The Molecules Gateway : A Homogeneous, Searchable Database of 150k Annotated Molecules from Actinomycetes

Journal of Natural Products

Simone, Matteo; Iorio, Marianna; Monciardini, Paolo; Santini, Massimo; Cantù, Niccolò et al

https://doi.org/10.1021/acs.jnatprod.4c00857

JOURNAL OF NATURAL PRODUCTS

pubs.acs.org/jnp

Article

The Molecules Gateway: A Homogeneous, Searchable Database of 150k Annotated Molecules from Actinomycetes

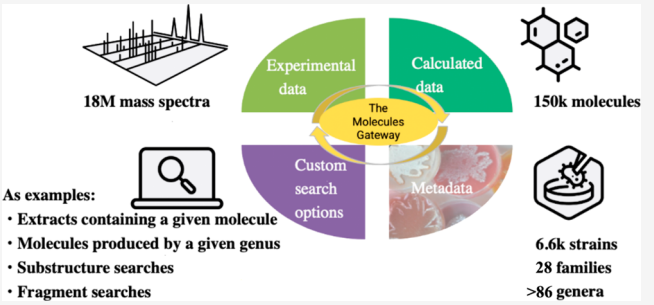
Matteo Simone,[¶] Marianna Iorio,[¶] Paolo Monciardini,[¶] Massimo Santini, Niccolò Cantù, Arianna Tocchetti, Stefania Serina, Cristina Brunati, Thomas Vernay, Andrea Gentile, Mattia Aracne, Marco Cozzi, Justin J. J. van der Hooft, Margherita Sosio, Stefano Donadio, and Sonia I. Maffioli*

Cite This: *J. Nat. Prod.* 2024, 87, 2615−2628

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Natural products are a sustainable resource for drug discovery, but their identification in complex mixtures remains a daunting task. We present an automated pipeline that compares, harmonizes and ranks the annotations of LC-HRMS data by different tools. When applied to 7,400 extracts derived from 6,566 strains belonging to 86 actinomycete genera, it yielded 150,000 molecules after processing over 50 million MS features. The web-based Molecules Gateway provides a highly interactive access to experimental and calculated data for these molecules, along with the metadata related to extracts and producer strains. We show how the Molecules Gateway can be used to rapidly identify known hard to find microbial products, unreported analogs of known families and not yet described metabolites. The Molecules Gateway, which complements available repositories, contains annotated MS data, both acquired and computationally processed under an identical workflow, making it suitable for global analyses which reveal a large and untapped chemical diversity afforded by actinomycetes.



Natural products (NPs), molecules produced by living organisms, continue to provide a longstanding source of diverse drugs and drug leads.[1] The entire chemical complexity of an organism can be captured by processing the entire organism, or a single part (leaves, roots etc.), with appropriate methods to prepare extracts. The number of molecules present in a extract may vary from a few (e.g., organic extraction of a tissue or organ) to several hundred molecules (e.g., full extracts of microbial cultures).[2,3] In all cases, molecules are usually of unknown nature and concentration. While properly prepared extracts are suitable for most drug discovery programs, the identification of the molecule(s) responsible for the observed biological signal requires the time-consuming processes of deconvolution, i.e., reducing the number of molecules present in the sample to be tested, and dereplication, i.e., matching the observed properties of the active molecule(s) to those reported in available databases.[4] Hence, knowing *a priori* the composition of extracts would greatly accelerate the drug discovery process.

Untargeted analysis of NP extracts typically involves separation with liquid chromatography (LC) coupled with mass spectrometry (MS), which provides metabolic features, characterized by mass-to-charge (m/z) ratios measured in MS1 experiments, along with m/z values of molecular fragments (MS2 fragmentation)[5,6] with a single metabolite typically detected as multiple metabolic features with the same retention time (t_R) but different m/z values.[7,8] Structural annotation of small molecules is a bottleneck: a metabolite's candidate molecular structure is usually determined by comparing its mass spectral features with mass spectral databases such as GNPS[9] or mzCloud.[10] It is worth to note that care is needed when matching and/or comparing spectra recorded with different acquisition parameters (e.g., resolution, collision energy, ionization), while useful information, such as relative t_Rs, is not available from these resources. Furthermore, computational metabolomics workflows exist that take candidate structures from structure databases: however, the structure collections used will impact the relevance and plausibility of structure annotations by these workflows. Manual annotation of mass spectra represents a daunting task, as it is time-consuming and dependent on expert knowledge, thus hampering scalability and high-throughput.[11] Subjectivity and human error may also introduce biases from researchers lacking expertise in particular classes of NPs. Addressing these limitations necessitates the use of standardized laboratory procedures, the improvement and synergistic combination of

Received: July 25, 2024
Revised: October 18, 2024
Accepted: October 18, 2024
Published: October 25, 2024

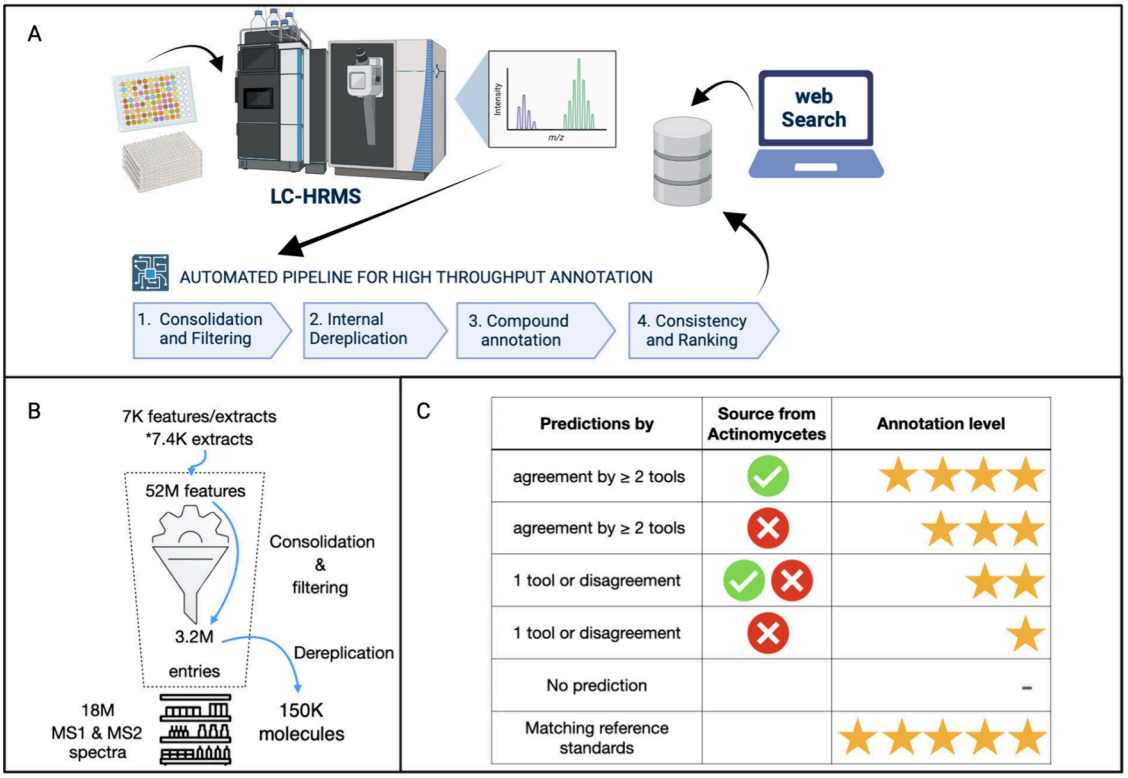


Figure 1. Panel A: Schematic of the process for analyzing microbial fermentation extracts and creating a web-searchable database, along with the logical flow of the annotation pipeline for processing HR-MS data. Panel B: summarized inputs and outputs after analyzing 7,400 actinomycete extracts. Entries denote unique molecule-extract combinations. Panel C: organization of annotation results into four different levels. See also Table 2 for further details.

Table 1. Performance of the Annotation Tools on Validated Molecules from Reference Strains[a]

			Compound Discoverer			MolDiscovery			MS2Query			
Reference strain	Code	number of manually curated molecules	correct	wrong	na	correct	wrong	na	correct	wrong	na	Correct for at least 1 tool
Streptomyces sp. ID38640	ID38640	15	5	4	6	8	0	7	7	8	0	13
Streptomyces rimosus	ATCC10970	15	7	1	7	9	4	2	3	12	0	11
Streptomyces mobaraensis	ATCC15003	11	3	5	3	6	2	3	0	11	0	7
Streptomyces coelicolor	M145	4	3	1	0	4	0	0	0	4	0	4
Streptomyces venezuelae	ATCC15439	3	1	0	2	1	0	2	2	1	0	3
Streptomyces griseus	DSM40236	1	0	1	0	0	1	0	0	1	0	0
Streptomyces glaucescens	GLA000	6	6	0	0	6	0	0	4	2	0	6
Streptomyces avermitilis	NRRL 8165	4	1	0	3	1	0	3	1	3	0	2
	TOT	**59**	**26**	**12**	**21**	**35**	**7**	**17**	**17**	**42**	**0**	**46**
	%		**44**	**20**	**36**	**59**	**12**	**29**	**29**	**71**	**0**	**78**

[a]na: annotation not available.

automated annotation tools and appropriate mass spectral reference databases.

The demand for high-throughput methods has led to the development of various tools to automatically annotate molecules present in complex mixtures. Here, we developed a comprehensive workflow, which integrates leading compound discovery tools (Compound Discoverer,[12] MolDiscovery[13] and MS2Query[14]), evaluates and ranks their prediction and integrates these with additional computational tools, biological criteria and metadata. We applied this pipeline to 7,400 extracts derived from 6,566 actinomycetes, leading to the annotation of 150,000 molecules, loaded in a web-based database called "Molecules Gateway". The database, which can be employed for different searches, highlights the existence of a large, untapped chemical diversity from a relatively small number of actinomycetes.

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628

RESULTS AND DISCUSSION

Design of the Automated Pipeline. We designed an automated annotation pipeline with the objective of performing systematic annotation of a large number of microbial fermentation extracts. After LC-HRMS analysis, the pipeline processes data through four sequential steps, as depicted in Figure 1A: 1) consolidation and filtering, whereby different m/z values originating from the same molecule are grouped into a single entity, followed by filtering out signals below fixed MS peak thresholds; 2) internal dereplication, whereby molecules identical with compounds observed in previously analyzed extracts do not enter the following steps, but are assigned the annotations of their matching molecules while the relevant data and metadata are added; 3) compound annotation, which exploits the annotation capabilities of Compound Discoverer, MS2Query and MolDiscovery, with their outputs harmonized and combined into a single output file; and 4) consistency and ranking, which weighs in the molecular formula (MF) predicted by SIRIUS[15] and the t_R calculated by jp^{2}rt tool[16] and the biological consistency of the annotation (i.e., whether the molecule has been reported from the same biological source, actinomycetes in our case). At the end, experimental and calculated data, along with the associated metadata relative to the producer strains and extracts, enter a web-based, searchable database, the "Molecules Gateway", as described below.

Concept Validation. We evaluated the performance of Compound Discoverer, MolDiscovery and MS2Query in the annotation of a selected number of metabolites present in extracts derived from eight well-characterized *Streptomyces* strains: *S. coelicolor* and *S. avermitilis*, reference strains with extensive metabolite analyses;[17] *Streptomyces* sp. ID 38640,[18] *S. rimosus*, *S. mobaraensis*, *S. venezuelae*, *S. griseus* and *S. glaucescens* (see Table 1).

