NP Analyst: An Open Online Platform for Compound Activity Mapping

Sanghoon Lee, Jeffrey A. van Santen, Nima Farzaneh, Dennis Y. Liu, Cameron R. Pye, Tim U. H. Baumeister, Weng Ruh Wong, and Roger G. Linington*

Cite This: ACS Cent. Sci. 2022, 8, 223-234

ABSTRACT: Few tools exist in natural products discovery to integrate biological screening and untargeted mass spectrometry data at the library scale. Previously, we reported Compound Activity Mapping as a strategy for predicting compound bioactivity profiles directly from primary screening results on extract libraries. We now present NP Analyst, an open online platform for Compound Activity Mapping that accepts bioassay data of almost any type, and is compatible with mass spectrometry data from major instrument manufacturers via the mzML format. In addition, NP Analyst will accept processed mass spectrometry data from the MZmine 2 and GNPS open-source platforms, making it a versatile tool for integration with existing discovery workflows. We demonstrate the utility of this new tool for both the dereplication of known compounds and the discovery of novel bioactive natural products using a challenging low-resolution antimicrobial bioassay data set. This new platform is available at www.npanalyst.org.

INTRODUCTION

Traditionally, natural products discovery has been a linear endeavor, with projects completed on a sample-by-sample basis. This approach has been successful in discovering some of our most valuable therapeutics such as Taxol, rapamycin, and cyclosporine A, but is becoming increasingly inefficient as rates of rediscovery rise. The development of accurate and accessible omics technologies, including genome sequencing, untargeted metabolomics, and high-throughput screening, is changing the discovery landscape for this field. Omics-data integration between these platforms has the potential to improve both the speed and accuracy of compound discovery by leveraging information from orthogonal data types at the system level, rather than evaluating individual samples sequentially. Currently, however, there are few open access tools that integrate metabolomics and bioassay data sets, limiting applications of these approaches in the natural products community.

Several approaches exist for integrating biological activity data with untargeted metabolomics data. Bioactivity-based molecular networking predicts the bioactivity of each MS feature by calculating the Pearson correlation between activity profiles and intensity profiles for each feature in the sample set using a combination of open-source tools and custom R scripts. Ory et al. developed a method termed Find Biactive COMpounds (FiBiCo) which combines results from four different statistical models (Spearman, F-PCA, PLS, PLS-DA) to prioritize MS features that correlate positively with bioactivity profiles. This tool is written in R and is available from the Supporting Information of the original article. Finally, Olivon et al. have developed a strategy that incorporates metabolomic, taxonomic, and bioactivity data into a single data matrix for bioactive compound prioritization. This approach color codes molecular networks based on the biological activities of active fractions to permit manual prioritization of...
MS features. Several of these methods are labor-intensive to implement and require tailored and customized workflows; none are available as stand-alone platforms for data integration that include both data analysis and data visualization components.

Previously we developed an approach, termed Compound Activity Mapping, to directly predict bioactive constituents and modes of action from complex mixtures using a combination of image-based screening and untargeted metabolomics.\textsuperscript{11} We now present NP Analyst, a versatile, open access platform for Compound Activity Mapping (www.npanalyst.org). Importantly, NP Analyst is designed to work with biological data from any assay platform and accepts mass spectrometry in several common data formats including the standard open mzML format,\textsuperscript{12} output peak lists from MZmine 2,\textsuperscript{13} and network files from the Global Natural Products Social molecular networking platform (GNPS).\textsuperscript{14} The inclusion of these input formats makes NP Analyst compatible with bioassay data from most bioassay types and MS data from all of the major instrument manufacturers. We have developed new strategies for processing metabolomics data and integrating these results with bioactivity profiles, and have packaged these tools in an open access online environment. NP Analyst provides the research community with a new discovery platform designed to accelerate compound dereplification, highlight priority metabolites for isolation, and generate global network views of biologically active chemical space for large extract libraries.

\section*{RESULTS}

\textbf{Platform Design.} Originally, Compound Activity Mapping was designed to work with one specific biological assay (cytological profiling) and relied on highly customized in-house scripts for data integration.\textsuperscript{11,15} In designing the NP Analyst platform (Figure 1), we identified three key attributes required for adoption by the natural products community:

- an open, freely accessible online interface
- a workflow capable of accepting data from any biological assay
- mass spectrometry data import functions compatible with both open data formats for raw data and common mass spectrometry data processing packages for processed data

Developing NP Analyst as an online resource offered several advantages over desktop deployment. It eliminated the need to support multiple operating systems and allows dynamic allocation of storage and processing resources, making the platform both responsive and scalable. In addition, updates and upgrades can be made immediately available without the need for software updates, improving the user experience.

The online interface includes a suite of data validation and quality control checks to assist users with data upload and formatting. This feedback identifies issues with input data and allows users to make corrections before to job submission, reducing job failures related to data format issues. Each results file is assigned a unique job number which is used to create a permanent hyperlink to the results pages, providing a facile mechanism for collaborators to retrieve and share data.

Compound Activity Mapping works by comparing the distribution of mass spectrometry features in the sample set against the biological signatures of each of these samples. This requires processing of mass spectrometry data to generate a complete list of all unique mass spectrometry features ($m/z$ vs retention time pairs) in the sample set and the distribution of these features among the samples in the set. To ensure that NP Analyst would work for the broadest cross section of users we created a new data processing pipeline that requires only MS\textsuperscript{1} level MS data in the standard mzML open data format. This strategy is sufficient to describe the vast majority of unique mass spectrometry features in the data set without requiring MS\textsuperscript{2} data of a specific format from one of the myriad possible data acquisition modes. This approach also enables laboratories without MS\textsuperscript{2} capability (e.g., UPLC-TOF instruments) to use the NP Analyst platform.

