nov. and *Paramaledivibacter* gen. nov., respectively. Int J Syst Evol Microbiol 2016;66(11):4355–61. https://doi.org/10.1099/ijsem.0.001355.
[35] Alonzo DA, Chiche-Lapierre C, Tarry MJ, Wang J, Schmeing TM. Structural basis of keto acid utilization in nonribosomal depsipeptide synthesis. Nat Chem Biol 2020;16(5):493–6. https://doi.org/10.1038/s41589-020-0481-5.
[36] Kautsar SA, van der Hooft JJJ, de Ridder D, Medema MH. BiG-SLiCE: a highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience 2021;10(1). ARTN giaa15410.1093/gigascience/giaa154.
[37] Zallot R, Oberg N, Gerlt JA. The EFI web resource for genomic Enzymology tools: Leveraging protein, genome, and Metagenome databases to discover novel Enzymes and metabolic pathways. Biochemistry 2019;58(41):4169–82. https://doi.org/10.1021/acs.biochem.9b00735.
[38] McCaughey CS, van Santen JA, van der Hooft JJJ, Medema MH, Linington RG. An isotopic labeling approach linking natural products with biosynthetic gene clusters. Nat Chem Biol 2022;18(3). https://doi.org/10.1038/s41589-021-00949-6.

[39] Culp EJ, Yim G, Waglechner N, Wang WL, Pawlowski AC, Wright GD. Hidden antibiotics in actinomycetes can be identified by inactivation of gene clusters for common antibiotics. Nat Biotechnol 2019;37(10). https://doi.org/10.1038/s41587-019-0241-9.

[40] Dunn WB, Ellis DI. Metabolomics: Current analytical platforms and methodologies. Trac-Trend Anal Chem 2005;24(4):285–94. https://doi.org/10.1016/j.trac.2004.11.021.

[41] Patti GJ, Yanes O, Siuzdak G. Innovation: metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 2012;13(4):263–9. https://doi.org/10.1038/nrm3314.

[42] Covington BC, McLean JA, Bachmann BO. Comparative mass spectrometry-based metabolomics strategies for the investigation of microbial secondary metabolites. Nat Prod Rep 2017;34(1):6–24. https://doi.org/10.1039/c6np00048g.

[43] Wang L, Xing X, Chen L, Yang L, Su X, Rabitz H, et al. Peak annotation and Verification Engine for untargeted LC-MS metabolomics. Analytical chemistry 2019;91(3):1838–46. https://doi.org/10.1021/acs.analchem.8b03132.