Consolidation and dereplication were performed by Compound Discoverer, and for this validation Compound Discoverer was limited to utilizing only the default databases (Chemspider, NPAtlas and mzCloud library). Signals corresponding to known molecules from the eight reference strains were manually searched for within the three software outputs thus selecting MS1 m/z values and MS2 fragmentation scans related to the target molecules. 59 molecules were manually annotated (see details in Supporting Information Table S1) based on a reference standard (four molecules) or by manually checking the consistency of the experimental data in terms of MS1 and MS2 fragmentation data, type of adducts and UV absorption, together with the presence of the expected biosynthetic gene cluster in the genome of the producing strain. For each software the correct and wrong predictions for each strain are reported together with the number of absence of prediction. Note that, differently from Compound Discoverer and MolDiscovery, MS2Query is also able to detect analogs, as it evaluates MS2 fragmentation irrespective of a match between the experimental exact mass and the exact mass of the molecule in its library. During the concept validation MS2Query was run in its original configuration that includes both exact matching and analogue searching. Combined with the lenient thresholds used, this explains why MS2Query always produce an annotation.

The annotation was considered correct when it identified the exact molecule, an isomeric congener or, in the case of MS2Query, a highly similar molecule (i.e., Desferrioxamine G, entry 51 of Table S1, was correctly assigned by both Compound Discovery and MolDiscovery and assigned as the isomeric congener Deferrioxamine E by MS2Query). The details for each molecule are listed in Table S1. In the last column the number of correct annotations from at least one tool is reported for each reference strain. Molecules present in more than one strain are here considered only once but detailed in Table S2.

Manual curation of the LC-HRMS data from the corresponding eight extracts led to the identification of 51 metabolites which were previously reported in the literature for these strains. In addition, we identified eight molecules derived from the unfermented medium (see below). Overall, 22 molecules appeared in multiple extracts. When the HRMS raw data (on average, 7,000 features per extract) were processed through the first two steps of the annotation pipeline, different adducts originating from the same molecule (a total of 196 adducts from the 59 molecules) were correctly merged (Table S1) and the 22 molecules appearing in multiple extracts were correctly dereplicated (Table S2). Of note, the 59 molecules were always detected, at least in their most abundant m/z values, above a 10E7 peak area and generally showed a peak quality around 8 (a Compound Discoverer parameter indicating the peak quality). Next, we tested the performance of Compound Discoverer, MolDiscovery and MS2Query in annotating the 59 molecules. MolDiscovery exhibited the highest accuracy at 59%, followed by Compound Discoverer and MS2Query. Of note, each tool led to a significant number of false positives. The annotation outputs are reported in Table S1 and summarized in Table 1. While no single tool was able to accurately identify all the 59 molecules, 46 of them (78%) were successfully identified by at least one tool, suggesting that an acceptable accuracy level could be achieved by a judicious choice of the most likely annotation.

Implementation of the Automated Annotation Pipeline. The annotation pipeline was implemented as a series of consecutive rounds of analysis and data processing, with each round consisting of 80 different extracts (i.e., the content of one 96-well microtiter plate suitable for bioassay screening). The input for each round of annotation was the raw HRMS data files, which underwent signal consolidation and filtering. By applying peak area and quality thresholds (see above) and by considering only signals with MS2 fragmentation, we observed a 10-fold reduction in the number of molecules entering step 2 (Figure 1A), with a lower risk of including MS artifacts. For dereplication, MS2 fragmentation data from newly analyzed extracts were compared with those present in the Dereplication Library, which initially included data from 452 reference molecules (Table S3) and was augmented with all newly observed molecules at the end of each 80-extract round (see Experimental Section). Molecules not matching entries in the Dereplication Library entered step 3 of the annotation pipeline, which involved parallel computational analysis utilizing Compound Discoverer, MolDiscovery, and MS2Query (see Experimental Section for details). To manage the different output formats of the three tools, and compare and score their results into an automatic workflow, the Soupy software was developed (Figure S1 and S2 and Tables S4−S5). In principle, each tool can annotate a molecule or not, thus from zero to three different structural predictions can be present. Instances without any prediction were labeled "unknown", while molecules with one to three predictions entered the phase of "Consistency and Ranking", which ultimately selected the most likely annotation. The criteria for consistency took into consideration the biological source, the SIRIUS-calculated MF and the jp^{2}rt-calculated t_R of the annotated molecules (Table S6). Following a decision tree (Figure S3), annotations were grouped into 12

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628

Table 2. Results from Manual Curation of 100 Molecules: 20 Randomly Selected for Each Grouped Annotation Bins as Defined in Table S6 and Figure S3 (Except for 5a/b, Undecided) and Their Corresponding Annotation Levels[a]

			Manual curation result			
Group	Annotation bins	Source from Actinomycetes	Likely	Insufficient	Unlikely	Annotation level (stars)
A	1a: ≥2 tools agreement (InChIKey)	yes	17 (85%)	-	3 (15%)	4
C	2a: ≥2 tools agreement (similarity name)	yes	15 (75%)	2 (10%)	3 (15%)	4
B	1b: ≥2 tools agreement (InChIKey)	no	5 (25%)	2 (10%)	13 (65%)	3
D	2b: ≥2 tools agreement (similarity name)	no	8 (40%)	-	12 (60%)	3
E	3a−c and 4a−c: 1 tool or disagreement	both	3 (15%)	-	17 (85%)	2

[a]Details are reported in Table S7. Each tool can annotate a molecule or not, thus from zero to three different structural predictions can be present. Agreement results when at least two tools propose the same chemical structure (same InChIKey, as in Groups A and B) or chemically related molecules (name similarity higher than 75% as described in Experimental Section, as in Groups C and D). Group E contains molecules annotated by a single tool together with molecules for which the annotations proposed by two or more tools are in disagreement. Groups A and C are further discriminated from Groups B and D depending whether the proposed molecule is known to be produced by an Actinomycete or not, while biological source was not considered for group E.

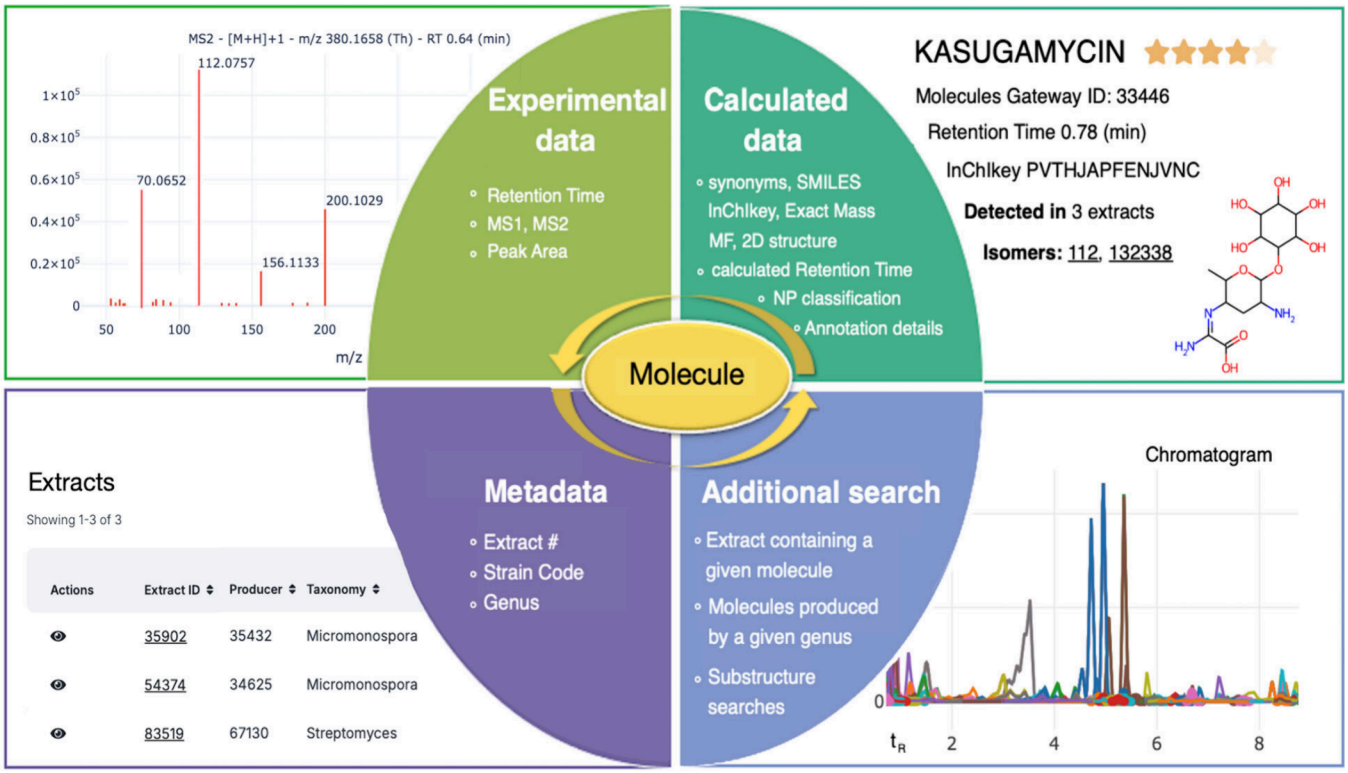


Figure 2. Schematic representation of the Molecules Gateway. All information linked to each molecule is interconnected. It includes experimental data (top left), calculated data (top right), metadata (bottom left) and additional search inputs and outputs (bottom right). A visual example is shown for each quadrant.

different bins (Table S6). In order to evaluate its scalability, we applied the annotation pipeline to extracts derived from 6,566 strains from our collection of about 45,000 actinomycetes.[19] These strains—classified by partial 16S rRNA gene sequence—represent at least 86 genera belonging to 28 distinct actinomycete families (Figure S4). Among the processed set, *Streptomyces* strains account for 2,000 isolates, and other important contributors are strains belonging to the families *Micromonosporaceae*, *Streptosporangiaceae*, *Thermomonosporaceae*, *Nocardiaceae* and *Pseudonocardiaceae*. Most strains were cultivated in a single fermentation medium, leading to 7,400 extracts prepared from the same number of cultures. Step 1 of the automated annotation pipeline reduced the complexity of the HR-MS data for each extract from about 7,000 features to about 400 grouped metabolite features (called molecules hereafter). Iterative dereplication of each 80-extract set against the Dereplication Library ultimately led to 150,777 distinct molecules representing 3,197,300 unique molecule-extract pairs. The average extract thus contributed 432 molecules, of which 20 unique. Overall, we processed over 52 million distinct MS features leading to a database of 18 million MS1 and MS2 spectra (Figure 1B). To assess the reliability of the annotations, we manually curated 100 randomly selected molecules, specifically 20 for each set of grouped annotation bins (Table S6). Manual curation involved, besides checking the consistency of the MS2 fragmentation with the proposed structure, using additional data (MS adducts and UV consistency, presence of congeners in the same extract, Molecular networking; see Experimental Section) that were not considered during automated annotation. Apart from four instances with