To extend this approach beyond our original study using image-based screening data, we have developed a new algorithm for data integration that is agnostic to bioassay data format. Unlike the previous strategy, which required normalized data sets with values between $\pm 1$ and +1 and a precise number of features, this new approach will accept data sets with any number of biological features and accepts different assay data formats (inhibition/no inhibition, percent growth, etc.). This eliminates the requirement for users to have access to sophisticated screening infrastructure (e.g., high-content imaging microscope) and permits the use of existing biological screening data in the NP Analyst platform. The only requirement is that the bioactivity file contains multiple biological features to create biological fingerprints for each sample. These can either be multiple readouts from a single assay (e.g., gene expression profiles) or single readouts from multiple assays (e.g., activity across a panel of bacterial pathogens). These data should be provided as a flat CSV file containing one row for each sample in the set and one column for each bioassay readout. For full instructions on bioassay file formatting requirements see the online documentation at https://liningtonlab.github.io/npanalyst_documentation/NPAnalyst/file-import/

\textbf{Algorithms.} NP Analyst contains three main functions; an optional step to process the mass spectrometry data to create a list of unique MS features, feature score determination based

---

\textbf{Figure 1.} Structure of the NP Analyst platform. Users input both mass spectrometry files and biological activity data for sample sets, which are scored to prioritize compounds with strong predicted biological activities for isolation and secondary screening.
on bioactivity profiles, and generation of output files for visualization.

**Mass Spectrometry Data Processing.** Mass spectrometry data can be uploaded in one of three formats: as individual mzML files for each sample, as a comprehensive peak list obtained through third-party software (e.g., MZmine 2), or as a GNPS-derived graphML network file. In the first case (mzML), NP Analyst will perform replicate comparison and feature alignment on the individual files as discussed below. In the latter two cases (MZmine 2 and GNPS), data processing and alignment are performed in these external platforms, and only the final aligned peak list is supplied to NP Analyst. This allows users to either process their data in GNPS or MZmine 2 using existing workflows or use the alignment and replicate comparison tools built into the NP Analyst platform.

For the mzML input, files are uploaded as peak picked centroided data. If technical replicates are available, then samples are first compared to identify signals that are present consistently between replicates. For n replicates, the default requirement is present in n − 1 samples. Because both retention times (rt) and m/z values can vary between analytical runs, a processing method to align replicate signals is required. We designed an alignment method based on R-trees that dynamically groups signals between samples into groups that conform to allowed tolerances of rt and m/z values (**Experimental Section**). This approach ensures that groups remain limited to the defined errors in both dimensions by subdividing groups that grow to include members with too wide a range of rt of m/z values. An advantage of this R-trees approach is that it is easily extensible to additional dimensions. This permits future incorporation of additional data types, such as drift times/CCS values from ion mobility spectrometry. Following replicate comparison, consensus signals from each sample are aligned using the same R-trees method. This analysis yields a single file containing all the unique mass spectral features in the sample set, describing these signals’ distribution within the set.

For MZmine 2 data, preprocessing, replicate comparison, and alignment are performed using in-built tools in the MZmine package, and a single aligned peak list is exported into NP Analyst without further manipulation. For GNPS data the MZmine package, and a single aligned peak list is exported into and alignment are performed using in-built tools in the standard GNPS graphML network containing all MS features that pass the Activity and Cluster Score cutoffs during the job submission process. The analysis workflow calculates both scores for every MS feature and retains MS features that meet the minimum values for both parameters. Increasing the cutoff values for either score reduces the complexity of the results files by removing MS features that do not have strong bioactivity profiles and/or do not correlate with specific biological phenotypes. These two parameters can be used in concert to select either MS features with strong biological profiles (Activity Score) or features with consistent activity profiles independent of spectrum of activity (Cluster Score). The range of Activity Score values depends on the bioassay data format. For bioassay data that has been normalized to a scale of 0 to 1, the maximum Activity Score is equal to the number of bioassay parameters. By contrast, Cluster Score values always range from −1 to +1. In practice, setting low positive cutoff values (0.3 for Activity Score, 0.1 for Cluster Score) is sufficient to remove many of the inactive features, simplifying both data visualization and data interpretation. The definition of “high” scores is dependent on the type of bioassay data used, but as a general guide values greater than 30% of maximum for Activity Score and 50% of maximum for Cluster Score can be considered “high”.

**Data Visualization. Network Visualization.** NP Analyst provides three complementary data visualization options: network, scatter plot, and community visualizations. The network view (**Figure S1**) displays samples and associated mass spectrometric features, distributed by chemical relatedness. Extract nodes (squares) are connected by edges to their associated mass spectrometry features (circles), with the network containing all MS features that pass the Activity and Cluster Score filters. The value of this representation is that samples are grouped together in the network based on the presence of shared bioactive mass spectrometric features. MS features that possess high interconnectivity within each group derive from molecules predicted to be responsible for the observed biological phenotypes. These can include different adducts and in-source fragments from the same molecule or features from groups of related molecules with similar biological profiles. Therefore, the network view is valuable for describing the range and breadth of predicted bioactive metabolites within the sample set and how they are distributed between samples.

As part of the analysis pipeline the network is divided into communities using the Louvain method for community detection. Each community is given a unique color and community ID used in the Community view (**Figure S2**). These communities contain highly interrelated nodes indicative of shared biologically active MS features and are a helpful resource for selecting priority extracts for further investigation. Switching the “Color by Community” toggle in calculating the sum of the mean of the squares of the bioactivity values for each extract containing a given mass spectrometric feature. For example, if a feature is present in three samples, the Activity score is calculated by averaging the squares of the values in each column of the bioassay data and summing the resulting values.

Cluster Score is a measure of the consistency of the biological fingerprints between all of the samples that contain a given mass spectrometric feature. It is determined by taking the average Pearson similarity scores between the biological fingerprints of all extracts containing a given mass spectrometric feature.

Users can define Activity and Cluster Score cutoffs during the job submission process. The analysis workflow calculates both scores for every MS feature and retains MS features that meet the minimum values for both parameters. Increasing the cutoff values for either score reduces the complexity of the results files by removing MS features that do not have strong bioactivity profiles and/or do not correlate with specific biological phenotypes. These two parameters can be used in concert to select either MS features with strong biological profiles (Activity Score) or features with consistent activity profiles independent of spectrum of activity (Cluster Score). The range of Activity Score values depends on the bioassay data format. For bioassay data that has been normalized to a scale of 0 to 1, the maximum Activity Score is equal to the number of bioassay parameters. By contrast, Cluster Score values always range from −1 to +1. In practice, setting low positive cutoff values (0.3 for Activity Score, 0.1 for Cluster Score) is sufficient to remove many of the inactive features, simplifying both data visualization and data interpretation. The definition of “high” scores is dependent on the type of bioassay data used, but as a general guide values greater than 30% of maximum for Activity Score and 50% of maximum for Cluster Score can be considered “high”.

