# A linked Genomics Sequencing and Mass Spectrometry multi-modal dataset and models for Streamlined Natural Products Discovery in Microbial Strain Libraries

Dillon W. P. Tay[+,*,a,e], Winston Koh[+,*,b,e], Shi Jun Ang[c,e,f], Zicong Marvin Wong[c,e], Yi Wee Lim[a,e], Elena Heng[d,e], Naythan Z. X. Yeo[a], Krishnan Adaikkappan[a], Fong Tian Wong[*,a,d,e], and Yee Hwee Lim[*,a,e,g]

**Affiliations:**

a. Institute of Sustainability for Chemicals, Energy and Environment, Agency for Science, Technology and Research, 8 Biomedical Grove, #07-01 Neuros Building, Singapore 138665, Republic of Singapore.
b. Bioinformatics Institute, Agency for Science, Technology and Research, 30 Biopolis Street, #07-01 Matrix Building, Singapore 138671, Republic of Singapore.
c. Institute of High Performance Computing, Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore.
d. Institute of Molecular and Cell Biology, Agency for Science, Technology and Research, 61 Biopolis drive, #07-01 Proteos, Singapore 138673, Republic of Singapore.
e. Singapore Integrative Biosystems and Engineering Research, Agency for Science, Technology and Research, Singapore, Republic of Singapore.
f. Department of Chemistry, National University of Singapore, 3 Science Drive 3, Singapore 117543, Republic of Singapore.
g. Synthetic Biology Translational Research Program, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117597, Republic of Singapore.

*** Corresponding authors**

D. T.: dillon_tay@isce2.a-star.edu.sg

W. K.: Winston_Koh@bii.a-star.edu.sg

F.T.W.: wongft@imcb.a-star.edu.sg

Y.H.L.: lim_yee_hwee@isce2.a-star.edu.sg

+ These authors contribute equally to this work.

**Keywords:** Natural product discovery, protein language model, molecular language model, mass spectroscopy, actinomycetes, prediction, structural elucidation

**ABSTRACT**

Natural products (NP) are instrumental in drug development, but their discovery and validation remain challenging and laborious despite advances in both genomic and analytical technologies. In this study, we demonstrate the use of an integrated multi-modal characterization of a microbial strain library for enhanced natural product discovery. This characterization utilizes language- and transformer-based models, integrated through a cross-validate and rank approach to search a mass spectrometry (MS)-genome multi-modal dataset with high confidence. MS data are analysed using an in-house developed tandem mass spectral MS/MS to structural elucidation workflow (WISE) that features a combination of molecular language and transformer-based models to predict corresponding molecular structures. Simultaneously, the related genomic data is pre-processed using the protein language model (ESM2) to extract meaningful embeddings. As a proof of concept, these models and pre-processed linked MS-genome datasets were applied and validated for the rapid identification of microbial strains capable of producing three diverse natural product compounds with precision ranging from 75-100%. Our findings demonstrate the transformative potential of linked MS-genome datasets at the strain-level to accelerate natural product discovery. This approach can expand the range of biotechnological innovations beyond what is currently known and curated, while also greatly reduce the resources and effort needed for discovery.

## INTRODUCTION

Natural products have been instrumental in drug discovery and therapeutic development, offering unparalleled chemical diversity and unique biological activities[1]. These compounds, derived from a wide range of organisms—including bacteria, fungi, plants, and marine life—have underpinned the development of numerous drugs, such as antibiotics[2], immunosuppressants[3], and anticancer agents[4]. Despite their tremendous potential, the rate of new natural product discoveries has been bottlenecked by inefficiencies in screening and more pertinently, characterizing and mining within the vast amount and diversity of potential biosynthetic sources.

In response to these challenges, advancements in genomics[5,6] and chemical analytics (e.g. $^{13}C$, deuterium or phosphonate labelling, and other studies), including mass spectrometry technologies[7], have provided unprecedented opportunities to scale natural product exploration. Sequencing and curated annotation of microbial genomes have revealed the vast potential of biosynthetic gene clusters (BGCs), which encode pathways for synthesizing novel bioactive compounds. However, a critical challenge remains: effectively linking genomic data to molecular structures to predict and validate natural product production. Existing tools, such as antiSMASH[8] and PRISM[9], have been instrumental in addressing this challenge by automating the identification, annotation of BGCs and predicting their products classes based on genomic sequences. These tools provide foundational insights but are constrained by predefined rules of known BGCs. Recent advancements in deep learning-based approaches have been expanding the toolkit for BGC annotation and prediction. For instance, models like DeepBGC[10] and BiGCARP[11] leverage embeddings from protein language models (e.g., ESM1-B) to improve the predictive performance of BGC-

product relationships. These approaches highlight the potential of machine learning to overcome traditional bottlenecks, but still face limitations in linking genomic and MS data for comprehensive structural elucidation.

High resolution mass spectrometry and tandem mass spectrometry (MS/MS) are frequently used to characterize natural products produced by microbes. Improvement in mass spectrometry technologies and their sensitivities have even enabled detection of low yielding compounds. While MS/MS data provides signature fingerprint information about the molecules, structural elucidation of individual compounds from complex sample mixtures remain a long-standing bottleneck[12] for natural product discovery. Often, de novo structural elucidation of a novel natural product requires tedious isolation of the molecule, multiple analytical characterization techniques such as 1D and 2D NMR, as well as expert interpretation. Alternative methods rely on matching MS/MS fingerprints against reference libraries, but these face significant limitations, such as the limited scope of current MS/MS reference libraries[13,14] despite recent advancements in deep learning matching approaches such as MS2DeepScore[15] and integrated analytics[16]. These technologies are limited by curated/known characterized compounds (~400K; COCONUT[31]) and struggle to keep pace with the expanding complexity of the natural chemical space[17] (~114M, PubChem, accessed Jan 2025), leaving >98% of MS fingerprints unidentified in untargeted metabolomics experiments[18].

To accelerate natural product discovery, we describe a framework that combines a two-pronged bio- and chem- informatics analysis to enable scalable inference of secondary metabolites from a library of microbial strains (Figure 1). Mass spectral[19]

and genomic data from a set of 54 native strains and their derived mutant strains were processed using a suite of language- and transformer-based technologies. We employ the ESM2[20] protein language model to generate embeddings from protein sequences of our genomic dataset, which overcomes the challenges of a lower-quality fragmented genomic dataset while enabling scalable and efficient protein sequence comparisons through vector databases. For LC-MS/MS data obtained from fermentation extracts, an in-house technology encompassing (1) molecular language processing, (2) transformer-based structure-to-MS/MS models, and (3) a data processing workflow was developed. Leveraging (1) and (2) enables direct structural elucidation of both known and unknown compounds from LC-MS/MS data of complex mixtures. The integration of generative molecular language model with predicted spectra enhances the utility of LC-MS/MS data, creating opportunities to explore previously inaccessible chemical space. Here, we demonstrate the utility of these models to pre-process the linked MS-genome datasets in the rapid identification of producers of three diverse natural products, valinomycin, surfactin B, and neomycin B, from this collection of strains.