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628

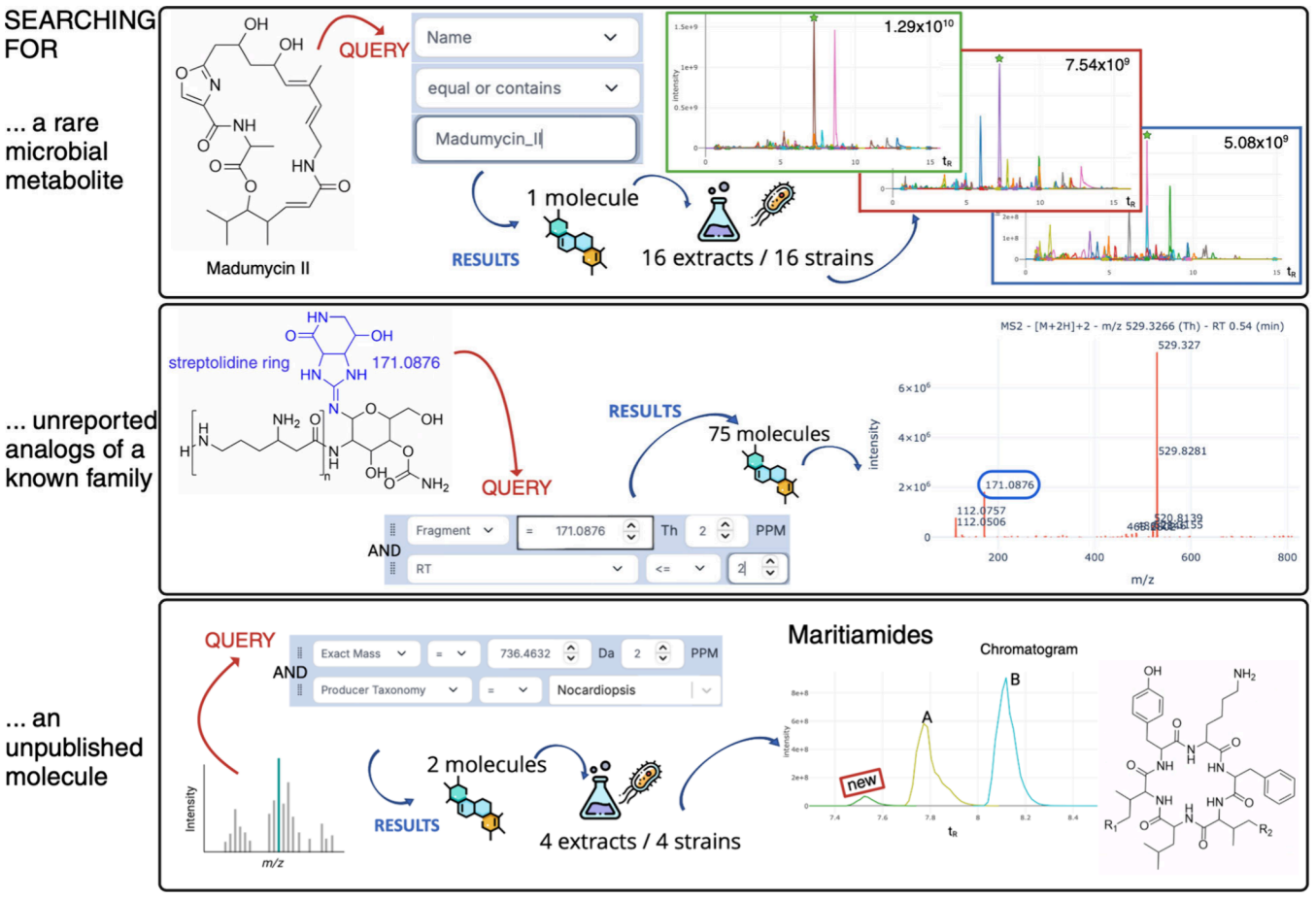


Figure 3. Possible uses of the Molecules Gateway. The figure shows the question asked, the search logics and an example of the outputs. Details for each example can be found as Supporting Information as videos.

insufficient data to assist manual curation, we could categorize the automated annotations of the 96 molecules as likely or unlikely, as summarized in Table 2 (details in Table S7).

As expected, reliability was highest (75−85%) when two annotation tools agreed and the annotation was consistent with the biological source. Consistency with biological source was a more reliable indicator of annotation accuracy also when a single tool had annotated a molecule. On the basis of logical considerations and of the results from manual curation, the annotations were grouped into four levels that qualitatively provide an indication of their reliability (Figure 1C). When at least two tools agreed, annotations were scored with 4 stars when the biological source was coherent, and with 3 stars when not. When tools were in disagreement, or there was a prediction by a single tool, the annotation received two stars when coherent with the biological source or when at least two consistency were present, and one star when not. One-star annotation molecules were actually labeled as "undecided", given the limited reliability of this category. Of note, among the 100 molecules, we observed three cases when possible duplicate molecules escaped the dereplication process (Table S7). If this figure translates to a larger scale, a failure rate of 3% in dereplication would be quite acceptable.

The Molecules Gateway. The results from the automated annotation pipeline entered the Molecules Gateway, centered around the annotated molecules. Each molecule, identified by a specific ID, is accompanied by experimental data (detected adducts with their MS1 and MS2 data, associated area for each occurrence in different extracts, and t_R), the associated metadata (a list of the extracts where it was identified, along with the code and genus of the extract-generating strain(s)), and, depending on the annotation level a certain number of calculated data (Figure 2).

The latter can include the proposed annotation, MF and chemical structure, InChIKeys, SMILES, name and synonyms, and NP classification for molecules classified at annotation levels 2 to 5. In addition, it includes the annotation calls made by each tool, along with the consistency with the biological source, with the predicted versus experimental t_R, and with the annotation-derived versus the SIRIUS-calculated MF. Molecules labeled as "undecided" contain all the information as above except for the proposed annotation, while molecules labeled as "unknown" include solely the m/z value, the experimental t_R and, for masses below 850 Da, the SIRIUS-calculated MF. In addition, the Molecules Gateway provides a visual representation of the chromatogram for each extract, along with a list of all annotated molecules detected in each extract. This information can be useful for assessing the reliability of the annotation (e.g., by observing congeners of the annotated molecule) or for selecting the best strain for further work. Most experimental and calculated data are searchable in the Molecules Gateway, including substructure searches and MS2 fragment searches. Searches with experimental data span both known and unknown molecules, while searches with calculated data are limited to fully annotated molecules, with the exception of MFs, which can be performed for any molecule with exact mass below 850 Da. In

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628

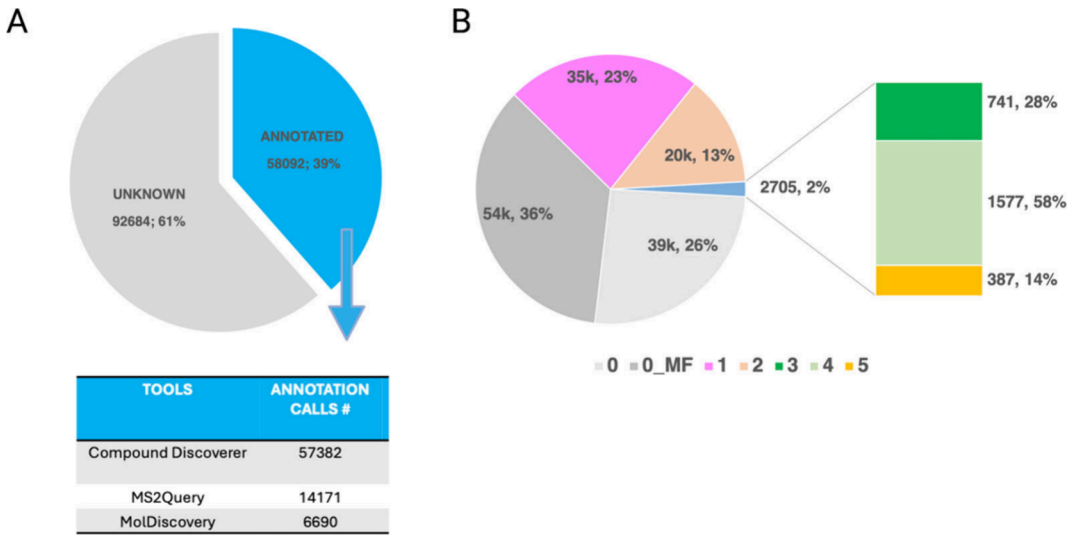


Figure 4. Overview of the 150,777 molecules listed in the Molecules Gateway. Panel A: Actual number of predictions by the different annotation tools. Panel B: Distribution of entries by annotation level.

addition, the Molecules Gateway flags whether a molecule matched any of the 1,031 distinct molecules detected in unfermented media (see Experimental Section).

Using the Molecules Gateway. With its web-based, multiple search options, the Molecules Gateway can be used in many ways, ultimately leading to the identification of desired molecule(s), of the associated extract(s) and of the producing strain(s). We report below selected examples of using the Molecules Gateway, as schematized in Figure 3 and detailed in the Supporting Information.

A first example involves searching for a rare microbial product, i.e., one with few commercial suppliers and few reports in the literature. Madumycin II, the simplest type A streptogramin, originally isolated from an *Actinoplanes*[20] and a peptidyl transferase center inhibitor,[21] can fit into this category. Searching for "madumycin II" yields one hit (ID 56, with the highest reliability level) that appears in 16 distinct extracts, all from *Actinoplanes* strains. Selecting a specific extract leads to a web page with the corresponding chromatogram, displaying all detected molecules. This enables assessing the relative abundance of madumycin II, the presence of related molecules and the overall complexity of the metabolomes across the different strains, aiding in the selection of the most suitable strain for further work. A second example pertains to searching for a yet undescribed analogs of a known metabolite family. As a specific case we took streptothricin, a molecule described in 1942[22] and the subject of numerous studies.[23] Hallmarks of streptothricins are the streptolidine ring, which results in a diagnostic MS^2 fragment of m/z 171.0876, and a high hydrophilicity, leading to low t_Rs in reversed phase chromatography. Searching the Molecules Gateway for "Fragment = 171.0876" and "$t_R \leq 2$ min" led to 75 hits, mostly from *Streptomyces* strains. Many of these hits contain m/z 171.0876 as a major MS2 fragment, as streptothricins do. In addition to several known members of the streptothricin family, many hits are labeled as "unknown", suggesting they represent variants not yet reported in the scientific literature, even for a frequently occurring molecule known for many decades. A third example consists in looking for a specific molecule not yet described in the scientific literature, e.g., a molecule under investigation in a scientist's laboratory. To illustrate this scenario, we searched for maritiamides A and B,[24] antibacterial cyclic hexapeptides containing a Val and Ile residue, respectively, isolated from a *Nocardiopsis maritima* strain and reported in the scientific literature in 2024, thus absent from the libraries used in the present work. Searching for exact masses of 763.4632 and 777.4789 Da and restricting the search to the genus *Nocardiopsis* led to two hits (ID 69072 "undecided" and ID 69653 "unknown"), present in four distinct extracts, two from *Nocardiopsis* strains and two from *Streptomyces* strains. The fragmentation patterns of these hits are consistent with the maritiamides A and B structures, respectively. In addition, all extracts contain a third molecule (ID 69724 "unknown") whose exact mass and fragmentation pattern are consistent with an additional maritiamide congener in which both Ile residues in maritiamide B are replaced by Val (Figure S5). A noteworthy characteristic of the complex in that it consists of comparable amounts of maritiamides A and B in the *Nocardiopsis* extracts, while maritiamide A is the predominant form in the *Streptomyces* extracts (Figure S6). The above examples illustrate how the Molecules Gateway can be readily queried to identify desired metabolites, analogs or unknown molecules, leading in many cases to the identification of multiple producer strains, also from different actinomycete genera. Having additional producing strains may help, for example, in the identification of the biosynthetic gene cluster, in detecting the presence or absence of analogs, in finding better producer strains, or in increasing the chances of finding a strain amenable to genetic manipulation.