**Bioactivity Profile Integration.** MS features in the unique feature list are scored for the strength and consistency of their predicted biological profiles using the Activity Score and Cluster Score metrics employed in our previous study. NP Analyst has been configured to automatically adjust to the dimensionality of the bioassay data file (e.g., 5 bioassay features vs 250) and is capable of handling data from most assay readouts (IC50, percent inhibition etc.). In addition, Boolean data types (e.g., True/False, Inhibition/no inhibition, Yes/No) are also accepted. The only requirements are that each data column must contain a consistent data type (i.e., no columns containing a mixture of numerical and Boolean data) and that missing data must be represented by empty cells rather than “not tested” or other text entries.

Activity Score is a measure of the strength of the phenotype for a given mass spectrometry feature. It is determined by
the network view changes the network color scheme from activity color-coding to node colors based on community membership, which is valuable for understanding how the full sample set is distributed between communities and how those communities are interconnected. Finally, clicking on any node highlights that node and the nodes to which it is directly connected, providing a mechanism for exploring the network. The graph is interactive, permitting zoom and pan functions, as well as autolabeling based on zoom level.

**Scatter Plot Visualization.** The scatter plot visualization (Figure S3) presents a plot of the bioactive MS features (retention time vs m/z ratio), providing a more targeted, compound-centric data representation of the full data set. This plot can be filtered by Activity Score and Cluster Score to retain only features with strong bioactivity predictions. In addition, users can select samples of interest from the interactive list below the plot to retain only MS features from a defined subset of samples. This provides an interactive interface for assessing the predicted bioactive mass spectrometry features for any sample or set of samples in the data set and is particularly valuable when combined with information from the Community view.

**Community Visualization.** The community visualization provides data on bioactive MS features for specific communities from the Louvain community detection algorithm (Figure 2). Each community page contains a network view of the extract and MS feature nodes in that community, two scatter plots (rt vs m/z and Cluster Score vs Activity Score), and a bioactivity heatmap for the extracts in the community. Together these plots allow users to assess the biological similarities between samples within each community, select bioactive MS features that interconnect these samples, and determine the rt and m/z values for these features for subsequent dereplication and compound isolation.

**User Interface.** The user interface is designed to assist users with data import, quality control, and visualization, while prioritizing ease of use. The platform requires no registration or login details and is W3C compliant to ensure functionality on all major operating systems and browsers. To start a new analysis, users optionally enter an email address (in order to
Upon completion, the results section is displayed, including interactive tabs for scatter plot, network, and community views, as well as a downloads page to export the results files. Export options include a graphML network file for use in network visualization tools (Cytoscape, Gephi) and CSV for use in spreadsheets and graph plotting tools (e.g., Excel, Tableau, JMP, Spotfire).

For advanced users, the underlying processing algorithms are freely available as a command line tool via a Docker container at https://github.com/liningtonlab/npanalyst. The advantage of local installation is that it eliminates the slow step of uploading mass spectrometry data to the online server; a particular issue with mzML files. The disadvantages are that (i) programming experience is required to deploy the Docker container, (ii) some of the client-side quality control steps for individual files built into the web interface are not included in the command line tool, and (iii) the online interface cannot be used for data visualization.

**Bioactive Compound Discovery.** In order to evaluate the value of NP Analyst for bioactive compound discovery, we analyzed a set of 925 prefractions from our in-house marine Actinobacterial strain library. For biological profiling we deliberately selected a low-density bioassay data set that would test the performance limits of the platform. Data from our previously developed BioMAP antibacterial profiling platform had been combined to generate a data set containing inhibition/no inhibition results against 15 bacterial pathogens (6 Gram + ve, 9 Gram − ve) for all 925 prefractions. This bioassay offers a maximum of 32,768 unique phenotypes (2^{15}). However, antibacterial compounds tend to have broad activity against either Gram + ve or Gram − ve strains (or both), meaning that the effective number of phenotypes is...
Discovery of Dracolactam C. Community 12 contained 23 prefractions with very similar biological profiles (Figure 4A). Sixteen of the 23 prefractions in this community were connected by a single MS feature with an m/z of 452.2788 and a retention time of 3.12 min. The molecule responsible for this MS feature was therefore prioritized for isolation. Refermentation of the producing organism followed by mass-guided purification yielded a compound with a precursor [M + H]+ peak at 470.2888 and a prominent [M − H2O + H]+ peak at 452.2791. Dereplication against the Natural Products Atlas database18 and comparison of the 'H NMR data against the published literature identified this compound as the polyene macrolactam micromonolactam (1).19,20 However, during the isolation of this metabolite we identified two compounds from this fraction with similar MS features, one of which was isobaric with micromonolactam. To determine the identity of the active species unequivocally, we isolated these two additional metabolites and identified them using extensive 1D and 2D NMR experiments. One of these compounds was the known metabolite dracolactam A (2),15 which is proposed to derive from the intramolecular Diels−Alder cyclization of micromonolactam (Scheme S1). The second was a new compound, dracolactam C (3), which was identified as a different intramolecular cyclization product of micromonolactam (Scheme S1). For a full description of the structure elucidation of this new compound, see the Supporting Information and Figure S4. Screening of all three compounds (Figure 4B) in the BioMAP panel revealed that micromonolactam possessed a similar antibacterial profile to the profile for community 12, while dracolactams A and C were largely inactive (Table 1).

These bioassay results align with the prediction from NP Analyst, which prioritized a single MS feature (m/z 452.2788; rt 3.12 min). Inspection of the original UPLC-HRMS metabolomics data for these fractions and comparison against the retention times for compounds 1−3 using the same analytical method revealed that all three compounds were present in multiple prefractions in the library. However, only the distribution of micromonolactam aligned with the activity profile for community 12 (Figure 4C), explaining why only one of these metabolites was predicted as the active component. This represents the first instance of reported biological activity for micromonolactam.