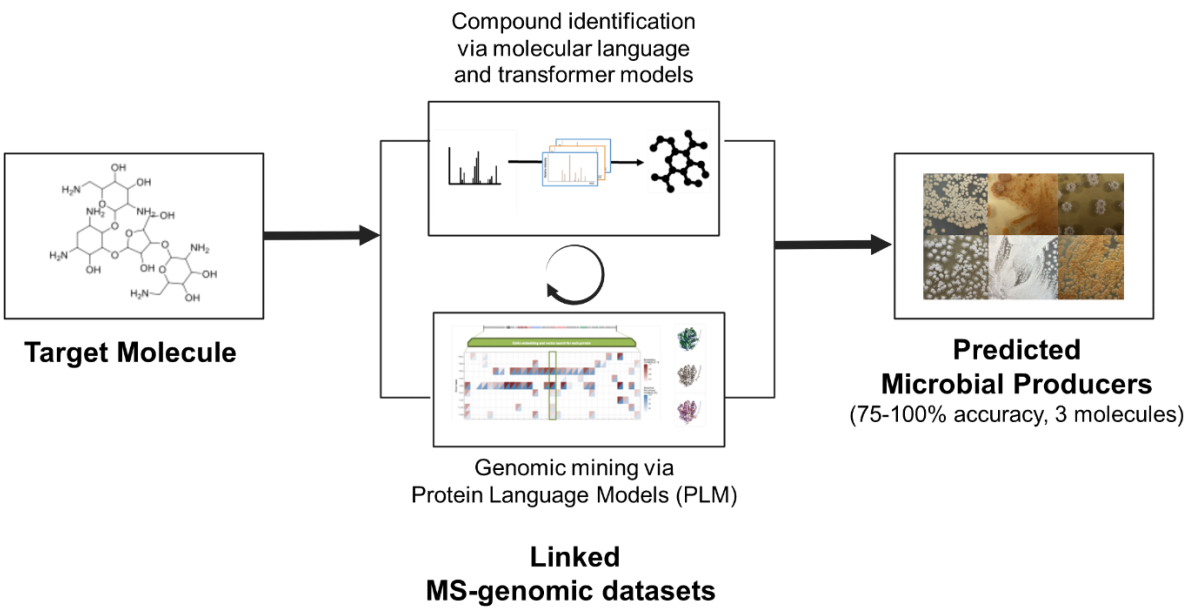

**Figure 1.** Concept of a linked genomics and mass spectrometry (LCMS/MS) multi-modal dataset and predictive models for streamlined discovery of natural products in microbial strain libraries. Putative producers of compounds of interest are identified through co-occurrence of the best ranked strains *via* protein language models embeddings and the highly ranked strains predicted to produce the target molecules *via* our in-house developed tandem mass spectral MS/MS to structural elucidation workflow (WISE).

Collectively, the advancements presented here represent a paradigm shift in analytical methodologies and natural product discovery. They are particularly impactful when working with large microbial libraries. Our approach disrupts the conventional mindset of being limited to identify only well-characterized entities by transforming it to generate informed hypotheses about all features. This bridges the gap between data generation and actionable insights with increased confidence and ultimately facilitates the discovery of relevant strains and corresponding compounds.

## RESULTS

### 1. A highly diverse foundation dataset for illustrating linked multi-modal (MS-genomic) analysis

In our previous study, we reported a flexible and reliable method to stimulate antibiotic production in a diverse set of 54 actinobacteria from the A*STAR National Organism Library[21] using an integrase-mediated multi-faceted approach[19]. We generated 124 different combinations of activators and strains (459 mutants) across 54 actinobacterial strains, which consistently outperformed the respective wild-type strains in terms of metabolite diversity. Fermentation of these native and engineered strains in 3 to 5 different media, resulted in 2,138 fermentation extracts. We utilized high-throughput liquid chromatography-tandem mass spectrometry (LC-MS/MS) to analyse the chemical composition of these extracts, yielding 2,138 distinct LC-MS/MS sample data[22]. The 2,138 distinct LC-MS/MS corresponding data were processed through in-house developed MS/MS to structure technology (*vide infra*). Through GNPS[23] molecular networking analysis, we previously identified 743 different metabolites grouped into 69 clusters, each containing at least two metabolites, and an additional 126 orphan metabolites[22]. Here, these 54 wild-type strains were further sequenced and characterized using Illumina technology (Table S1). This linked multi-modal MS-genomic dataset will form the basis of this study for the rapid identification and prioritization of producer strains of targeted natural products.

### 2. Compound identification from LC-MS/MS data via molecular language and structure-to-MS/MS models

Structural elucidation of compounds in natural product discovery and untargeted metabolomics has long been hampered by the reliance on the limited available curated

reference MS/MS databases[12,17,18]. Current methods[13,14,24] seek to overcome this limitation by developing models that predict MS/MS spectra from known molecular structures, thus extending searches beyond the limited spectral library to more comprehensive molecular structure databases. However, precisely identifying molecular structures from raw spectral data via such predictive models remains challenging (8~13% accuracy, **Table S2**). Generally, the application of these methods in discovery metabolomics is constrained to a filtering mechanism that excludes known structures, rather than a tool to discover novel metabolites. It is noteworthy that expanding MS reference libraries experimentally is a resource-intensive endeavour, creating a barrier to annotating especially underexplored sources.

To address this critical bottleneck, we describe the integration of three technologies comprising (1) molecular language processing[25], (2) structure-to-MS/MS models[26], and (3) a data processing workflow[27], synergistically designed to directly elucidate molecular structures from LC-MS/MS data of mixtures (e.g., fermentation extracts). Molecular language processing generates novel compound structures, expanding the chemical space to include possibilities beyond known entities. Predictive structure-to-MS/MS models then simulate reference spectra for both these novel structures as well as known structures without reference spectra. The data processing workflow ties these elements together, handling complex LC-MS/MS data of mixtures by extracting, denoising, and deconvoluting it to reveal individual compound MS/MS spectra that can then be automatically annotated by the generated MS/MS reference database. While our approach can ultimately be harnessed to identify novel metabolites, here we demonstrate its utility on three known metabolites as a proof of concept. Extending the workflow beyond targeted metabolite identification is anticipated to be a

straightforward process, as the generation of novel structures is inherently integrated into the workflow. Nevertheless, further investigation is required to address potential limitations, such as optimizing the balance between the number of generated structures and computational efficiency to ensure effective and accurate identification.

### 2.1 Molecular language processing for novel compound discovery

To enable the identification of novel structures, we leverage molecular language processing, a generative technology that simulates the "language" of molecular structures. Previously, we had disclosed a long short-term memory (LSTM) recurrent neural network[25] trained in ~400K natural product structures from the public repository, COCONUT[28]. More than 67 million novel natural product-like compounds were sampled from the trained model, expanding the chemical space by two orders of magnitude expansion, which notably also covers previously uncharted physiochemical space.

### 2.2 Predictive transformer-based structure-to-MS/MS models for reference spectra generation

Beyond employing transformer-based structure to MS/MS predictive models[13], these models have also been adapted to include collision energy as a parameter[26]. Applied collision energies influence the extent of molecular fragmentation, significantly affecting the resulting observed MS/MS spectra. By separately fine-tuning models to spectra within ±5 eV from each other (i.e., one model per 10 eV bin), we improved cosine similarity between predicted and experimental MS/MS spectra from an average of 59% to 75%. The fine-tuned models are then used to predict the MS/MS spectra for both known and generated structures to serve as a reference database for spectral

matching and structural elucidation. Due to the high (75% on average) alignment with experimental spectra, there is enhanced reference reliability and marked improvement in the accuracy of computational structural elucidation compared to state-of-the-art approaches (**Table S2**).