Exploring the Chemical Diversity from 6,566 Actinomycetes. The Molecules Gateway represents a data set of 150,777 molecules annotated by the same workflow, produced by a diversified set of actinomycetes and present in homogeneously prepared extracts, which were analyzed under identical conditions. Thus, the data set is amenable to different analyses (as exemplified in Figures 4−7) that can provide insights into the chemical diversity of actinomycetes, including the existence of genus-specific and frequently occurring molecules. In terms of annotation statistics, Compound Discoverer, MolDiscovery and MS2Query predicted molecules at very different rates due to their different annotation logics and the size and pertinence of the employed databases (Figure 4).

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628

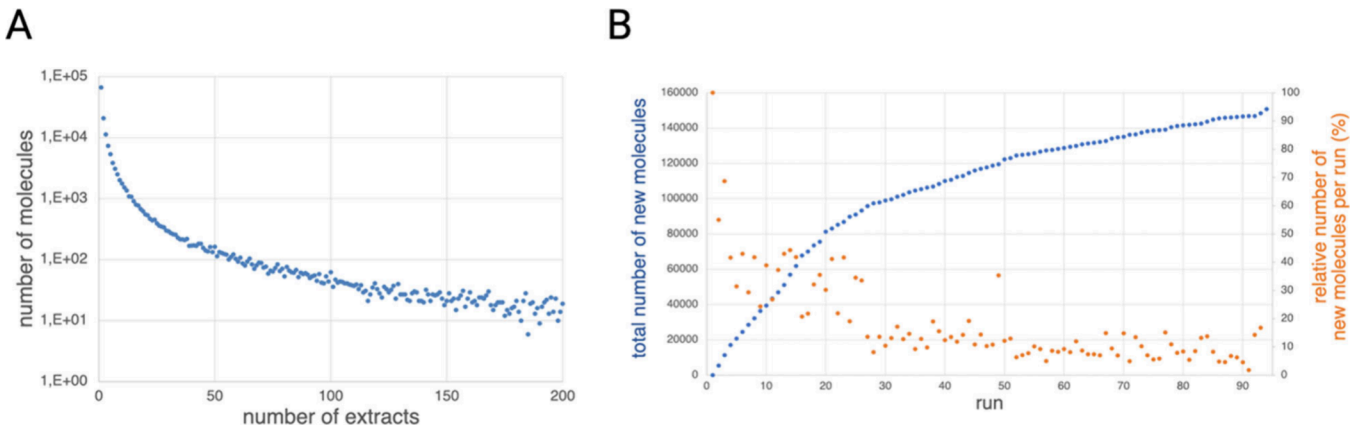


Figure 5. Panel A: Number of molecules found in a given number of extracts. Molecules present in >200 extracts are omitted. Panel B: Incremental growth of the Molecules Gateway.

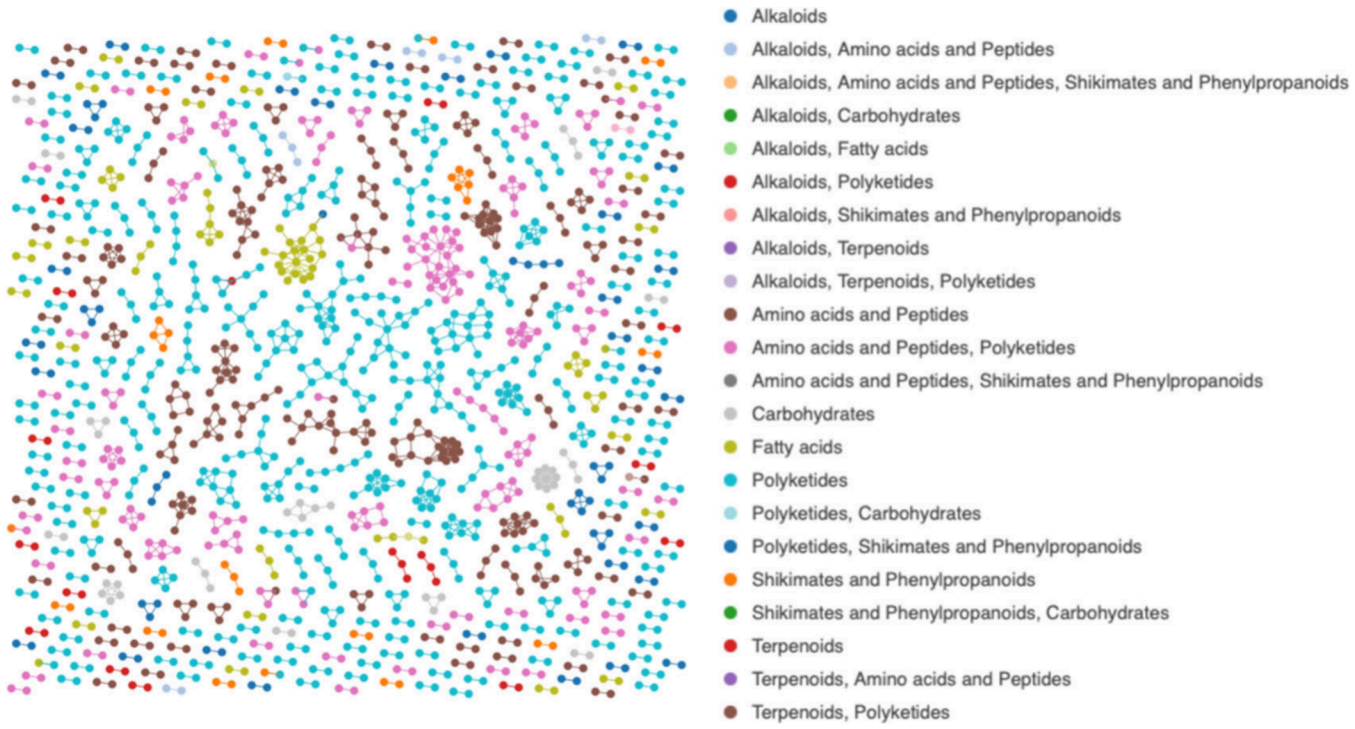


Figure 6. Chemical relatedness and originating biosynthetic class: each dot indicates a molecule, and each color shows the originating biosynthetic class according to NPclassifier.[25] Clustering was performed using Morgan Fingerprint[26] (with Radius set at 5) and Tanimoto Similarity[27] (with cosine set at 0.7). In the figure the 1,417 molecules arranged in families are represented while the remaining 4,243 molecules as single nodes are represented in Figure S8. Interactive version of Figure 6 and Figure S8 are available as Supporting Information.

As expected from the large number of molecules, most molecules belong to the 0-star category (i.e., unknown), with about 60% of them having a SIRIUS-calculated MF (Figure 4B). In terms of molecules distribution, just 3,120 out of 150 K molecules in the Molecules Gateway were detected in more than 200 extracts and are likely to include, in addition to the 1,031 molecules present in the unfermented media, primary metabolites and frequently occurring specialized metabolites. The remaining molecules are distributed in the extracts in an exponentially decaying manner, as expected for specialized metabolites (Figure 5A).

Looking at how the Dereplication Library increased in size after each 80-extract analysis can provide an indication of how further the database can grow by processing additional strains. During the first 20 runs, the number of new molecules increased at a rate of about 4,000 molecules per run, and then stabilized after the 30th run to about 1,200 per run, with no sign of a decreasing rate after 92 runs (Figure 5B). The relative number of new molecules added to the database varied with different runs, reflecting the varying contribution by different actinomycete genera. These data suggest that additional molecules can be added to the database at a rate of 10−20 new molecules per strain, depending on the chosen genus.

The distribution of exact masses and t_Rs for the 58,093 molecules with annotation shows that there is no obvious bias in t_R and exact mass (Figure S7). In terms of chemical diversity, Figure 6 and Figure S8, report 1,417 molecules arranged in families and 4,243 molecules as single nodes, respectively, along with the corresponding NP classification. Overall, these data indicate that the annotation pipeline did not introduce any

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628

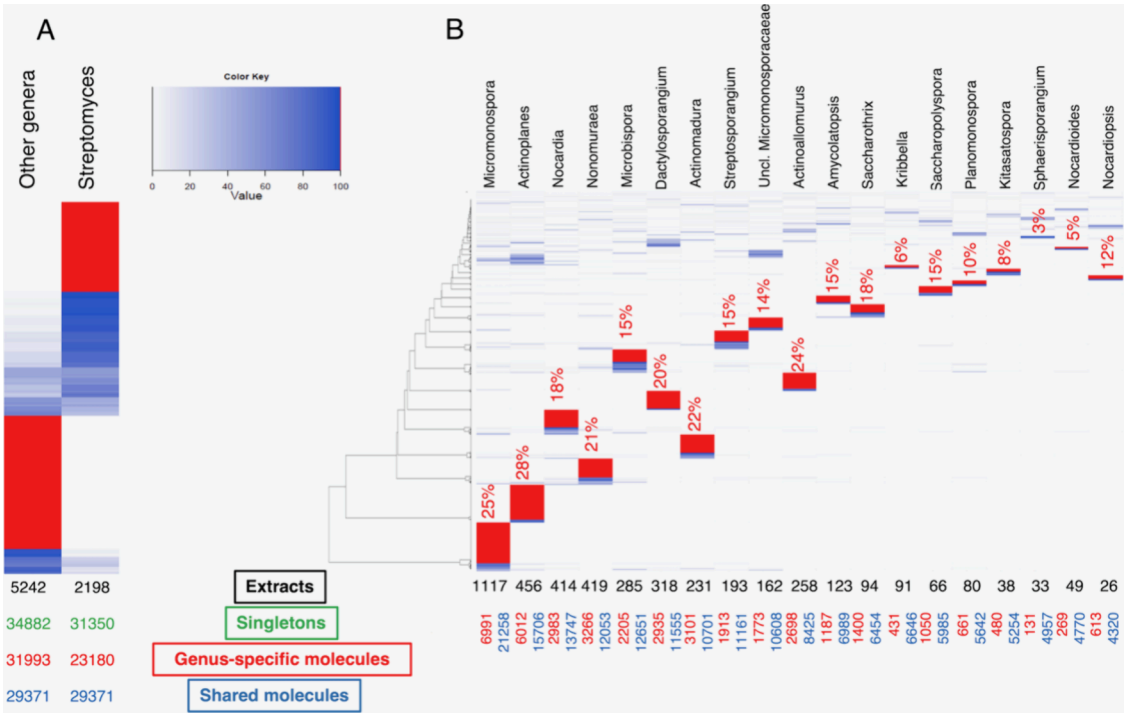


Figure 7. Heatmaps representing the distribution of molecules after UPGMA-based hierarchical clustering. Group-specific molecules (i.e., those found only in *Streptomyces* or only in genera different from *Streptomyces* for panel A, or in a single genus for panel B) are indicated by red bars, while gradient blue indicates increasing abundance for shared molecules (i.e., those found in more than one group). Panel A shows the distribution of molecules in *Streptomyces* and in the bulk of non-*Streptomyces* genera. The number of singletons (molecules found in a single extract) is reported below in green type. Panel B shows the distribution of genus-specific molecules among the non-*Streptomyces* genera contributing at least 5,000 molecules. Panel B also reports the number of extracts (black type), the genus-specific molecules as number and percentage (red type), and the number of shared molecules (blue type) for each group. For both panels, annotated molecules, detected in more than 2 and less than 500 extracts, were grouped according to genus and Ward-based hierarchical clustering in R (hclust). A distribution matrix, representing the specificity of each molecule to each genus, was used to create heatmaps in R-4.3.3 (R Core Team, 2023) using gplots and colorspace packages and a modified version of heatmap.3 function[28] available at https://github.com/TV27/heatmap-metabo/blob/main/revheatmap3-R.

obvious bias, with molecular families or biosynthetic classes greatly over- or underrepresented.