Dereplication of Collismycin Compound Family. Community 15 (Figure 5A) presented a more complicated scenario. The community contained three prefractions with broad-spectrum antibiotic activities and a large number of candidate bioactive features. However, six of these features possessed strong Activity Scores (large diameter red nodes connecting prefractions RLUS-2108C and RLUS-2108D in Figure 5A, highlighted rows in Figure 5E). A review of the retention times for these features in the scatter plot view (Figure 5B) highlighted three sets of features with related retention times (2.64, 3.42, and 3.97 min) that were consistent with adducts and fragments from three separate molecules (Figure S5). This was supported by the peak shapes for extracted ion chromatograms (EICs) for these features in the original MS data, which were grouped into three sets (Figure 5D).

Isolation and characterization of two of these metabolites by 1D and 2D NMR experiments (Figures S22−S25) identified the related bipyridyl compounds collismycin A (4) and SF2738D (5) (Figure 5C).15,21,22 The third metabolite was not present in sufficient quantity for analysis by NMR but was isobaric with collismycin A and possessed very similar MS3 fragmentation and UV absorbance spectra (Figure S6), consistent with the known stereoisomer collismycin B.22 Surprisingly, screening of collismycin A and SF2738D in the BioMAP assay revealed that, while collismycin A possessed antibacterial activity against a range of strains, SF2738D was completely inactive. This result is in line with previously

Figure 4. (A) Community 12 network. Prefractions illustrated as square gray nodes. Predicted bioactive MS features illustrated as circular red nodes. m/z values for priority MS feature node annotated on the network. Peripheral nodes removed for clarity. Full community presented in Figure S2. (B) Structures of isolated compounds. (C) Extracted ion chromatogram (EIC) traces for predicted bioactive MS feature (top), micromonolactam (middle), and dracolactam C (bottom) illustrating alignment between the predicted bioactive constituent and the purified bioactive metabolite (micromonolactam) and the absence of predicted bioactivity for the inactive isomer dracolactam C.
Table 1. Comparison of BioMAP Screening Data for Selected Communities

| Tested compound | Community 12 | Community 15 | Community 16 |
|-----------------|--------------|--------------|--------------|
| BioMAP results¹ | | | |
| B. subtilis | E. faecium | L. ivanovii | MRSA | S. aureus | S. epidermidis |
| | | | | Y. pseudotuberculosis |
| | | | | |

¹Each color (HEX code) indicates activity against the test organism in the BioMAP screening panel. (1) Red (ff6600) – A. baumannii, (2) Orange (ff6066) – B. subtilis, (3) Tangerine Yellow (ff00cc) – K. aerogenes, (4) Fluorescent Yellow (#ccff00) – E. coli, (5) Bright Green (#66ff00) – E. faecium, (6) Lime Green (#00ff80) – L. ivanovii, (7) Spring Green (#00ff66) – MRSA, (8) Bright Turquoise (#00ffcc) – O. anthropi, (9) Deep Sky Blue (#00ccff) – P. aeruginosa, (10) Navy Blue (#0066ff) – P. alcalifaicens, (11) Blue (#0000ff) – S. aureus, (12) Electric Indigo (#6600ff) – S. epidermidis, (13) Electric Purple (#cc00ff) – S. enterica, (14) Hot Magenta (#ff00cm3) – Y. cholaera, (15) Vivid Raspberry (#ff0066) – Y. pseudotuberculosis. *No screening results available for collismycin A or amychelin C against S. epidermidis.

Published screening data for these compounds, which did not identify any antimicrobial activity for SF2738D (Table 1).²¹

This result highlights one of the limitations of the NP Analyst platform. In cases where inactive compounds are always coproduced with bioactive molecules in microbial cultures it is not possible to differentiate between the activities of the two metabolites. The central premise of the method is that metabolites that are present in both active and inactive fractions will have low Activity and Cluster Scores. However, inactive metabolites that are only present in active fractions will have high Activity and Cluster Scores, provided that the active fractions they are present in all have similar biological profiles. This issue is particularly acute for small communities with low numbers of samples (<10), as the chance of coexpression of active and inactive metabolites is higher if the number of samples is low. Therefore, users should take care to consider the values for Cluster and Activity Scores when selecting priority MS features to isolate from small communities containing multiple candidate masses. Specifically, users should examine the connections between MS features and samples, and the activity profiles of those samples on the Communities page. This information should be used in combination with the relative magnitudes of the Activity and Cluster Scores to select priority molecules for downstream isolation and biological evaluation. It is recommended that users prioritize features with the highest interconnectivity between samples with similar activity profiles, and that MS features with additional connections to inactive samples are given lower priority.

**Discovery of Amychelin C.** Finally, community 16 contained 5 prefractions, 2 of which possessed very similar bioactivity profiles and were connected by 3 mass spectrometry features. Evaluation of the scatter plot and EIC traces for these prefractions indicated the presence of two related molecules, one of which contained both the precursor [M + H]+ feature at m/z 753.3243 and a prominent mass fragment at m/z 623.2319 (Figure 6A). Searching the Natural Products Atlas database did not yield any candidate structures for this compound pair, so this community was prioritized for further investigation.