### 3.3 The Workflow for Intelligent Structural Elucidation (WISE) for complex mixture analysis

To further translate these capabilities into actionable insights, we developed a data processing workflow[27] that leverages the reference MS/MS database generated from molecular language processing and structure-to-MS/MS predictive modelling to process complex mixture LC-MS/MS data to molecular structures in a scalable manner. Untargeted metabolomics analyses often involve complex mixtures, necessitating the use of denoising and deconvolution techniques to identify individual compounds and their associated MS signals. Our workflow achieves this by performing analyses at the individual *m/z* level, identifying compounds based on their elution pattern over time (**Figure 2A**). This approach enables us to separate MS signals into three categories: (1) “Noise-like”: signals with similar intensities across the entire time range, indicating a noise or background signal, (2) “Fragment-like”: signals with more than 1 major peak, usually localized around the same retention time. This indicates a common scaffold or fragment from compound analogs eluting at different distinct times, and (3) “Compound-like” signals following a regular peak shape distribution, in line with compound elution from the column. Such deep deconvolution at the individual *m/z* level not only enables removal of background signals, but also allows separation of overlapping compound signals, a common problem faced by peak-picking methods.

The MS/MS spectra of the deconvoluted compound-like signals are compared against the generated MS/MS reference database in three metrics system of (1) MS/MS spectra similarity, (2) monoisotopic mass difference, and (3) isotopic distribution similarity, to give a consensus score. MS/MS reference spectra are predicted using the above-described transformer-based models, while monoisotopic masses and isotopic distributions are calculated based on natural isotopic abundances. Identified compound precursor *m/z* is used to shortlist plausible structures that are then more rigorously scored based on specific molecular structure (MS/MS spectra) and molecular formula (monoisotopic mass, isotopic distribution). Ultimately, this comparison produces a list of potential candidate structures, ranked by their consensus score (i.e., WISE score). The top scoring candidate is then proposed as the predicted molecular structure (**Figure 2B**).

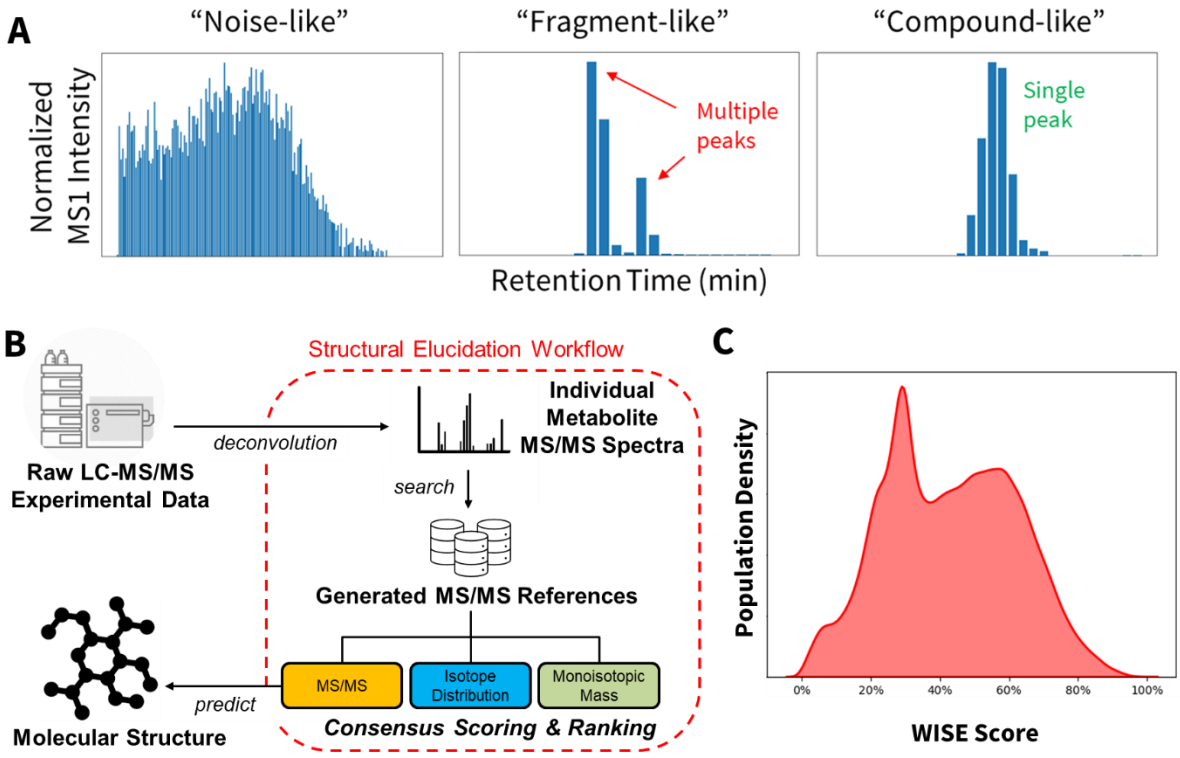

**Figure 2.** (A) Graphs of normalized MS1 intensity of selected *m/z*'s (mass-to-charge ratio) against retention time (min) for three different signal types – "noise-like", "fragment-like", and "compound-like". (B) Overview of the cheminformatics analysis utilizing the Workflow for Intelligent Structural Elucidation (WISE) to annotate individual molecules from raw liquid chromatography-tandem mass spectral (LC-MS/MS) data of fermentation extracts. (C) Distribution of WISE scores for 29,112 compounds identified by WISE from the 2,138 curated LC-MS/MS dataset.

To benchmark our Workflow for Intelligent Structural Elucidation (WISE), we evaluated the Top 1 consensus score ranking of 4,505 natural products from public repositories from the National Institutes of Health (NIH) and Global Natural Products Social Molecular Networking (GNPS) databases[23], to which WISE demonstrated an accuracy of 42%. This represents a substantial improvement over existing methods such as NEIMS[24] and CFM-ID 4.0[14], which achieve accuracies of 8.6%, and ICEBERG[13], which demonstrates an accuracy of 12.9%, highlighting WISE's superior performance in accurately elucidating natural product structures. In this study, WISE is deployed to deconvolute and annotate the 2,138 distinct LC-MS/MS dataset generated from fermenting 124 distinct actinobacteria strains. In this way, we identified 29,112 unique compounds with an average WISE score of 43% (**Figure 2C**). Due to the differences in processing algorithms, the number of unique compounds identified through WISE are nearly 40-fold larger than that previously processed via GNPS molecular networking analysis. Despite its robust performance, WISE inherently depends on data quality, which poses certain limitations. For instance, low-abundance metabolites may produce fewer MS/MS signals or have their signals obscured by random noise, leading to poorer matches with the generated database. This may explain the spike

around ~30% score observed in **Figure 2C**. Additionally, WISE relies on the stability of the molecules being analysed. If a molecule fragments prematurely during ionization, only its fragments will be detected, potentially compromising annotation accuracy. This limitation, however, can be mitigated by optimizing ionization settings, such as employing softer ionization techniques to preserve molecular integrity.

### 3. Simplified genomic library mining via integration with embeddings from protein language models

In parallel, we also explored the use of a protein language model for genomic mining simultaneously. The field of genomic library mining has traditionally relied on sequence similarity to predict biological functions[29] . This approach is grounded in the principle that amino acid sequence determines structure, and thus, similar sequences are likely to share similar structures, which, in turn, can be expected to exhibit similar biological activity, ultimately leading to similar metabolite production. Recent advances in protein language models (PLMs) have significantly enhanced the capabilities of *in silico* biophysical analysis, especially in structure-related tasks such as secondary structure prediction and contact prediction. Embeddings from these models have even been shown to be capable of extracting remote homology information that may be obscured in primary sequences, enabling a more sophisticated understanding of protein function[30,31].

Building on these advancements, we decided to focus on protein-level embeddings rather than embeddings derived from entire BGCs. Protein-level embeddings offer higher resolution, as they encapsulate functional and structural characteristics. This approach also circumvents the need for high-quality contiguous genome assemblies,

as protein sequences can be annotated even from fragmented or incomplete genomes as demonstrated in our current genomic dataset (**Table S1**). By leveraging protein embeddings, we can take advantage of the sophisticated representations provided by protein language models to identify functional and structural similarities across diverse genomic datasets.