Finally, we wanted to establish the relative contribution of different genera to populating the Molecules Gateway. The number of contributed molecules varied greatly, from 199 molecules for *Embleya* (one single extract) to 83,901 molecules for *Streptomyces* (2,198 extracts), with a rough correlation between the number of extracts and the number of molecules (Figure 7).

We were also interested in metabolite distribution at genus level, which is the best taxonomic rank for comparative evaluation of biosynthetic potential.[29] Among the 83,901 molecules observed from *Streptomyces*, one-third (29,371) are shared with other genera, while the rest are *Streptomyces*-specific (Figure 7, Panel A). When a similar analysis was performed on the remaining 96,246 molecules from non-*Streptomyces* genera, almost 70% of them (66,875) were not encountered in *Streptomyces*. Of note, the number of singletons (i.e., molecules observed in a single extract) was similar in both groups (Figure 7, Panel A). When the non-*Streptomyces* genera were analyzed individually, each genus contributed genus-specific metabolites ranging from 3% for *Sphaerisporangium* to 28% for *Actinoplanes* (Figure 7, Panel B). These data indicate that most if not all genera produce genus-specific molecules, while pointing to the most prolific genera in terms of genus-specific molecules. In this respect, genus diversification was empirically employed decades ago to reduce rediscovery rates in bioassay-based screens.[30]

CONCLUSIONS

After preparing 7,400 extracts derived from 6,566 actinomycetes and processing 52 million MS data, the Molecules Gateway with its 150,777 molecules represents one of the largest resources for analyzing and searching microbial products. The homogeneity of wet-lab and computational methods used to create the Molecules Gateway make it an attractive resource to accelerate research projects. Other databases of experimental MS data exist along with community-curated annotations: they range from 16- to 45-thousand molecules and include between 0.1 and 2 million MS spectra.[31−33] For example, Metabolights, GNPS/MassIVE and microbeMASST are valuable resources for the scientific community but their data remain intrinsically heterogeneous notwithstanding recent efforts at harmonization.[34] In addition, being community-based, their content reflects the interest of the scientific community and its willingness to contribute data, leading to possible biases. We believe the Molecule Gateway can nicely complement these resources, while providing a single place to access the physical samples that originated the data, along with the producing microorganisms. Any MS-based annotation system suffers from the intrinsic limitation that stereoisomers and positional isomers cannot be distinguished, since they possess the same exact mass (MS1) and may display extremely similar or even undistinguishable mass fragmentations (MS2). Thus, in the absence of reference standards, stereoisomers and positional isomers will lead to identical annotations for molecules with (potentially) different t_Rs. This limitation is

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628

obviously present in the Molecules Gateway, notwithstanding our efforts at using reference standards. At the same time, the complexity of the automated annotation pipeline and the mere scale of the processed data, can lead to mis-annotations. For example, the utilized annotation tools have intrinsic limitations. Compound Discoverer operates on the largest database, which can be expanded but not tailor-reduced, resulting in many annotations that are not biologically relevant and unlikely to be correct. The accuracy of MS2Query, which operates on libraries of MS2 data from properly annotated molecules, is expected to increase when run against expanded and biologically relevant libraries. We observed that MolDiscovery is more reliable with some metabolite classes, e.g., unmodified peptides or oligosaccharides, that obey well-defined fragmentation rules. In addition, the annotation tools are known to suffer from an annotation bias that favors the known chemical space over the unknown.[35] Nonetheless, this annotation bias did not seem to have a profound effect on the annotations, as most molecules in the database belong to the unknown category (Figure 4B). Finally, we observed that Compound Discoverer occasionally failed to properly dereplicate molecules as the size of the Dereplication Library increased, leading to redundant entries in the Molecules Gateway. Our sampling of 100 molecules suggests this phenomenon is limited to a small percentage. Notwithstanding these limitations, we believe the Molecules Gateway will be a useful tool, as experienced researchers can view experimental data and thus assess the validity of a suggested annotation and identify duplicated entries. Less experienced researchers can be guided by the annotation categories and by the consistency warnings associated with each molecule. The Molecules Gateway contains far more molecules than those reported from all types of microorganisms after about 80 years of worldwide efforts.[36] Samples in untargeted metabolomics studies typically contain thousands of different molecules, the vast majority of which remain unidentified.[37] Thus, it is not surprising that most molecules in the Molecules Gateway are listed as unknown. At the same time, literature data suggest that only around 10% of molecules can be annotated, especially in nonmodel organisms[35] and the best-in-class annotation methods can reach an annotation accuracy of around 40%.[37] Nonetheless, providing an annotation and a qualitatively indication of its reliability can help navigating large data sets, and expedite the identification of interesting molecules. The accumulation curve we observe in Figure 5B is consistent with literature trends, which show an appreciable number of novel NPs discovered every year.[38] These data therefore suggest that thousands of additional molecules can be identified by processing the remaining portion of NAICONS' library of actinomycetes or exploring additional fermentation media on the current strains. Genomic analyses suggest that only 3% of bacterial biosynthetic potential has been described, with *Streptomyces* as the most prolific actinomycete genus for potential undescribed metabolites.[29] As many genera are underrepresented in genomic databases, their metabolic potential could not be properly analyzed. Our metabolomic analysis suggest that some genera, for example *Actinoplanes*, *Micromonospora* and *Actinoallomurus* (Figure 7B), contribute a significant percentage of genus-specific molecules and can thus be considered prolific producers of potential novel metabolites, consistent with previous reports.[30,39,40] While the annotation pipeline was employed to analyze actinomycete-derived extracts, it can be applied to extracts derived from other bacteria, from eukaryotic microorganisms or from higher organisms after the proper reference libraries are implemented to assist annotation and to establish biological consistency. This broadening of scope would empower the Molecules Gateway to furnish a universally applicable framework, where data searches for NPs from different biological sources can be readily associated with the availability of the corresponding chemical samples and biological material for further research. Such a comprehensive expansion may engender a transformative paradigm in the realm of NPs discovery and exploitation, fostering interdisciplinary collaborations and catalyzing innovation across many sectors.

EXPERIMENTAL SECTION

General Experimental Procedures. Strain cultivation and extract preparation. The reference strains *Streptomyces coelicolor* M145, *S.avermitilis* NRRL 8165, *S. griseus* DSM 40236, *S. venezuelae* ATCC 15439, *Streptomyces* sp. ID38640, *S. rimosus* ATCC 10970, *S. mobaraensis* ATCC 15003 and *S. glaucescens* GLA000 and strains from the NAICONS collection were cultured as described.[18] For extract preparation, strains were cultivated for 3 days (*Streptomyces*) or for 7 days (other genera) at 30 °C and 200 rpm. Production media used were: SV2,[41] M8, G1/0 and INA5,[42] M8acid (M8 with pH adjusted to 5.6), AF/A[43] and Mare2Br (g/L; $MgSO_4.7H_2O$ 12.3, NaCl 11.7, soluble starch 10, glucose.H_2O 5, casein hydrolysate 2, $CaCl_2.2H_2O$ 1.1, yeast extract 1, meat extract 1, $CaCO_3$ 1, KCl 0.75, KBr 0.05, $MnCl_2.4H_2O$ 0.0079, $CuSO_4.5H_2O$ 0.0064, $ZnSO_4.7H_2O$ 0.0015, $FeSO_4.7H_2O$ 0.001). For extract preparation, 4 mL EtOH was added to a 2 mL sample of each culture in a 15 mL centrifuge tube. The tube was shaken for 1 h at 30 °C and centrifuged at 4000 rpm for 8 min. Then, 125 μL from each supernatant was transferred into individual wells of 96-well microtiter plates (80 extracts per well). Extracts from the seven unfermented production media were also prepared. **LC-HRMS analysis**. Samples (8 μL) were analyzed using a Vanquish UHPLC system (Thermo Fisher Scientific) with a YMC-Triart ODS column (3.0 × 100 mm, S-1.9 μm, 12 nm) coupled to an Orbitrap Exploris 120 High-Resolution Mass Spectrometer (HRMS, Thermo Scientific Scientific) with Untargeted Data-Dependent Acquisition (DDA) analysis. The mobile phase, which was delivered at 0.8 mL min−1 at 40 °C, consisted of 0.1% formic acid in H_2O (A), LCMS-grade MeCN (B) and LCMS-grade isopropyl alcohol (C). The runtime sample analysis was 23 min and the mass to charge (m/z) ratio (MS1scan) was measured in the range from 150 to 2000. Further details as reported in Vind et al. 2023.[44]

Libraries and Databases. Dereplication Library. We made use of the customizable function of the Compound Discoverer suite (mzVault) to build a library of annotated compounds, which includes fragmentation patterns, adduct types, t_R, SMILES and InChIKey of each molecule. The Dereplication Library initially consisted of 452 reference molecules: 92 reference standards and metabolites characterized in previous works and 360 manually curated molecules (Table S3). After each annotation run the Dereplication Library was enriched with newly observed molecules and updated from the Molecule Gateway (see below). **MS2Query Library**. The MS2Query default library was the GNPS[9] library as of December 15, 2022, containing 314k MS2 fragmentation corresponding to 24k unique molecules. MS2Query was then trained on the default library expanded with data from the 452 reference molecules. [This step took 24 h using a single light node (processor Intel E5−2683 V4 2.1 GHz, 4 core and 8 GB of RAM) on the HPC INDACO computing platform (www.indaco.unimi)]. **Bacterial Database (Bacterial-DB)**. The database to run MolDiscovery was built by retrieving entries from NPAtlas,[36] Antimarin[45] and ABL[46] leading to 37,549 molecules originating from bacterial sources. All MS2 fragmentations were predicted by processing the database using the default MolDiscovery parameters. **Natural Products Database (NP-DB)**. This database, which includes 164k unique molecules, was created by merging data from NPAtlas,[36] Antimarin,[45] Coconut,[47] Lotus[48] and ABL,[46] and includes key molecular attributes (SMILES, InChIKey, MF, names, exact mass, literature references and biological source, with 15,614 Actinomycete-

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628

derived molecules selectively flagged). The predicted t_R for each molecule was calculated using jp^{2}rt. **Extract Database (Extract-DB)**. It contains numerical codes for each extract associated with data on the producer strain code, its genus classification, the cultivation procedure and the plate number. The DB is used to retrieve and load metadata into the Molecules Gateway as described below.

Annotation Workflow. Tools validation was performed as described in Table 1, Tables S1 and S2. SIRIUS[15] (version 4.9.15) was tested on 219 reference molecules, and accurately calculated MFs for 100, 85 and 72% of molecules with masses up to 400, 600, and 850 Da, respectively (Table S8).

Soupy Software. In order to manage the entire workflow, we developed an ad-hoc software (named Soupy) that integrates several modules to support MS data analysis and signal annotation. The Soupy software was written in Python and the available commands are listed in Table S4. Each of the commands (written in *italic*) is detailed in the following paragraphs. The entire annotation pipeline was implemented as a Snakemake workflow[49,50] which allowed scalability and speed due to parallelization of the process. The Snakemake workflow was executed on Naicons Dell R620 server (SFF 2x E5−2630Lv2 with 6 cores, 2 threads per core, and 128GB RAM), processing one plate at a time. A total of 93 plates were analyzed using the iterative process reported in Table S4. A "ThermoRawFileReader",[51] was accessed through a docker wrapper[52] to achieve the conversion from raw files to mzML files. To facilitate data extraction from the Compound Discoverer output (cdResult file), Soupy incorporates the PyEDS library,[53] which enables programmatic access to cdResults files. Additionally, Soupy utilizes RDKit, a comprehensive collection of computer chemistry and machine learning tools written in both C++ and Python[54] to clean up the data, including the conversion from SMILES to InChIKey.