Fermentation of the producing organism, followed by mass-guided isolation by HPLC-MS, yielded 1.18 mg of an off-white solid with an m/z of 753.3053. 1D and 2D NMR analyses, coupled with high-resolution mass spectrometry, suggested a formula for this new metabolite of C_31H_45N_8O_14 (m/z 753.3053 [M + H]^+ (calcld. 753.3050)). Examination of the gCOSY, gHSQC, and gHMBC spectra identified three subunits (Figure 6B). Signals for subunit B were broad and asymmetric, and possessed integrations in multiples of three, suggesting three repeating motifs. Examination of the MS² spectrum revealed sequential neutral losses consistent with the loss of the amino acid subunits proposed in Figure 6C. This analysis, coupled with key gHMBC correlations between subunits (Figure S7A), afforded the planar structure for this new metabolite, amychelin C (6; Figure 6C). Amychelin C is related to a previously reported siderophore, amychelin (7), but differs by the inclusion of a methyl-oxazoline moiety in place of the oxazoline subunit present in amychelin. To determine the absolute configurations of the amino acid-derived stereocenters we performed Marfey’s analysis (Figure S8), which identified the presence of l-threonine, l-ornithine, and a 2:1 ratio of d- and l-serine.
Because amychelin C contained both D- and L-serine, an additional experiment was required to determine the position of the L-serine residue. Refermentation of the producing organism in the presence of a 1:1 mixture of [1,2,3-13C]-(L)-serine and [1,2,3-12C]-(D)-serine yielded an isotopically labeled version of amychelin C with an increased [M + H]+ signal of 3 Da. Interpretation of the MS fragmentation data for this isotopically labeled derivative (Figure S9) identified the position of isotopic labeling (Figure 6C), completing the full absolute configurational analysis. Interestingly, amychelin C is enantiomeric to amychelin at all shared chiral centers (Figure 6D). This finding supports a recent study on the evolutionary origins of this compound class which noted several examples with opposite configurations at some or all positions.24

Screening of purified amychelin C in the BioMAP assay recapitulated the spectrum of activity predicted from the original community, confirming this molecule as the predominant active component in this community (Table 1).

**DISCUSSION**

Prioritization of bioactive constituents from complex mixtures has been a longstanding challenge for the field of natural products.25 This is often the rate-limiting step for bioactive compound discovery and frequently leads to the rediscovery of known molecules.26,27 NP Analyst offers a target-agnostic platform for predicting biological activities of MS features directly from complex mixtures that addresses this bottleneck. Importantly, this platform is not tailored to a particular biological assay or target class, making it suitable for use with any numerical or Boolean assay readout. This extends its utility beyond drug discovery to other scientific areas. For example, this platform could be used to identify molecules related to behavioral phenotypes in chemical ecology studies, given response data for extract libraries in ecologically relevant assay systems.

In principle, every natural product will have a defined profile across all of the biological space,26,28 meaning that NP Analyst is not restricted to multiparametric assays but can also be used with individual results from a panel of different assay platforms. For example, laboratories may have results for the same extract library against a range of different bioassay targets (cancer cell lines, bacterial or fungal pathogens, protozoan parasites, viruses, etc.). Although acquired as individual results in each screen, these data are suitable for assembly into biological profiles for use in NP Analyst, provided that the data are normalized so that each column uses the same scale. Detailed instructions for data normalization are included in the Web site documentation (https://liningtonlab.github.io/npanalyst-documentation/). This tool is therefore immediately applicable in laboratories that have legacy bioassay results, provided that...
companion mass spectrometry data also exist for these samples. For optimal performance, users should carefully consider the quality of the input MS data that they import. While it is possible to run NP Analyst without any data preprocessing, this typically increases the number of candidate bioactive MS features, because raw mass lists can include many "junk" MS features. Replicate comparison can significantly reduce noise peaks in untargeted metabolomics data sets and is strongly recommended where possible. In addition, it is useful to consider the likely limit of detection of biological assays and to select an appropriate signal intensity cutoff for the MS data. While it is tempting to select the minimum possible cutoff value so as not to miss any potentially important features, if the assay is insensitive this may dramatically increase the complexity of the resulting networks without including any additional biologically informative MS features.

An advantage of the NP Analyst approach is that communities are created based on bioactive MS feature distribution, rather than sample activity profiles. Therefore, molecules with different structures but similar biological profiles will form separate communities, even if the samples that contain them have indistinguishable biological fingerprints. For example, the prefractions containing micromonolactam are part of a large group of 79 prefractions with activity against the same 6 pathogens (S. aureus, MRSA, P. alcalifaciens, B. subtilis, E. faecium, L. ivanovii). However, the micromonolactam-containing cluster includes just 23 prefractions. The remaining prefractions with this biological profile are distributed across other communities in the network, including several (such as community 3) that have clear candidate MS features for future development that are distinct from micromonolactam.

In cases where communities do not have clear individual bioactive MS features, an effective strategy for prioritizing bioactive molecules is to identify sets of active MS features with the same retention time in the plot of retention time vs m/z ratio in the community view (Figure 2C). It is widely recognized that many compounds form multiple MS features during the ionization process (adducts, fragments, and multiply charged species). Because NP Analyst scores each MS feature independently, all MS features from a given bioactive molecule will have similar Activity and Cluster Scores but different m/z ratios. These appear as a vertical "stripe" of features with the same retention time in the retention time vs m/z ratio plot (Figure 2C), which provides retention time and MS properties for bioactive molecules for direct isolation without the need for further bioassay-guided fractionation. Examples of this phenomenon are communities 3 and 15, both of which contain multiple MS features for individual priority molecules.

Overall, NP Analyst offers a powerful suite of data visualizations for exploring the bioactive component of extract libraries. The scatter plot, network, and community views provide three complementary viewpoints that should be used interdependently to identify priority MS features for further chemical and biological evaluation. Provision of the raw data from all three visualizations in the downloads page allows users to manipulate these data in other third-party data visualization tools (e.g., Tableau), enabling the development of tailored data analysis workflows as required.

Figure 6. (A) Community 16 network. Prefractions illustrated as square gray nodes. Predicted bioactive MS features illustrated as circular red nodes. Circle diameter proportional to Activity Score. Peripheral nodes removed for clarity. Full community presented in Figure S2. (B) Substructures determined from NMR data. (C) Comparison of MS fragmentation patterns for unlabeled and labeled amychein C, illustrating the position of labeled [1,2,3-13C]-(L)-serine (red moiety) and 13C-labeled carbons (blue circles). (D) Structures of amychein C (6) and amychein (7).
CONCLUSION

A new open online platform has been developed for the direct prediction of metabolite bioactivity profiles from complex mixtures. This platform accepts a wide range of bioassay data types and is compatible with both the mzML mass spectrometry open data format and output files from two commonly used open-source mass spectrometry data processing platforms (GNPS and MZmine 2). This platform is available at www.npanalyst.org.