To first rank and prioritize protein sequences of interest, we employed our workflow that utilizes PLM embeddings to measure sequence similarity. We utilized a vector database powered by approximate nearest neighbour (ANN) algorithms (e.g., Faiss), enabling rapid retrieval of relevant embeddings. This embedding also addresses the challenge of a significantly larger search space as protein sequences are inherently more extensive than gene clusters. This efficiency is also critical when scaling up to large datasets, where search complexity grows exponentially. After initial ANN-based searches, we employed the Smith-Waterman algorithm for precise sequence alignment and homology quantitation, ensuring high specificity in downstream analyses. For each query protein, we calculated the distance between its embedding vector and the embedding vectors in the index. This distance serves as a measure of similarity, where lower values correspond to greater similarity and result in a higher rank in the search results (**Figures 3A, 4A, 5A**). Faiss's approximate nearest neighbour (ANN) algorithms were used to optimise the efficiency of the distance calculations, enabling rapid searches across our large dataset. With fragmented genomes and a PLM approach, the removal of cluster localisation constraints may lead to the possibility that genes are coincidentally found within the same genome but do not work in concert with one another. To overcome this limitation, strains are also ranked in terms of total similar proteins observed from BGC as a means of prioritising

possible producers, rather than looking at individual protein hits. By combining high-resolution protein embeddings with scalable and efficient search strategies, our strategy provides the biological starting point for cross-modal retrieval that bridges the gap between biological and chemical modalities.

### 4. A co-occurrence prioritization strategy to identify natural products producers leveraging multi-modality of linked MS-genome data

To enhance discovery and tackle the challenges of over-predictions from the individual models, we proposed a multi-modal framework that combines and validates protein language model embeddings with WISE scores derived from MS/MS signals. We achieved the integration of genomic and chemical data through a cross-validate and rank approach, which helped identify strains that are highly likely to align both genomic and chemical profiles (**Figures 3D, 4D, 5D**). This cross-validation across different data modalities increased the confidence in strain prioritisation and identification.

For a compound of interest, we first began by identification of enzymes and proteins linked to its biosynthesis. If the exact biosynthetic pathway of the target compound is unknown, plausible enzymes and proteins that may be linked to its biosynthesis will be used as search vectors. In either scenario, this review provides critical insights into the biological context and highlights relevant genomic and protein sequences. Using this information, we applied vector search techniques to identify and rank strains from our dataset that contain proteins with embeddings similar to those identified in the literature review. Simultaneously, we performed a search for the target compound across the strain library of LC-MS/MS data processed by WISE. Putative strains that

harbour the LC-MS/MS data of the target compound were ranked according to the corresponding WISE consensus score.

To illustrate this concept, the following case studies demonstrate the application of a co-occurrence prioritization strategy aimed at accelerating the discovery and identification of potential producers of three targeted compounds: valinomycin, surfactin B, and neomycin B. We show that by utilizing a co-occurrence strategy which involves a ranking score matrix comprising PLM embeddings and WISE consensus scores, we are able to identify producers at high accuracy. Accuracy ranges from 75% to 100% for these three examples; valinomycin (12 out of 14 correlated strains), surfactin B (5 out of 5 correlated strains), and neomycin B (3 out of 4 correlated strains) within collection of diverse strains in this study (**Figures 3 – 5, Table S3, S5-7**). These case studies were attempted without any pre-validation against external models, use of reference MS/MS data, or high-quality genomic information and annotation, thus demonstrating the approach as a powerful strategy to uncover putative producers.

**CASE STUDY 1: Valinomycin**

Valinomycin is a nonribosomal peptide cyclodepsipeptide produced by a biosynthetic gene cluster consisting of two nonribosomal peptide synthases (miBiG: BGC0000453)[32]. The main mode of action of this macrocyclic ionophore is mainly contributed to its binding and transportation of potassium ions and thus has the potential to disrupt the intracellular ion concentration of the cell. Since its discovery in 1955, it has received continuous attention due to its unique chemical structure and broad biological activities for antitumour, antibacterial, antifungal and more recently

there is a spotlight on its antiviral activity, especially its activity against coronaviruses[33]. Its more prevalent producers are *Streptomyces* species.

In this first example, we used valinomycin as a first demonstration due to extensive characterisation of valinomycin producers in our prior study[19]. The cheminformatics-driven analysis predicted 17 strains (out of 54 strains) as potential producers of valinomycin with WISE consensus scores exceeding 35%. Embedding-based protein alignments predicted 30 strains as containing putative valinomycin biosynthetic genes or proteins with PLM scores exceeding 0.75. At the intersection of chemical and genomic-based prediction, 14 strains were in consensus and proposed as high-confidence leads as producers of valinomycin (**Figure 3D** – **coloured box**). Out of these 14 strains, experimental validation confirmed the presence of valinomycin in 12 strains[22] of which three strains (T265, A1532, A1090) only produced valinomycin in their activated forms (**Tables S5, S9**).