The annotation pipeline (Figure S1) was run as follows:

1. The Consolidation and Filtering Step: each 80-extract plate was analyzed together with three H_2O:MeOH 1:1 blanks with the untargeted metabolomics workflow from Compound Discoverer (Version 3.3 SP2; Figure S2). Parameters applied to each node of the workflow are available as Supporting Information. Consolidation was performed by the "Group Compounds" node, which receives inputs from the following nodes: "Detect Compounds", "Align timeTimes" and "Select Spectra") using the default parameters. Filtering was performed on the Compound Discover output (.cdResult file) using the Soupy software (see below), selectively extracting the signals with t_R 0.5 to 15 min, peak area $\geq 10^7$ and peak rating ≥ 8 (a Compound Discoverer parameter indicating the peak quality). Signals present in blanks were discarded.
2. Internal Dereplication Step: this involved using the "Group Compounds" and "Search mzVault" nodes—with the latter receiving input form the former—of Compound Discoverer (Figure S2) to recognize the same molecule when present in different extracts of the same run (intrarun dereplication) and of different runs (inter-run dereplication using the Dereplication Library). Compounds that matched the Dereplication Library were exported by the *cdresult2matches* command as a matches.json file.
3. The Annotation Step: this was performed by Compound Discoverer using the original raw HRMS data files, using the "Assign Compound Annotation" node with parameters detailed in Figure S2, adding the NP-DB to the "Search Mass Lists" node and assigning priority for matches inside this database. The computation time ranged from 2 to 9 h. For running MolDiscovery and MS2Query, the original raw HRMS data files were converted into mzML files by using the Soupy *raw2mzml* command (Table S4). MolDiscovery[13] (2.6.0-beta version) was launched via the Soupy *mzml2moldiscovery* command (Table S4) by setting both the product ion and the parent mass threshold equal to 0.01 and by using the preprocessed Bacterial-DB. Running MolDiscovery took around 1 h on a Dell R620 server (SFF 2x E5−2630Lv2 and 128 GB of RAM). MS2Query[14] (Version 0.3) analysis was performed on each single. mzML file by applying default parameters and positive ion mode, and run on the HPC INDACO by the Soupy commands *mzml2indaco*, *slurm2indaco* and *indaco2ms2query* (Table S4). A single light node was used for each mzML file with processor Intel E5−2683 V4 2.1 GHz, 4 core and 8 GB of RAM. Running MS2Query took around 2 h. MS2Query was implemented to account for exact matches only (within 2 ppm mass difference between experimental and theoretical neutral mass). After each annotation run, the Compound Discoverer outputs are an open-format SQLite database (.cdResult), where each molecule can have information ranging from a complete annotation, to MF only or no match. The compounds were extracted by the *cdresults2compound* command as a compounds.json file that contained MS1 and MS2 data, t_R, exact mass, as well as, when available, compound name and predicted MF. The compounds.json file is the input for the *compound2summary* command, which completed and harmonized the information for each compound-based annotation exported from Compound Discoverer. MolDiscovery produces a single .tsv file for the 80 extracts and MS2Query individual .csv files for each extract. Additionally, the Compound Discoverer output is structured on compound-level logic (i.e., after signal consolidation and assigning a unique index named Compound_ID), while MolDiscovery and MS2Query operate at the MS2-scan level, potentially resulting in different annotations for the different adducts originating from a single compound. The MolDiscovery. tsv file (significant_matches_plate.tsv, see Table S4) was the input for the *moldiscovery2summary* command, which cleaned up and harmonized the scan-based annotations selecting those with a score higher than 50. Since MS2Query renames the scan numbers from the raw files, the *mzml2summary* command exported a list of "extract−scan number pairs" for all MS2 fragmentation found in the MS2Query 80 mzML files and saved a single mzml.tsv summary file. Starting from the mzml.tsv summary and the 80 csv files, the *ms2query2summary* command produced a single ms2query.tsv summary which aligns the MS2Query output to each "extract−scan number pairs" in the original mzML file and ultimately cleaned up and harmonized the annotations from MS2Query possessing a score higher than 0.63. SIRIUS was run, by using the output from the *cdresult2compound* command (compounds.json file) as input: the *compounds2mgf* command extracted an mgf file containing MS signals for molecules with molecular mass lower than 850 Da, irrespective of their annotation, and selecting protonated single charge adduct types only. Then, the *mgf2sirius* command run SIRIUS with the configuration reported in Table S8 and the output file formula_identification.tsv, with predicted MFs and associated scores, was converted into a sirius.tsv file by *sirius2summary*, which filtered results for scores greater than 0 and converted the MF according to the Hill system.[55] The four. tsv summary files were the input for the *compute_annotations* command, which returned annotation on the filtered and unmatched compounds. The annotations were selected after aligning the summaries to squash multiple predictions and running the decision tree with consistency checks and ranking, as described below. For aligning the summaries, the outputs from each software were aligned using a multilevel index system (MultiIndex) in which the first level included the Compound_ID, the second level the Extract Number and the third level the Compound Discoverer-scan number (progressively numbered during the MS data acquisition). In order to retrieve the most reliable annotations from MolDiscovery and MS2Query (which perform scan-based annotations), multiple data were squashed into a single representative element by relying on four sequential criteria: highest binned score, SIRIUS agreement, most frequent InChiKey, lowest Compound Discoverer scan number. For score binning, MS2Query scores (scoring range [0.3,1]) were split into 20 uniform bins, while MolDiscovery scores (scoring range [50,220]) were split into three bins for scores under 100 and into 17 bins for scores above

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628

100.[56] The MFs computed by MS2Query or MolDiscovery were individually compared with those predicted by SIRIUS, and scored as 1 = agreement; −1 = disagreement; 0 = no predicted formula or disagreement when exact mass >600 Da. For the InChIKey frequency, we considered the scans yielding the same InChIKey from a specific Compound Discoverer_ID, excluding scans with missing prediction. Finally, the annotation associated with the lowest Compound Discoverer scan number was preferred. Squashing resulted in a simple index based on Compound Discoverer_ID. An example is reported in Table S5.

4. Consistency and ranking Step: the three criteria considered, the biological source, the MF and the t_R of the annotated molecules, were part of the decision tree (Figure S3, Table S6), executed by the *compute_annotations* command and resulting into 12 annotation bins. Each annotation resulted in a 3-tuple of values within {−1,0,1}. The biological consistency was rated as 1 when the biological source field in the NP-DB was True; 0 if the InChIKey was not present in the NP-DB or no InChiKey was assigned; or −1 if the biological source field in the NP-DB was False. Similarly, the t_R consistency was rated as 1 if the predicted t_R was within 1.4 min of the experimental t_R; −1 if the difference exceeded 1.4 min; and 0 if no InChIKey was proposed or if the proposed InChiKey was not present in the NP-DB. The proposed MF was scored as 1 in case of agreement with the SIRIUS prediction; *−1* in case of disagreement; and 0 if at least one tool did not provide a MF. The arrangement aided in their lexicographic sorting, where consistency 3-tuples were categorized into (not mutually exclusive) classes as follows: class **T** comprised tuples with at least two occurrences of 1, in any position; class **A** is a subset of **T**, with tuples with a 1 in the first position (i.e., the biological source consistency is 1) and at least one additional 1; class **O** comprises tuples with solely a 1 in the first position (Table S6). These classes were employed in subsequent steps, to ascertain consistency winners. To determine the "winning" annotation, we also differentiated scenarios when an InChIKey was predicted from those when no prediction was made. The InChIKey match winner: if the number of predicted InChiKeys was greater than 1, and 2 or 3 of them are identical, the winner was selected according to the name assigned, in the order, by MolDiscovery, Compound Discoverer, and MS2Query. The resulting annotation were assigned to bins **1a** and **1b**, depending on whether there was consistency or not, respectively, with the biological source (Figure S3 and Table S6). The Name similarity winner: if the number of InChIKeys was greater than 1 but they were different, the predicted compound names were pairwise compared. If the similarity exceeded 75%, both names were included; after all the comparisons the set can contain two or three names. [Note that since similarity is not transitive, a list containing three names does not necessarily mean that they are all similar in pairs.] To assign the molecule's name, the winner was subsequently determined according to the name assigned, in order, by MolDiscovery, Compound Discoverer, and MS2Query. The resulting annotation were assigned to bins **2a** and **2b**, depending on whether there was consistency or not, respectively, with the biological source (Figure S3 and Table S6). The consistency winners: when the number of InChiKeys was greater than 1, but they were not identical and the name similarity was lower than 75%, a set of four winners was established as detailed in Table S6. Accordingly, the annotation bins **3a**−**b**, **4a**−**b** were generated, together with **5a** (no consistency, or a single consistency different from biological source). The single InChIKey case: if the number of InChIKey was 1 (i.e., only one annotation was proposed), the annotations were assigned to bin **3c when in** consistency class **A**, in bin **4c** when in consistency class **O** or in bin **5b** when there was no consistency with the biological source.

Molecules Gateway. The web application Molecules Gateway was developed in Python using the Django framework using the PostgreSQL engine with the RDKit cartridge extension to store and search molecular structures. The database structure was defined using Django ORM, the *django-rdkit* library provided an implementation of the custom fields and functions of RDKit. The web application defers asynchronous long tasks to a distributed queue implemented using the Celery library. Using the process described in the previous steps, the following information is provided for every compound: the tool (Compound Discoverer, MolDiscovery or MS2Query) whose *prediction* is used in the final annotation; the *annotation bin*, a set of *warnings* that highlight discordant annotations or consistencies. Moreover, a total of 1031 molecules were identified and labeled as "component of unfermented media" by matching between the signals originated from extracts of unfermented media and the Reference Library obtained from all analyzed 6,5k extracts in the Molecules Gateway. Users can browse the Molecules Gateway using two search pages, simple and advanced. User queries are performed on a materialized view, a PostgreSQL table that acts like a computed cache to preaggregate and assemble data for user visualization. This view is manually refreshed using custom-developed scripts, as it takes one-2 h to produce the result. The advanced search is built using React.js, *react-querybuilder* for composing the UI, and *react-ketcher* to draw molecular structures; the query is then serialized as JSON and sent to Django. The query is validated using JSON Schema and then is converted to Django ORM to actually perform the search in the database. **Data upload**. The *catalog_export* command compiled a plate.JSON file containing all the information to be imported into the Molecules Gateway. It received as input: the matches.json file produced by *cdresult2matches*; the compounds.json produced by *cdresults2compound*; the excel file produced by *compute_annotations*; the sirius.tsv file produced by *sirius2summary*; the directory housing the IDmap file generated by *vaults_export*. Finally, *catalog_export* appended the compounds lacking annotation within the "compound.json" labeling as "unknown" and appending the SIRIUS-generated MF when exact mass was lower than 850 Da. Additionally, Soupy used the information from comnpound.json file, about the presence of compounds in extracts, to query the Extracts-DB and retrieve the information on producer strain and its taxonomy. The *Export2catalogue* command use the plate.JSON file produced by *catalog_export* command and the corresponding Compound Discoverer Results file (cdResult file) as inputs. For molecules matched against the Dereplication Library it appended the list of extracts in which they are detected, while new entries were created for novel annotation and unknown molecules with the list of extracts in which they were detected. MS1 and MS2 fragmentations are extracted from cdResult file for all the adducts. To avoid uploading wrongly assigned adducts, the theoretical neutral mass is calculated and if differs from the experimental one more than 2 ppm the adduct is discarded. After the upload to Molecule Gateway is complete, integrity checks, image generation of chemical structure (using RDKit software), NPclassification (using the REST API provided by NP Classifier[25]), and synonym search (using Pubchem REST APIs) tasks are launched. **Updating the Dereplication Library from Molecules Gateway**. Dereplication Library was updated after completing each annotation and uploading run, using the *catalogue2-vault* command to allow the next Dereplication step. For each entry the mzVault file contained the experimental MS2 fragmentations, t_Rs together with name, SMILES, InChIKey, measured neutral mass, precursor mass and related adduct type. To allow the tracking of the next matches, the same commands also generated a map file (IDmap) that correlate the unique Molecules Gateway IDs to the Dereplication Library indexes.