We validated this new platform by analyzing a “low-resolution” antimicrobial bioassay data set for 925 natural product prefractions. From these results, three communities were selected for further study, leading to the isolation of three classes of bioactive metabolites including two new compounds (dracolactam C (3) and amychelin C (6)). In addition, this analysis afforded accurate predictions of biological activities for all three compound classes, including the first reported biological activity for the polyene macrolactam micromonolactam (1). Together, these results demonstrate the utility of this new platform as a rapid, accurate, and flexible strategy for the discovery of novel bioactive natural products from complex mixtures.

EXPERIMENTAL SECTION

General Experimental Procedures. Optical rotations were measured on a Model 341 Polarimeter (PerkinElmer). Ultraviolet absorption spectra were recorded on a Cary 300 UV–vis spectrophotometer (Agilent Technologies). HR-ESI-MS and MS² fragmentation spectra were recorded on a SYNAPT UPLC-EI-qTOF (Waters). NMR spectra were measured on an AVANCE II 600 MHz spectrometer equipped with a 5 mm QNP cryoprobe (Bruker). MPLC (CombiFlash, Teledyne ISCO) was carried out on RediSep Rf solid load cartridge (5 g, Teledyne ISCO). HPLC separations were performed on either a Waters autopurification system equipped with a SQ Detector 2 quadrupole MS detector or an Agilent 1200 series HPLC equipped with a binary pump and a diode array detector using either Synergi Fusion-RP or Kinetex XB-C18 columns (Phenomenex).

Collection of Samples. Sediment samples were collected into sterile 15 mL Falcon tubes by SCUBA. Sample RL09-219 was collected from Crowbar Canyon, CA. Samples RL12-176 and RL12-145 were collected from Bell Point and Dinner Island, respectively, under permit number 12-034 from the Washington Department of Fish and Wildlife.

Isolation of Bacteria. Sediment samples were plated onto Actinobacteria-specific isolation media with added antifungal and Gram-negative antibacterial agents (cycloheximide and nalidixic acid; 50 mg/L each) by radial stamping with sterile cotton swabs. Morphologically distinct colonies were picked and replated on DiCO marine broth agar plates repeatedly until pure isolates were obtained. Isolated colonies of Actinobacteria were subjected to liquid medium culturing using our standard fermentation conditions and cryopreserved as glycerol stock solutions at −80 °C.

Isolate Fermentation, Extraction, and Prefractionation. Bacterial frozen stocks were inoculated on solid media (10.0 g of glucose, 5.0 g of NZ-amine, 1.0 g of CaCO₃, 2.00 g of starch, 5.0 g of yeast extract, 20.0 g of agarose, and 1.0 L of water) and incubated at room temperature until discrete colonies became visible.

The colonies were inoculated into 40.0 mL culture tubes containing 7.0 mL of liquid media (31.2 g of Instant Ocean, 10.0 g of soluble starch, 4.0 g of peptone, 2.0 g of yeast extract per 1 L of water). Liquid cultures were incubated at room temperature and shaken at 200 rpm. After 3 days, 3.0 mL of the small-scale cultures were used to inoculate 60.0 mL of the same media in 250.0 mL wide-neck Erlenmeyer flasks with a metal spring and milk filter top. After 5 days, 50.0 mL of the medium scale cultures were inoculated into 1.0 L of media in 2.8 L wide-mouth Fernbach flasks containing a large spring with 20.0 g of Amberlite XAD-16 absorbent resin. Large-scale cultures were fermented for 7 days, and the cells and resin were filtered using Whatman glass microfiber filters and washed with sterile water. The filtered cells and resin with the filter paper were extracted with 1:1 dichloromethane (CH₂Cl₂) and methanol (MeOH). The organic extracts were removed from the cells and resin by vacuum filtration, and the extracts were evaporated under vacuum. The dried extracts were prefractionated by MPLC using a MeOH/H₂O step gradient system (10% MeOH wash (discarded), 20% MeOH, 40% MeOH, 60% MeOH, 80% MeOH, 100% MeOH, and 100% EtOAc) to afford six prefractions (A–F).

Optical rotations were measured on an AVANCE II 600 MHz spectrometer equipped with a SQ Detector 2 quadrupole MS detector or an Agilent 1200 series HPLC equipped with a binary pump and a diode array detector using either Synergi Fusion-RP or Kinetex XB-C18 columns (Phenomenex).

Compound Purification. Micromonolactam (1) and dracolactam A (2) and C (3) were isolated from prefractions RLUS-2152C and RLUS-2152D. RLUS-2152D was separated by a RP-HPLC (Phenomenex Synergi Fusion-RP 80A, 250 mm × 10.0 mm, 10 μm) and eluted using a gradient of 5% MeCN to 95% MeCN with 0.02% formic acid (0–22 min) at a flow rate of 8.0 mL/min to afford two subfractions (RLUS-2152D-1 and RLUS-2152D-2). RLUS-2152D-1 was purified by HPLC on a Phenomenex Kinetex XB-C18 column (100 mm × 4.6 mm, 26 μm, 1.0 mL/min) using an isocratic elution profile (25% MeCN with 0.02% formic acid) to afford micromonolactam (1) and dracolactam C (3). RLUS-2152C was subjected to an ODS HPLC (Phenomenex Kinetex XB-C₃₀, 100 mm × 4.6 mm, 26 μm) using an isocratic elution system (20% MeCN + 0.02% formic acid, 1.0 mL/min) to furnish dracolactam A (2).

RLUS-2108C was loaded on a RP-HPLC (Phenomenex Synergi Fusion-RP, 250 mm × 10.0 mm, 10 μm) using a gradient of 5% MeCN to 95% MeCN with 0.02% formic acid at a flow rate of 8.0 mL/min gave calsimycin A (4, RLUS-2108C-1) and two fractions (RLUS-2108C-2 and RLUS-2108C-3). The RLUS-2108C-2 fraction was purified by a RP-HPLC (Phenomenex Kinetex XB-C₃₀, 100 mm × 4.6 mm, 26 μm) using an isocratic separation (50% MeCN + 0.02% formic acid, 1.0 mL/min) to afford SF2738D (5).