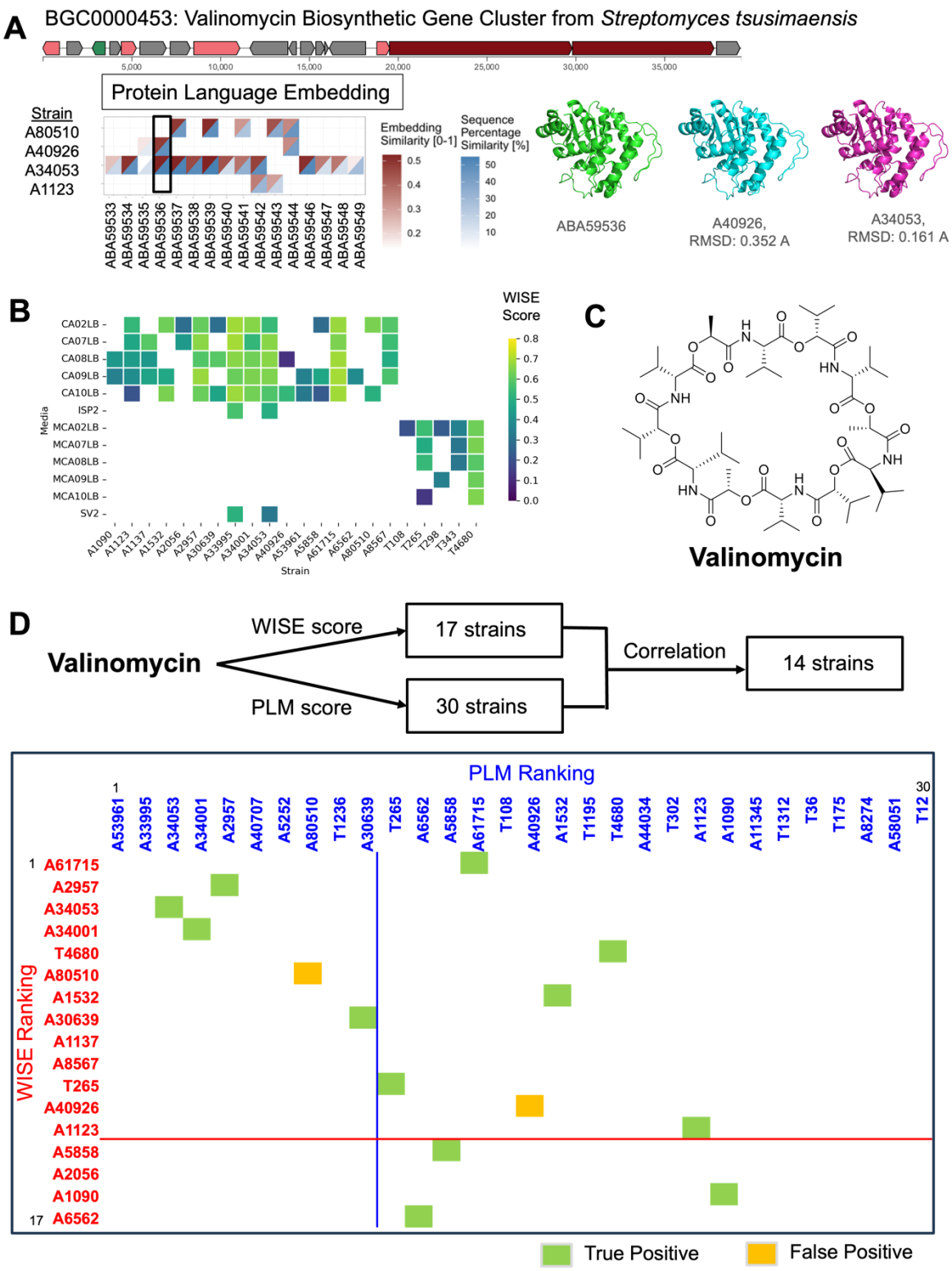


**Figure 3. Case Study: Identification of valinomycin producers.** (A) The biosynthetic gene cluster BGC0000453, responsible for valinomycin production, was analysed by embedding each protein component using the ESM2 protein language model. These embeddings were then queried against a precomputed vector database

of protein embeddings derived from the full dataset in this study. Nearest neighbour algorithms were employed to identify the top-scoring matches, with the results presented in the left inset. The sequences corresponding to the highest-ranked embeddings were retrieved, and pairwise sequence alignments confirmed that the embeddings captured the underlying sequence similarity. A reductase from the cluster (ABA59536.1) was also selected for *in silico* structural prediction using ESMfold. The predicted structure was aligned with known structural data, revealing a high degree of structural similarity, further validating the model's utility in capturing both sequence and structural relationships. (B) Heatmap of the WISE scores of 21 strains-media combination, derived from the liquid chromatography-tandem mass spectrometry (LC-MS/MS) data of fermentation extracts. Only strains with at least 1 media combination with a non-zero WISE score are shown. Fermentation extracts from strain-media combinations that do not have hits for valinomycin are indicated by blanks. Activated strains here are referred to by their parent strain instead of individually (details of the activated strains are in **Table S9**). (C) Structure of valinomycin. (D) Workflow of the co-occurrence strategy and matrix ranking table of the top-ranked strains processed by cheminformatics (WISE) and bioinformatics (PLM) with at least 35% WISE consensus score or 0.75 PLM score, respectively. Strains that have been cross-validated are highlighted in green or yellow, indicating whether these compounds were present (green) or absent (yellow) in experiments. Lines indicating higher confidence thresholds of 50% WISE consensus score (in red) and 2.5 PLM score (in blue) are drawn.

**CASE STUDY 2: Surfactin B**

Surfactins are cyclic lipopeptide biosurfactants produced by a biosynthetic gene cluster containing multiple non-ribosomal peptide synthetases[34]. Its primary mechanism involves reducing surface and interfacial tension and the disruption of cellular membranes, which contribute to its strong antimicrobial and antiviral activities. Since its discovery in the 1980s, surfactins have garnered interest for their amphiphilic structure and a broad spectrum of biological activities, including antibacterial, antifungal, antiviral, and antitumor effects. The most common producers of surfactins are *Bacillus* species, making these microorganisms primary sources for its commercial and research-based production.

The cheminformatics-driven analysis predicted only six strains (out of 54 strains) as producers of surfactin B with WISE scores exceeding 35%, whilst embedding-based protein alignment pointed out 35 strains with putative surfactin pathways (PLM scores >0.75). At the intersection of chemical and genomic-based prediction, 5 strains were in consensus and proposed as high-confidence leads as producers of surfactin B (**Figure 4D** – **coloured box**). Experimental validation[22], including NMR analyses[13], showed that surfactin B was indeed found in the five predicted strains, of which two strains (A1123, T1416) only produced surfactin B in their activated forms[19] (Tables S6, S10). There is also a ten-fold difference in production levels of surfactin B among these strains. Here, we also noted that these are new genus for surfactin B producers, where A1123 and A80510 are *Thermoactinomyces*, whilst the other three strains are putative new *Streptomyces* producers (**Table S8**). M*icromonospora*[35] has previously been identified as a producer of surfactin. One strain, known as *Streptomyces tendae*, has also been previously recognized as a surfactant producer, though surfactin was not previously identified in the mixture of surfactants observed[36].

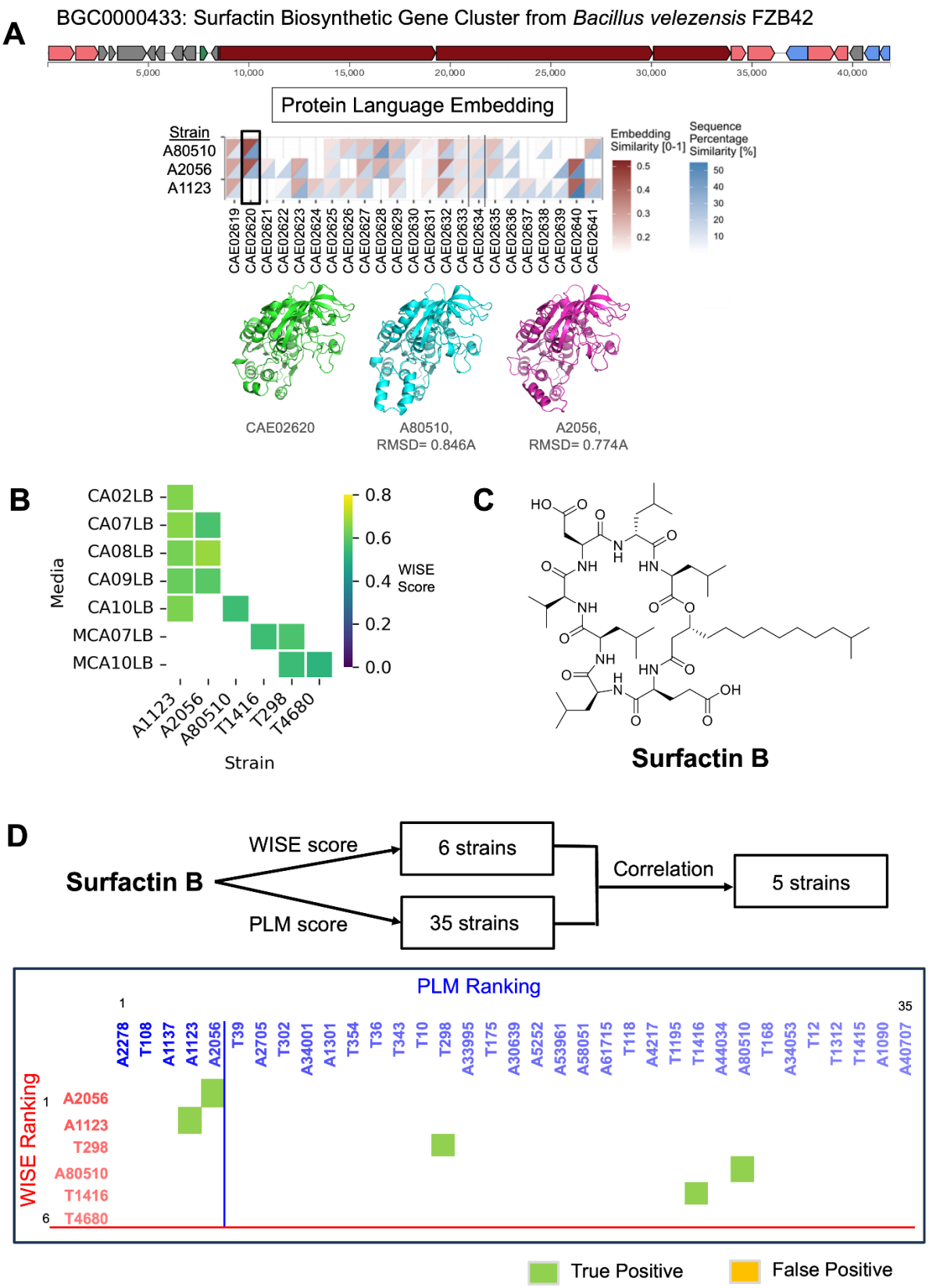


**Figure 4. Case Study: Identification of surfactin B producers.** (A) The biosynthetic gene cluster BGC0000433, responsible for surfactin production, was analysed by