Quality Control of Annotations. In order to establish the reliability of the annotations in the Molecules Gateway, we grouped the annotation bins into six logical categories and randomly picked 20 equally spaced molecule IDs within each of the categories excluding the last (bins **5a** and **5b**; Table S6). Manual curation was performed by evaluating the following criteria: consistency of the MS2 fragmentation with the proposed chemical structure (categorized as not-consistent, not-plausible, plausible and consistent, or insufficient data); consistency of the detected adducts with the proposed chemical structure; presence of annotated congeners in the same extract; presence of a

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628

consistent UV spectrum in the chromatogram; and molecular networking in GNPS. For the last item, we exported a single MS2 fragmentation per each adduct type for the 22,912 molecules with chemical structure—divided in roughly equal subgroups on the basis of t_R—and run a GNPS analysis (cosine ≥0.7). When one of the 20 molecules clustered in a network (see Supporting Information) the similarity in chemical structure was considered consistent with the annotation. Manual validation of MS2 fragmentation was used as the key criterion for evaluating accuracy of the annotation, with the additional criteria serving to reinforce or challenge the MS2 fragmentation outcomes. Consequently, MS2 fragmentation deemed "plausible/consistent" or "not-plausible/not-consistent" led to an evaluation of the annotation being either "likely" or "unlikely", respectively (Table S7). Fragmentation data of the 100 molecules submitted to manual curation are available as Supporting Information.

■ ASSOCIATED CONTENT

Data Availability Statement

Molecular Network used for quality control is available on GNPS (http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=af0dac3761cb4a46bba18f019cd4abcc; http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=da256bb97e8c405da4be1e5b53256e47; http://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=67b0d8e68e704323ad4f690ef610b536). Videos illustrating the described uses of the Molecules Gateway are available through the following link: https://micro4all.com/video-tutorial/. Interactive version of Figure 6 and Figure S8 are available at the following links: https://micro4all.com/moldiv/ and https://micro4all.com/moldiv-singletons/, respectively. The version of the heatmap.3 function used for Figure 7 Panels A and B is available at https://github.com/TV27/heatmap-metabo/blob/main/revheatmap3-R. The software jp^{2}rt to predict t_R is available through GitHub (https://github.com/mapio/jp2rt/tree/v0.2.3) or upon request. Software Soupy is available upon request.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jnatprod.4c00857.

Figures S1−S8 (PDF)
Tables S1−S8 (XLSX)
MS1 and MS2 fragmentation data of 59 molecules considered for Concept Validation (TXT)
MS1 and MS2 fragmentation data of 100 molecules considered for Quality Control (TXT)
MS1 and MS2 fragmentation data of 48 molecule-sresulted likely in Quality Control (TXT)

■ AUTHOR INFORMATION

Corresponding Author

Sonia I. Maffioli − *NAICONS SRL, 20139 Milan, Italy;* orcid.org/0000-0001-5489-5995; Email: smaffioli@naicons.com

Authors

Matteo Simone − *NAICONS SRL, 20139 Milan, Italy;* orcid.org/0009-0006-6025-6261
Marianna Iorio − *NAICONS SRL, 20139 Milan, Italy;* orcid.org/0000-0001-8669-5875
Paolo Monciardini − *NAICONS SRL, 20139 Milan, Italy;* orcid.org/0000-0002-8727-2791
Massimo Santini − *University of Milan, 20122 Milan, Italy*
Niccolò Cantù − *Code Atlas SRL, 20025 Legnano, Italy*
Arianna Tocchetti − *NAICONS SRL, 20139 Milan, Italy*
Stefania Serina − *NAICONS SRL, 20139 Milan, Italy;* orcid.org/0009-0006-3457-6202
Cristina Brunati − *NAICONS SRL, 20139 Milan, Italy;* orcid.org/0009-0007-4791-3455
Thomas Vernay − *NAICONS SRL, 20139 Milan, Italy; University of Milano-Bicocca, 20126 Milan, Italy*
Andrea Gentile − *NAICONS SRL, 20139 Milan, Italy;* orcid.org/0000-0002-1224-1130
Mattia Aracne − *Code Atlas SRL, 20025 Legnano, Italy*
Marco Cozzi − *Code Atlas SRL, 20025 Legnano, Italy*
Justin J. J. van der Hooft − *Wageningen University, 6708 PB Wageningen, The Netherlands;* orcid.org/0000-0002-9340-5511
Margherita Sosio − *NAICONS SRL, 20139 Milan, Italy;* orcid.org/0000-0003-4297-6936
Stefano Donadio − *NAICONS SRL, 20139 Milan, Italy;* orcid.org/0000-0002-2121-8979

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jnatprod.4c00857

Author Contributions

¶M. Simone, M. Iorio, and P. Monciardini contributed equally.

Funding

This project was supported by the Italian Government MIUR Funding No. DM60066 and has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie Grant Agreement No. 955626.

Notes

Molecules Gateway is a commercial resource provided by NAICONS Srl (www.micro4all.com/molecules-gateway/) accessible to subscribers as listed in www.micro4all.com/pricing/. A DEMO version of the Molecule Gateway is freely available at www.micro4all.com/molecules-gateway/.

The authors declare the following competing financial interest(s): All authors are employees, shareholders and/or members of the advisory board of NAICONS Srl. Specifically, S.D. is co-founder, CEO and President of the Board; M.So. is employee, co-founder, COO and Board Member; S.I.M., M.Si., M.I. and P.M. are co-founders and employees; S.S. and C.B. are employees and shareholders; A.G., A.T. and T.V. are shareholders and former employees; M.Sa., M.C. and M.A. are shareholders; J.J.J.v.d.H. is shareholder and member of the Scientific Advisory Board. M.Si., M.I., S.I.M. and S.D. are listed as inventors on an Italian patent application (102023000024171) filed by NAICONS Srl.

■ ACKNOWLEDGMENTS

We thank A. Nourifar, D. Ferreira, C. Yammine and M. Barrili for help with extracts preparation; M. Monga, L. Spampinato, T. Weber and K. Duncan for suggestions and stimulating discussions; T. Neus, A. Alessi and M. Ducci for informatic support.

■ REFERENCES

(1) Newman, D. J.; Cragg, G. M. *J. Nat. Prod.* **2020**, *83* (3), 770−803.
(2) Zhang, Q.-W.; Lin, L.-G.; Ye, W.-C. *Chin Med.* **2018**, *13* (1), 20.
(3) *Natural Products Isolation*; Sarker, S. D., Nahar, L., Eds.; Methods in Molecular Biology; Humana Press: Totowa, NJ, 2012; Vol. *864*. DOI: 10.1007/978-1-61779-624-1.
(4) De Medeiros, L. S.; De Araújo Júnior, M. B.; Peres, E. G.; Da Silva, J. C. I.; Bassicheto, M. C.; Di Gioia, G.; Veiga, T. A. M.; Koolen, H. H. F. Discovering New Natural Products Using Metabolomics-Based