Amychelin C (6) was purified from RLUS-2105D by a RP-HPLC (Phenomenex Synergi Fusion-RP, 250 mm × 10.0 mm, 10 μm) using a gradient elution profile (5% MeCN to 95% MeCN + 0.02% formic acid, 8.0 mL/min).

Antimicrobial Screening Methods and Data. Antimicrobial susceptibility tests were performed using a miniaturized high-throughput assay adapted from the broth microdilution method outlined by the Clinical and Laboratory Standards Institute (CLSI). Bacterial test strains were individually grown on fresh Nutrient Broth (NB, ATCC Medium 3) agar, Tryptic Soy Broth (TSB, ATCC Medium 18) agar, or Brain Heart Infusion (BHI, ATCC Medium 44) agar,

https://doi.org/10.1021/acscentsci.1c01108
ACS Cent. Sci. 2022, 8, 223–234
respectively (Table S4), as recommended by the American Type Culture Collection (ATCC) cultivation protocol. Individual colonies were used to inoculate 3 mL of sterile NB, TSB, or BHI media and grown overnight with shaking (200 rpm; 37 °C). Listeria ivanovii (ATCC BAA-139) and Streptococcus pneumoniae (ATCC 49619) were incubated overnight but not shaken (37 °C; 5% CO₂). Saturated overnight cultures were diluted in Cation-Adjusted Mueller-Hinton Broth (CAMHB, BBL BD) or 1:1 CAMHB/BHI media according to turbidity to achieve approximately 5 x 10⁸ CFU of final inoculum density and dispensed into sterile clear polystyrene 384-well microplates (Thermo Scientific 265202) with a final screening volume of 30 μL. DMSO solutions of test compounds and antibiotic controls were prepared as 1:1 dilution series and pinned into each assay plate (200 μL) using a high-throughput pinning robot (Tecan Freedom EVO 100) to achieve final screening concentrations ranging from 128 μM to 3.91 nM per compound. In each 384-well plate, lane 1 was reserved for DMSO vehicle, culture medium; lane 2 reserved for DMSO vehicle, culture medium, and target bacteria; lanes 23 and 24 reserved for antibiotic controls, DMSO vehicle, culture medium, and target bacteria. After compound pinning, assay plates were read as T₀ at OD₆₀₀ using an automated plate reader system (Thermo Scientific Spinnaker Microplate Robot; BioTek Synergy Neo2 plate reader) and then every hour for 20 hours (T₁−T₃₀), while incubating in an ambient room temperature carousel (25 °C). Resulting growth curves from the dilution series of each compound and control were used to determine their minimum inhibitory concentration (MIC) values following standard procedures.

Safety Statement. Caution! Biological agents Staphylococcus aureus, methicillin-resistant Staphylococcus aureus (MRSA), Vibrio cholerae, Salmonella enterica, Pythium, Pseudomonas aeruginosa, and Versinia pseudotubercolosis were handled following BSL-2 protocols. No other unexpected or unusually high safety hazards were encountered in the work reported.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acscentsci.1c01108.

Structure elucidation descriptions and NMR data for compounds 3 and 6, chemical characterization data for compounds 1–6, experimental methods and results for antibacterial screening, and high-resolution version of the full NP Analyst network presented in Figure S1 (PDF).

AUTHOR INFORMATION

Corresponding Author

Roger G. Linington — Department of Chemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada; orcid.org/0000-0003-1818-4971; Email: rliningt@sfu.ca

Authors

Sanghoon Lee — Department of Chemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

Jeffrey A. van Santen — Department of Chemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

Nima Farzaneh — Department of Chemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

Dennis Y. Liu — Department of Chemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

Cameron R. Pye — Unnatural Products Inc., Santa Cruz, California 95060, United States

Tim U. H. Baumeister — Department of Chemistry, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada

Weng Ruh Wong — Department of Chemistry and Biochemistry, University of California, Santa Cruz, Santa Cruz, California 95064, United States; Present Address: Microchemistry, Proteomics & Lipidomics Department, Genentech Inc., 470 E Grand Ave, South San Francisco, CA 94080

Complete contact information is available at: https://pubs.acs.org/10.1021/acscentsci.1c01108

Author Contributions

All authors have given approval to the final version of the manuscript.

Funding

This work was supported by NIH U41-AT008718 and NSERC Discovery.

Notes

The authors declare the following competing financial interest(s): JAvS is a consultant for Unnatural Products (a peptide-based biotech startup). CRP is the CTO of Unnatural Products (listed in affiliations).

REFERENCES

(1) Wani, M. C.; Taylor, H. L.; Wall, M. E.; Coggan, P.; McPhail, A. T. Plant Antitumor Agents. VI. Isolation and Structure of Taxol, a Novel Antileukemic and Antitumor Agent from Taxus Brevifolia. J. Am. Chem. Soc. 1971, 93, 2325–2327.

(2) Sehgal, S. N.; Baker, H.; Vézina, C. Rapamycin (Ay-22,989), a New Antifungal Antibiotic. II. Fermentation, Isolation and Characterization. J. Antibiot. (Tokyo) 1975, 28, 727–732.

(3) Rüegger, A.; Kuhn, M.; Lichti, H.; Loosli, H.-R.; Huguenin, R.; Quíquez, C.; von Wartburg, A. Cyclosporin A, a Peptide Metabolite from Trichoderma Polysporum (LINK Ex PERS.) Rifai, with a Remarkable Immunosuppressive Activity. Helvetica 1976, 59, 1075–1092.

(4) Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Lintoning, R. G. Retrospective Analysis of Natural Products Provides Insights for Future Discovery Trends. Proc. Natl. Acad. Sci. U. S. A. 2017, 114, 5601–5606.

(5) Louwen, J. J. R.; van der Hooft, J. J. J. Comprehensive Large-Scale Integrative Analysis of Omics Data to Accelerate Specialized Metabolite Discovery. mSystems 2021, 6, e000726–21.