embedding each protein component using the ESM2 protein language model. These embeddings were then queried against a precomputed vector database of protein embeddings derived from the full dataset. Nearest neighbour algorithms were employed to identify the top-scoring matches, with the results presented in the left inset. The sequences corresponding to the highest-ranked embeddings were retrieved, and pairwise sequence alignments confirmed that the embeddings captured the underlying sequence similarity. Additionally, an alcohol dehydrogenase from the cluster (CAE02620.1) was selected for *in silico* structural prediction using ESMfold. The predicted structure was aligned with known structural data, revealing a high degree of structural similarity, further validating the model's utility in capturing both sequence and structural relationships. (B) Heatmap of the WISE scores of 6 strains-media combination derived from the liquid chromatography-tandem mass spectrometry (LC-MS/MS) data of fermentation extracts. Only strains with at least one media combination that has a non-zero WISE score are displayed. Fermentation extracts from strain-media combinations that do not have hits for surfactin B are indicated by blanks. Activated strains are referred to by their parent strain instead of individually (details of the activated strains are in **Tables S3, S10**). (C) Structure of surfactin B. (D) Workflow of the co-occurrence strategy and matrix ranking table of the top-ranked strains processed by cheminformatics (WISE) and bioinformatics (PLM) with at least 35% WISE consensus score or 0.75 PLM score, respectively. Strains that have been cross-validated are highlighted in green or yellow, indicating whether these compounds were present (green) or absent (yellow) in experiments. Lines indicating higher confidence thresholds of >50% WISE consensus score (in red) and 2.5 PLM score (in blue) are drawn.

**CASE STUDY 3: Neomycin B**

Neomycin B is an aminoglycoside antibiotic produced by the *Streptomyces fradiae* species, encoded by a biosynthetic gene cluster (miBiG: BGC0001085). This antibiotic binds to the bacterial 30S ribosomal subunit[37], disrupting protein synthesis and leading to bacterial cell death. Since its discovery in 1949, neomycin B has remained a staple in the field of antibacterial agents, widely used for topical applications and as a treatment for infections caused by gram-negative bacteria[38]. While it is predominantly known for its antimicrobial properties, neomycin B has also recently demonstrated some antiviral activity[39]. One challenge with neomycin B is scaling its production for increasing interest and applications, including low yields in its native producer.

The cheminformatics-driven analysis predicted only six strains (out of 54 strains) as potential producers of neomycin B with WISE scores exceeding 35%, whilst embedding-based protein alignments predicted 11 strains as containing neomycin biosynthetic genes or proteins with PLM scores exceeding 0.75. Using integrated genomic and chemical screening methodologies, we prioritised four strains that exhibited biosynthetic potential for producing neomycin B. Experimental validation confirmed the presence of neomycin B in three strains, two of which (T343, T1236) produced neomycin B in their activated forms. One of these strains (T343) has been identified as *Streptomyces albogriseolus*, previously described as a neomycin producer. However, the other two *Streptomyces* species (T354, T1236) are newly identified producers of neomycin. In this library, we also have strains (T298, T265) related to *Streptomyces fradiae*, known producers of neomycin. Under the parameters of this study, although there were genome hits of high confidence from T298, neomycin was not ranked highly based on WISE, indicating the importance of data

quality to prevent missed identification. In this case, the lower production yields in T298 and T265 (**Table S11**) could have contributed to the low data quality of the mass spectra.

It should be noted that although GNPS contains a reference for neomycin B, it was not identified in their automated processing of our raw datasets. Instead, additional experimentation was required to confirm the presence of neomycin B (see Methods for validation), highlighting the limitations of GNPS's spectral matching algorithm in detecting certain compounds. This limitation underscores broader challenges in GNPS processing, which identified only 743 unique compounds from the curated LC-MS/MS dataset, compared to the 29,112 unique compounds identified through WISE. The significant disparity may stem from differences in the underpinning algorithms employed by GNPS and WISE. While GNPS may group related spectra by their MS/MS similarity, WISE performs individual structural identification for each spectra, potentially yielding a higher number of unique annotations.