Approaches. In *Microbial Natural Products Chemistry*; Pacheco Fill, T., Ed.; Advances in Experimental Medicine and Biology; Springer International Publishing: Cham, 2023; Vol. *1439*, pp 185−224. DOI: 10.1007/978-3-031-41741-2_8.
(5) Bauermeister, A.; Mannochio-Russo, H.; Costa-Lotufo, L. V.; Jarmusch, A. K.; Dorrestein, P. C. *Nat. Rev. Microbiol* **2022**, *20* (3), 143−160.
(6) Beniddir, M. A.; Kang, K. B.; Genta-Jouve, G.; Huber, F.; Rogers, S.; Van Der Hooft, J. J. J. *Nat. Prod. Rep.* **2021**, *38* (11), 1967−1993.
(7) Mahieu, N. G.; Patti, G. J. *Anal. Chem.* **2017**, *89* (19), 10397−10406.
(8) Dunn, W. B.; Erban, A.; Weber, R. J. M.; Creek, D. J.; Brown, M.; Breitling, R.; Hankemeier, T.; Goodacre, R.; Neumann, S.; Kopka, J.; Viant, M. R. *Metabolomics* **2013**, *9* (S1), 44−66.
(9) 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.; Porto, C.; Bouslimani, A.; Melnik, A. V.; Meehan, M. J.; Liu, W.-T.; Crüsemann, M.; Boudreau, P. D.; Esquenazi, E.; Sandoval-Calderón, M.; Kersten, R. D.; Pace, L. A.; Quinn, R. A.; Duncan, K. R.; Hsu, C.-C.; Floros, D. J.; Gavilan, R. G.; Kleigrewe, K.; Northen, T.; Dutton, R. J.; Parrot, D.; Carlson, E. E.; Aigle, B.; Michelsen, C. F.; Jelsbak, L.; Sohlenkamp, C.; Pevzner, P.; Edlund, A.; McLean, J.; Piel, J.; Murphy, B. T.; Gerwick, L.; Liaw, C.-C.; Yang, Y.-L.; Humpf, H.-U.; Maansson, M.; Keyzers, R. A.; Sims, A. C.; Johnson, A. R.; Sidebottom, A. M.; Sedio, B. E.; Klitgaard, A.; Larson, C. B.; Boya P, C. A.; Torres-Mendoza, D.; Gonzalez, D. J.; Silva, D. B.; Marques, L. M.; Demarque, D. P.; Pociute, E.; O'Neill, E. C.; Briand, E.; Helfrich, E. J. N.; Granatosky, E. A.; Glukhov, E.; Ryffel, F.; Houson, H.; Mohimani, H.; Kharbush, J. J.; Zeng, Y.; Vorholt, J. A.; Kurita, K. L.; Charusanti, P.; McPhail, K. L.; Nielsen, K. F.; Vuong, L.; Elfeki, M.; Traxler, M. F.; Engene, N.; Koyama, N.; Vining, O. B.; Baric, R.; Silva, R. R.; Mascuch, S. J.; Tomasi, S.; Jenkins, S.; Macherla, V.; Hoffman, T.; Agarwal, V.; Williams, P. G.; Dai, J.; Neupane, R.; Gurr, J.; Rodríguez, A. M. C.; Lamsa, A.; Zhang, C.; Dorrestein, K.; Duggan, B. M.; Almaliti, J.; Allard, P.-M.; Phapale, P.; Nothias, L.-F.; Alexandrov, T.; Litaudon, M.; Wolfender, J.-L.; Kyle, J. E.; Metz, T. O.; Peryea, T.; Nguyen, D.-T.; VanLeer, D.; Shinn, P.; Jadhav, A.; Müller, R.; Waters, K. M.; Shi, W.; Liu, X.; Zhang, L.; Knight, R.; Jensen, P. R.; Palsson, B. Ø.; Pogliano, K.; Linington, R. G.; Gutiérrez, M.; Lopes, N. P.; Gerwick, W. H.; Moore, B. S.; Dorrestein, P. C.; Bandeira, N. *Nat. Biotechnol.* **2016**, *34* (8), 828−837.
(10) M/Z Cloud: Advanced Mass Spectral Database. https://www.mzcloud.org.
(11) Russo, F.; Ottosson, F.; Van Der Hooft, J. J. J.; Ernst, M. Deep Learning Models for LC-MS Untargeted Metabolomics Data Analysis. In *From Computational Logic to Computational Biology*; Cantone, D., Pulvirenti, A., Eds.; Lecture Notes in Computer Science; Springer Nature Switzerland: Cham, 2024; Vol. *14070*, pp 128−144. DOI: 10.1007/978-3-031-55248-9_7.
(12) https://www.thermofisher.com/compounddiscoverer.
(13) Cao, L.; Guler, M.; Tagirdzhanov, A.; Lee, Y.-Y.; Gurevich, A.; Mohimani, H. *Nat. Commun.* **2021**, *12* (1), 3718.
(14) De Jonge, N. F.; Louwen, J. J. R.; Chekmeneva, E.; Camuzeaux, S.; Vermeir, F. J.; Jansen, R. S.; Huber, F.; Van Der Hooft, J. J. J. *Nat. Commun.* **2023**, *14* (1), 1752.
(15) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. *Nat. Methods* **2019**, *16* (4), 299−302.
(16) Santini, M.; Simone, M.; Iorio, M.; Maffioli, S. I. *Jp2rt*. https://github.com/mapio/jp2rt/tree/v0.2.3.
(17) Nett, M.; Ikeda, H.; Moore, B. S. *Nat. Prod. Rep.* **2009**, *26* (11), 1362.
(18) Iorio, M.; Davatgarbenam, S.; Serina, S.; Criscenzo, P.; Zdouc, M. M.; Simone, M.; Maffioli, S. I.; Ebright, R. H.; Donadio, S.; Sosio, M. *Sci. Rep* **2021**, *11* (1), 5827.
(19) Monciardini, P.; Iorio, M.; Maffioli, S.; Sosio, M.; Donadio, S. *Microbial Biotechnology* **2014**, *7* (3), 209−220.
(20) Chamberlin, J. W.; Chen, S. *J. Antibiot.* **1977**, *30* (3), 197−201.
(21) Osterman, I. A.; Khabibullina, N. F.; Komarova, E. S.; Kasatsky, P.; Kartsev, V. G.; Bogdanov, A. A.; Dontsova, O. A.; Konevega, A. L.; Sergiev, P. V.; Polikanov, Y. S. *Nucleic Acids Res.* **2017**, *45* (12), 7507−7514.
(22) Waksman, S. A.; Woodruff, H. B. *Experimental Biology and Medicine* **1942**, *49* (2), 207−210.
(23) Franck, E.; Crofts, T. S. *npj Antimicrob Resist* **2024**, *2* (1), 3.
(24) Lee, J.; Um, S.; Kim, E.-H.; Kim, S. H. *J. Nat. Prod.* **2024**, *87* (4), 733−742.
(25) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. B.; Van Der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. *J. Nat. Prod.* **2021**, *84* (11), 2795−2807.
(26) Morgan, H. L. *J. Chem. Doc.* **1965**, *5* (2), 107−113.
(27) Bajusz, D.; Rácz, A.; Héberger, K. *J. Cheminform* **2015**, *7* (1), 20.
(28) Griffith, D. M.; Veech, J. A.; Marsh, C. J. *J. Stat. Soft.* **2016**, DOI: 10.18637/jss.v069.c02.
(29) Gavriilidou, A.; Kautsar, S. A.; Zaburannyi, N.; Krug, D.; Müller, R.; Medema, M. H.; Ziemert, N. *Nat. Microbiol* **2022**, *7* (5), 726−735.
(30) Parenti, F.; Coronelli, C. *Annu. Rev. Microbiol.* **1979**, *33* (1), 389−411.
(31) Yurekten, O.; Payne, T.; Tejera, N.; Amaladoss, F. X.; Martin, C.; Williams, M.; O'Donovan, C. *Nucleic Acids Res.* **2024**, *52* (D1), D640−D646.
(32) Leao, T. F.; Clark, C. M.; Bauermeister, A.; Elijah, E. O.; Gentry, E. C.; Husband, M.; Oliveira, M. F.; Bandeira, N.; Wang, M.; Dorrestein, P. C. *Nat. Metab* **2021**, *3* (7), 880−882.
(33) Zuffa, S.; Schmid, R.; Bauermeister, A.; P. Gomes, P. W.; Caraballo-Rodriguez, A. M.; El Abiead, Y.; Aron, A. T.; Gentry, E. C.; Zemlin, J.; Meehan, M. J.; Avalon, N. E.; Cichewicz, R. H.; Buzun, E.; Terrazas, M. C.; Hsu, C.-Y.; Oles, R.; Ayala, A. V.; Zhao, J.; Chu, H.; Kuijpers, M. C. M.; Jackrel, S. L.; Tugizimana, F.; Nephali, L. P.; Dubery, I. A.; Madala, N. E.; Moreira, E. A.; Costa-Lotufo, L. V.; Lopes, N. P.; Rezende-Teixeira, P.; Jimenez, P. C.; Rimal, B.; Patterson, A. D.; Traxler, M. F.; Pessotti, R. D. C.; Alvarado-Villalobos, D.; Tamayo-Castillo, G.; Chaverri, P.; Escudero-Leyva, E.; Quiros-Guerrero, L.-M.; Bory, A. J.; Joubert, J.; Rutz, A.; Wolfender, J.-L.; Allard, P.-M.; Sichert, A.; Pontrelli, S.; Pullman, B. S.; Bandeira, N.; Gerwick, W. H.; Gindro, K.; Massana-Codina, J.; Wagner, B. C.; Forchhammer, K.; Petras, D.; Aiosa, N.; Garg, N.; Liebeke, M.; Bourceau, P.; Kang, K. B.; Gadhavi, H.; De Carvalho, L. P. S.; Silva Dos Santos, M.; Pérez-Lorente, A. I.; Molina-Santiago, C.; Romero, D.; Franke, R.; Brönstrup, M.; Vera Ponce De León, A.; Pope, P. B.; La Rosa, S. L.; La Barbera, G.; Roager, H. M.; Laursen, M. F.; Hammerle, F.; Siewert, B.; Peintner, U.; Licona-Cassani, C.; Rodriguez-Orduña, L.; Rampler, E.; Hildebrand, F.; Koellensperger, G.; Schoeny, H.; Hohenwallner, K.; Panzenboeck, L.; Gregor, R.; O'Neill, E. C.; Roxborough, E. T.; Odoi, J.; Bale, N. J.; Ding, S.; Sinninghe Damsté, J. S.; Guan, X. L.; Cui, J. J.; Ju, K.-S.; Silva, D. B.; Silva, F. M. R.; Da Silva, G. F.; Koolen, H. H. F.; Grundmann, C.; Clement, J. A.; Mohimani, H.; Broders, K.; McPhail, K. L.; Ober-Singleton, S. E.; Rath, C. M.; McDonald, D.; Knight, R.; Wang, M.; Dorrestein, P. C. *Nat. Microbiol.* **2024**, *9* (2), 336−345.
(34) El Abiead, Y.; Strobel, M.; Payne, T.; Fahy, E.; O'Donovan, C.; Subramamiam, S.; Vizcaino, J. A.; Zuffa, S.; Xing, S.; Mannochio-Russo, H.; Mohanty, I.; Zhao, H. N.; Caraballo-Rodriguez, A. M.; Gomes, P. W. P.; Avalon, N. E.; Dorrestein, P. C.; Wang, M. Enabling Pan-Repository Reanalysis for Big Data Science of Public Metabolomics Data. *ChemRxiv*, April 16, 2024. DOI: 10.26434/chemrxiv-2024-jt46s.
(35) De Jonge, N. F.; Mildau, K.; Meijer, D.; Louwen, J. J. R.; Bueschl, C.; Huber, F.; Van Der Hooft, J. J. J. *Metabolomics* **2022**, *18* (12), 103.
(36) van Santen, J. A.; Poynton, E. F.; Iskakova, D.; McMann, E.; Alsup, T. A.; Clark, T. N.; Fergusson, C. H.; Fewer, D. P.; Hughes, A. H.; McCadden, C. A.; Parra, J.; Soldatou, S.; Rudolf, J. D.; Janssen, E. M.-L.; Duncan, K. R.; Linington, R. G. *Nucleic Acids Res.* **2022**, *50* (D1), D1317−D1323.
(37) Bach, E.; Schymanski, E. L.; Rousu, J. *Nat. Mach Intell* **2022**, *4* (12), 1224−1237.
(38) Pye, C. R.; Bertin, M. J.; Lokey, R. S.; Gerwick, W. H.; Linington, R. G. *Proc. Natl. Acad. Sci. U.S.A.* **2017**, *114* (22), 5601−5606.
(39) Wagman, G. H. *Annu. Rev. Microbiol.* **1980**, *34* (1), 537−558.

(40) Iorio, M.; Gentile, A.; Brunati, C.; Tocchetti, A.; Landini, P.; Maffioli, S. I.; Donadio, S.; Sosio, M. *RSC Adv.* **2022**, *12* (26), 16640−16655.
(41) Zettler, J.; Xia, H.; Burkard, N.; Kulik, A.; Grond, S.; Heide, L.; Apel, A. K. *ChemBioChem.* **2014**, *15* (4), 612−621.
(42) Donadio, S.; Monciardini, P.; Sosio, M. Chapter 1 Approaches to Discovering Novel Antibacterial and Antifungal Agents. In *Methods in Enzymology*; Elsevier, 2009; Vol. *458*, pp 3−28. DOI: 10.1016/S0076-6879(09)04801-0.
(43) Cruz, J. C. S.; Iorio, M.; Monciardini, P.; Simone, M.; Brunati, C.; Gaspari, E.; Maffioli, S. I.; Wellington, E.; Sosio, M.; Donadio, S. *J. Nat. Prod.* **2015**, *78* (11), 2642−2647.
(44) Vind, K.; Brunati, C.; Simone, M.; Sosio, M.; Donadio, S.; Iorio, M. *ACS Chem. Biol.* **2023**, *18* (4), 861−874.
(45) Blunt, J.; Munro, M.; Laatsch, H. *AntiMarin Database*, 2006.
(46) Simone, M.; Monciardini, P.; Gaspari, E.; Donadio, S.; Maffioli, S. I. *J. Antibiot.* **2013**, *66* (2), 73−78.
(47) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. *J. Cheminform* **2021**, *13* (1), 2.
(48) Rutz, A.; Sorokina, M.; Galgonek, J.; Mietchen, D.; Willighagen, E.; Gaudry, A.; Graham, J. G.; Stephan, R.; Page, R.; Vondrášek, J.; Steinbeck, C.; Pauli, G. F.; Wolfender, J.-L.; Bisson, J.; Allard, P.-M. *eLife* **2022**, *11*, No. e70780.
(49) Mölder, F.; Jablonski, K. P.; Letcher, B.; Hall, M. B.; Tomkins-Tinch, C. H.; Sochat, V.; Forster, J.; Lee, S.; Twardziok, S. O.; Kanitz, A.; Wilm, A.; Holtgrewe, M.; Rahmann, S.; Nahnsen, S.; Köster, J. *F1000Res.* **2021**, *10*, 33.
(50) Köster, J.; Rahmann, S. *Bioinformatics* **2012**, *28* (19), 2520−2522.
(51) http://compomics.github.io/projects/thermorawfileparser.
(52) https://www.compomics.com/.
(53) https://github.com/thermofisherlsms/pyeds.
(54) https://www.rdkit.org/.
(55) Hill, E. A. *J. Am. Chem. Soc.* **1900**, *22* (8), 478−494.
(56) Fayyad, U.; Irani, K. *Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning*; 1993.

https://doi.org/10.1021/acs.jnatprod.4c00857
J. Nat. Prod. 2024, 87, 2615−2628