(6) Schorn, M. A.; Verhoeven, S.; Riddler, L.; Huber, F.; Acharya, D. D.; Aksenov, A. A.; Aleti, G.; Moghaddam, J. A.; Aron, A. T.; Aziz, S.; et al. A Community Resource for Paired Genomic and Metabolomic Data Mining. Nat. Chem. Biol. 2021, 17, 363–368.

(7) Wolfender, J.-L.; Litaudon, M.; Touboul, D.; Queiroz, E. F. Innovative Omics-Based Approaches for Prioritisation and Targeted Isolation of Natural Products — New Strategies for Drug Discovery. Nat. Prod. Rep. 2019, 36, 855–868.

(8) Nothias, L. F.; Nothias-Esposito, M.; da Silva, R.; Wang, M.; Protsyuk, I.; Zhang, Z.; Sarvepalli, A.; Leyssen, P.; Touboul, D.; Costa, J.; et al. Bioactivity-Based Molecular Networking for the
Discovery of Drug Leads in Natural Product Bioassay-Guided Fractionation. J. Nat. Prod. 2018, 81, 758 − 767.

(9) Ory, L.; Nazih, E. H.; Daoud, S.; Mocquard, J.; Bourjot, M.; Margueritte, L.; Deluc, M. A.; Bard, J. M.; Pouchus, Y. F.; Bertrand, S.; et al. Targeting Bioactive Compounds in Natural Extracts − Development of a Comprehensive Workflow Combining Chemical and Biological Data. Anal. Chim. Acta 2019, 1070, 29 − 42.

(10) Olibon, F.; Allard, P. M.; Koval, A.; Righi, D.; Genta-Jouve, G.; Neyts, J.; Apé, C.; Pannecoque, C.; Nothias, L. F.; Cachet, X.; et al. Bioactive Natural Products Prioritization Using Massive Multi-Informational Molecular Networks. ACS Chem. Biol. 2017, 12, 2644 − 2651.

(11) Kunita, K. L.; Glassey, E.; Lintoning, R. G. Integration of High-Content Screening and Untargeted Metabolomics for Comprehensive Functional Annotation of Natural Product Libraries. Proc. Natl. Acad. Sci. U. S. A. 2015, 112, 11999 − 12004.

(12) Martens, L.; Chambers, M.; Sturm, M.; Kessner, D.; Levander, F.; Shoefstahl, J.; Tang, W. H.; Römpp, A.; Neumann, S.; Pizarro, A. D.; et al. MzML - A Community Standard for Mass Spectrometry Data. Mol. Cell. Proteomics 2011, 10, R110.000133.

(13) Plaskal, T.; Casillo, S.; Villar-Briones, A.; Orešić, M. MZmine 2: Modular Framework for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular Profile Data. BMC Bioinformatics 2010, 11, 395.

(14) Wang, M.; Carver, J. J.; Phelan, V. V.; Sanchez, L. M.; Garg, N.; Peng, Y.; Nguyen, D. D.; Watrous, J.; Kapono, C. A.; Luzzatto-Knaan, T.; et al. Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking. Nat. Biotechnol. 2016, 34, 828 − 837.

(15) Schulze, C. J.; Bray, W. M.; Woerhmann, M. H.; Stuart, J.; Lokey, R. S.; Lintoning, R. G. Function-First” Lead Discovery: Mode of Action Profiling of Natural Product Libraries Using Image-Based Screening. Chem. Biol. 2013, 20, 285 − 295.

(16) Guttman, A. R-Trees: A Dynamic Index Structure for Spatial Searching. SIGMOD Rec. 1984, 14, 47 − 57.

(17) Wong, W. R.; Oliver, A. G.; Lintoning, R. G. Development of Antibiotic Activity Profile Screening for the Classification and Discovery of Natural Product Antibiotics. Chem. Biol. 2012, 19, 1483 − 1495.

(18) van Santen, J. A.; Jacob, G.; Singh, A. L.; Aniebok, V.; Balunas, M. J.; Bunsko, D.; Neto, F. C.; Castaño-Espriu, L.; Chang, C.; Clark, T. N.; et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Cent. Sci. 2019, 5, 1824 − 1833.

(19) Skellam, E. J.; Stewart, A. K.; Strongman, W. K.; Wright, J. L. C. Identification of Micromonolactam, a New Polyene Macrolactam from Two Marine Micromonospora Strains Using Chemical and Molecular Methods: Clarification of the Biosynthetic Pathway from a Glutamate Starter Unit. J. Antibiot. (Tokyo) 2013, 66, 431 − 441.

(20) Hoshino, S.; Okada, M.; Awakawa, T.; Asamizu, S.; Onaka, H.; Abe, I. Myclic Acid Containing Bacterium Stimulates Tandem Cyclization of Polyene Macrolactam in a Lake Sediment Derived Rare Actinomycete. Org. Lett. 2017, 19, 4992 − 4995.

(21) Gomi, S.; Amano, S.; Sato, E.; Miyadōh, S.; Kodama, Y. Novel Antibiotics SF2738A, B and C and Their Analogs Produced by Streptomyces Sp. J. Antibiot. (Tokyo) 1994, 47, 1385 − 1394.

(22) Shindo, K.; Yamagishi, Y.; Okada, Y.; Kawai, H. Collismycins A and B, Novel Non-Steroidal Inhibitors of Dexamethasone-Glucocorticoid Receptor Binding. J. Antibiot. (Tokyo) 1994, 47, 1072 − 1074.

(23) Seyedsayamdomst, M. R.; Traxler, M. F.; Zheng, S. L.; Kolter, R.; Clardy, J. Structure and Biosynthesis of Amychelin, an Unusual Mixed-Ligand Siderophore from Amycolatopsis Sp. A44. J. Am. Chem. Soc. 2011, 133, 11434 − 11437.

(24) Fukuda, T. T. H.; Helfrich, E. J. N.; Mevers, E.; Melo, W. G. P.; van Arnam, E. B.; Andes, D. R.; Currie, C. R.; Pupo, M. T.; Clardy, J. Specialized Metabolites Reveal Evolutionary History and Geographic Dispersion of a Multilateral Symbiosis. ACS Cent. Sci. 2021, 7, 292 − 299.