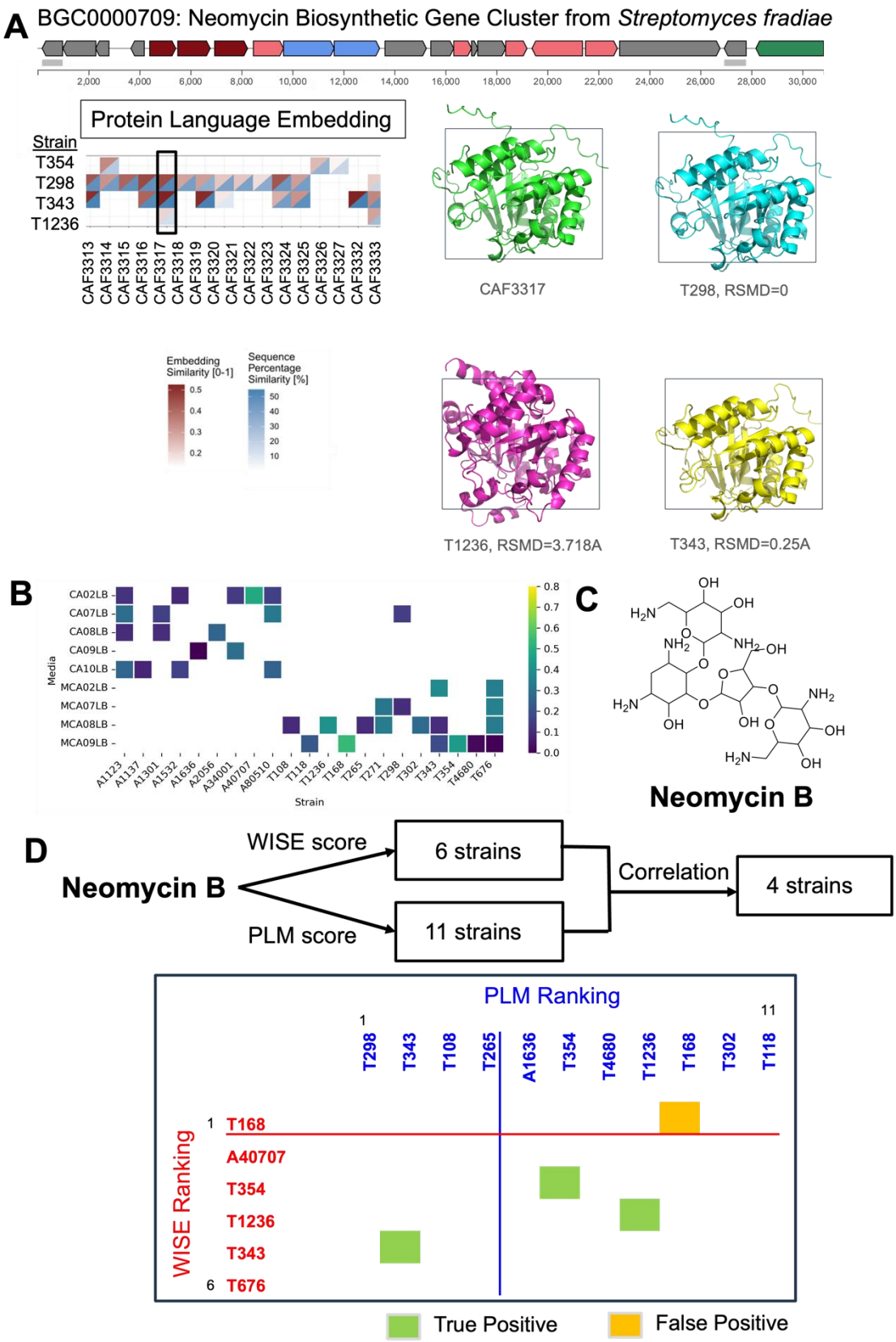


**Figure 5. Case Study: Neomycin B Screening.** (A) The biosynthetic gene cluster BGC00007089, responsible for neomycin production, was analysed by embedding

each protein component using the ESM2 protein language model. These embeddings were then queried against a precomputed vector database of protein embeddings derived from the full dataset. Nearest neighbour algorithms were employed to identify the top-scoring matches, with the results presented in the left inset. For this case, the sequences corresponding to the highest-ranked embeddings were retrieved, and pairwise sequence alignments confirmed that the embeddings captured the underlying sequence similarity. Additionally, an enzyme from the cluster (Fe-S oxidoreductase, CAF33317) was selected for in silico structural prediction using ESM. The predicted structure was aligned with known structural data, revealing a high degree of structural similarity, further validating the model's utility in capturing both sequence and structural relationships. (B) Heatmap of the WISE scores of 21 strains-media combinations derived from the liquid chromatography-tandem mass spectrometry (LC-MS/MS) data of fermentation extracts. Only strains with at least 1 media combination with non-zero WISE score are shown. Fermentation extracts from strain-media combinations that do not have hits for Neomycin B are indicated by blanks. Activated strains (if any) are referred to by their parent strain instead of individually (details of the activated strains are in **Tables S7, S11**). (C) Structure of neomycin B. (D) Workflow of the co-occurrence strategy and matrix ranking table of the top-ranked strains processed by cheminformatics (WISE) and bioinformatics (PLM) with at least 35% WISE consensus score or 0.75 PLM score, respectively. Strains that have been cross-validated are highlighted in green or yellow, indicating whether these compounds were present (green) or absent (yellow) in experiments. Lines indicating higher confidence thresholds of 50% WISE consensus score (in red) and 2.5 PLM score (in blue) are drawn.

## DISCUSSION

Our study demonstrates a transformative approach to compound and strain discovery that leverages genomic and chemical data integration through a multi-modal strategy, combining the strengths of protein language models, molecular language processing, and transformer-based models for compound identification. This methodology offers opportunities for substantial improvements in several longstanding challenges in natural product research.

### Advancements in genomic and MS analyses

Traditional methods for predicting biosynthetic gene clusters (BGCs), such as antiSMASH, primarily rely on the sequence homology of clustered genes. However, a significant challenge with these clustering approaches is their sensitivity to the quality of genomic assemblies. BGCs often span large genomic regions, and if the genomes are fragmented or incomplete, the accuracy of predictions can be compromised. Additionally, obtaining high-quality genomes for an extensive collection of GC rich strains can be both time-consuming and costly. On the other hand, traditional BLAST searches for individual genes are not scalable. To address these challenges, we explored using protein-level embeddings generated by the ESM2 language model. Unlike sequence homology, which emphasises conserved regions, these embeddings represent proteins as high-dimensional vectors, capturing structural and functional nuances that can extend beyond sequence similarity. This capability allows for more refined comparisons between proteins and facilitates the identification of biological functions that may elude homology-based methods. Additionally, protein-level embeddings are less dependent on genomic contiguity, requiring only accurate protein

annotations. This robustness makes them particularly advantageous for lower-quality genomics datasets, though some quality is still recommended.

Simultaneously, generative molecular language processing is pivotal in expanding the chemical space. By predicting the spectral references of generated novel molecules, we significantly broaden the range of compounds that can be explored. This synergistic integration of generative molecular language processing with structure-to-MS/MS predictions facilitates the expansion of reference databases into previously uncharted territories. This capability addresses the limitation of sparse spectral references in current databases, providing a robust tool for identifying novel compounds that may otherwise be overlooked. When integrated with the deep deconvolutive processing workflow, this suite of technologies transforms the analysis of large LC-MS/MS data into a scalable discovery process. Rather than focusing solely on the definitive identification of known compounds, this approach generates hypotheses guided by predictive models. These informed hypotheses offer valuable insights to the molecular structures corresponding to the mass spectral data that would have otherwise been left unannotated.

**Co-occurrence strategy – finding biological significance**

Individually, these genomic and chemical analytic advances can overcome limitations for rapid identification of patterns in their respective datasets (such as gene annotations or structural prediction). However, there remains a challenge in ensuring that these uncovered patterns are biologically relevant. On this note, data integration becomes important to pinpoint biological significance. The co-occurrence strategy we employed represents a paradigm shift in using linked genomic and chemical data. By

simultaneous screening of encoded proteins and predicted molecular structures through vector search, we created a framework for identifying producer strains that not only possess genomic features associated with the target compound but also possess relevant chemical profiles with high precision. We believe the integration of these multi-modal data sets facilitates a more comprehensive discovery process. When strains with matching genomic and chemical profiles are identified, and subjected to further validation, our case studies showed that the predictions are grounded in empirical evidence. This approach not only accelerates the discovery process but also improves the reliability of the results.

**Challenges and Future Directions**

As early adopters of integrating generative chemical and protein language advancements with multidisciplinary experimental datasets, we have observed significant potential and opportunities for accelerating discovery and hypothesis generation. However, our approach is not without challenges of data quality and model interpretability. In this study, predictions based on PLM or WISE scores alone grossly over-estimate the number of putative producer strains for each of the molecules studied. By applying a co-occurrence alignment with confidence cut off, we were able to greatly increase the hit rates of identifying putative producers, 75-100% with relatively few experiments. For examples, for valinomycin, 33 unique strains were predicted to be producers based on either PLM or WISE score alone whereas only 14 strains were prioritized based on aligned PLM and WISE scores and of which 12 strains were experimentally validated. For surfactin B, a whopping 36 unique strains were predicted to be producers based on either PLM or WISE score alone whereas only 5 strains were prioritized based on aligned PLM and WISE scores, all of which

were experimentally validated. Similarly, for neomycin B, 13 unique strains were predicted to be producers based on either PLM or WISE score alone whereas only 4 strains were prioritized based on aligned PLM and WISE scores, and of which 3 strains were experimentally validated.

By applying a confidence cut off, of course, there are limitations of potentially missing out on some of the producers which may not score highly due to poor data quality or cryptic clusters. In the case studies, the missed producers are 9/21 for valinomycin, 1/6 for surfactin B and 4/7 for neomycin B should all 54 strains in the library were individually studied. Like with all prediction methodologies, there remain limitations of false positives (2/14 for valinomycin and 1/4 for neomycin B) despite consensus in PLM and WISE scores alignment, possibly due to data quality and inherent bias in the processing algorithm. The complexity of interpreting high-dimensional data and the potential for errors in vector-based searches underscore the need for continued refinement of our algorithms and methodologies.

Future research will focus on enhancing the precision of our generative models and improving the integration of diverse data sources. Exploring the application of advanced deconvolution techniques for complex LC-MS/MS data and expanding our chemical database through continuous AI-driven predictions will be key to further advancing the field. In our studies, we also noticed that validation of predicted producers is frequently only made possible due to strain activation *via* overexpressing secondary metabolite regulators. With increasing utilisation of the current genomic driven discovery approaches, we are also starting to see disconnect between *in silico* prediction and actual production of the proposed compound. Often, heterologous

expression strategies are necessary to validate the chemical-biology predictions. Here, our studies suggest that by integrating gene cluster activation or metabolic regulation[8,40] into native strain libraries, our presented mining strategy could accelerate the identification and probability of finding a productive producer. Simultaneously, the advantage of being able to mine beyond curated chemical space of a microbial library also allows for much faster validation.

## MATERIAL AND METHODS

### Genomics sequencing dataset of the strain library

Multiplex sequencing of the 54 native strains was performed with Illumina. The short reads were subjected to quality assessment and *de novo* assembly (see Methods). The assemblies had a median length of 7.5Mb and N50 of 7.2kb. Clustering of BGCs into gene cluster families (GCFs) was performed using BiG-SCAPE[41], a tool that takes BGC sequences as input to build an interactive sequence similarity network. To facilitate a thorough comparison, we combined BGCs from the 124 strains with the latest available experimentally validated BGC repertoire MIBiG 3.0[42].

### Structure-to-MS/MS

Tandem mass spectra (MS/MS) of molecular structures were predicted using a transformer-based model adapted from ICEBERG[13] that takes as inputs: (1) molecular structure in SMILES format, (2) adduct type and (3) collision energy in eV, to simulate their corresponding MS/MS spectral reference. Spectral references were generated for both the COCONUT[28] and generated 67 million natural product[25] databases for 4 adduct types – $[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, $[M+NH_4]^+$ and a collision energy range from 1 – 129 eV to create an augmented MS/MS spectral reference database for structural elucidation.

### Workflow for Intelligent Structural Elucidation (WISE)

After processing and deconvoluting MS/MS spectra of individual compounds, the augmented MS/MS spectral reference database is shortlisted by filtering at ±0.005 *m/z* around the precursor *m/z* of the identified compound *m/z*. Consensus scores (WISE

scores) are then calculated between the identified compound and every shortlisted reference according to the following formula:

$$Consensus\ (WISE)\ Score = 0.5 \cdot (ms\ match) + 0.2 \cdot (iso\ match) + 0.3 \cdot (mass\ match)$$

Where,

*ms match* = weighted cosine similarity between experimental MS/MS and generated MS/MS reference

*iso match* = moderated cosine similarity between experimental MS1 isotopic distribution and calculated isotopic distribution

*mass match* = similarity scoring between experimental precursor *m/z* and calculated monoisotopic mass

Formulas for the individual match scores as follows:

$$\mathrm{ms\ match} = \frac{\sum_{k=1}^{M_{max}} m_k \sqrt{I_{qk}} \cdot m_k \sqrt{I_{ik}}}{\sqrt{\sum_{k=1}^{M_q} \left(m_k \sqrt{I_{qk}}\right)^2} \sqrt{\sum_{k=1}^{M_I} \left(m_k \sqrt{I_{ik}}\right)^2}}$$

$I_q$ and $I_i$ are vectors of *m/z* intensities representing the experimental spectrum and the predicted spectrum respectively. $m_k$ and $I_k$ are the *m/z* ratio and intensity found at *m/z* = k; $M_q$ and $M_i$ are the largest indices of $I_q$ and $I_i$ with non-zero values. $M_{max}$ is the larger of $M_q$ and $M_i$. The cosine similarity is weighted by *m/z* as the peaks in mass spectra corresponding to larger fragments are more characteristic and useful in practice for identifying[43].

$$iso\ match = \max\left(\left(\frac{\sum_i norm_exp_i \cdot norm_calc_i}{\sqrt{(\sum_i norm_exp_i^2) \cdot (\sum_i norm_calc_i^2)}} - 0.999\right) \cdot 1000, 0\right)$$

Norm_exp$_i$ and norm_calc$_i$ are MS1 intensities of the isotopic distributions divided by the maximum MS1 intensity in their distribution of the experimental and calculated isotopic distributions respectively. As the cosine similarity score of the isotopic distribution is inherently high due to the identical *m/z* values (M, M+1, M+2, M+3...etc) and minute differences between abundance distributions due to low natural abundance of heavier isotopes, the score is moderated to only consider similarities beyond 99.9% (i.e. a cosine similarity of 99.9882% is moderated to 88.2% instead). Anything below the 99.9% threshold is taken as 0%.

$$mass\ match = \frac{\max\left(0.005 - abs(mass_\exp - mass_calc), 0\right)}{0.005}$$

Mass_exp and mass_calc are the precursor *m/z* of the experimental compound observed and the calculated precursor *m/z* of the reference compound respectively. The previous mass filter of 0.005 *m/z* is taken here as the maximum possible deviation from the experimental precursor *m/z* (i.e. 0.005 *m/z* difference is 0% mass match).

The references are then ranked by their WISE scores and the top-ranking candidate is predicted as the molecular structure.

**Metabolomic characterization of the strain library**

Using the previously published 2,138 LC-MS/MS data[19], the dataset were re-processed using WISE that combines data processing, deconvolution, and structural elucidation in an end-to-end automated processing pipeline to detect metabolites present based on their MS1 signal intensities exceeding a threshold of 10,000 in their base peak chromatogram (BPC). The top 10 predicted candidate molecular structures per metabolite identified were taken to form a pool of metabolites from which searches were done to identify producers of valinomycin, surfactin B, and neomycin B.

**Case study experimental validation**

The presence of valinomycin and surfactin B in microbial strain fermentation extracts were validated via isolation and structural confirmation by NMR[19], as well as spectral matching of experimental LC-MS/MS data against the Global Natural Products Social Molecular Networking (GNPS)[23] spectral reference libraries[22]. The library spectra were filtered by removing all MS/MS fragment ions within +/- 17 Da of the precursor *m/z*. MS/MS spectra were window filtered by choosing only the top 6 fragment ions in the +/- 50Da window throughout the spectrum. The precursor ion mass tolerance was set to 0.02 Da and a MS/MS fragment ion tolerance of 0.02 Da. All matches kept between queried experimental spectra and reference library spectra were required to have a cosine similarity score above 0.7 and at least 6 matched peaks.

The presence of Neomycin B in microbial strain fermentation extracts was validated by spectral matching between experimental LC-MS/MS data against an in-house experimental reference MS/MS obtained from analyzing a commercial standard of Neomycin B, as well as the GNPS reference of Neomycin B (CCMSLIB00010114511). The dataset was filtered to those where WISE annotated it as a $[M+Na]^+$ adduct, had a precursor *m/z* of ±0.005 *m/z* of 637.3015062321 (the calculated theoretical monoisotopic mass of the $[M+Na]^+$ adduct of Neomycin B), had a retention time between 6.94 – 7.14 min. The short list was further filtered to keep only matches where the queried experimental spectra had on average at least 4 matched peaks that are at least 5% of base peak abundance and ±5 Da away from each other to both the commercial and GNPS reference spectra. Abundance threshold was set so that

matched peaks must be significant, and Da separation set so that isotopes of the same ion were not considered.

**Similarity searches Using kNN in protein language model embeddings**

We utilized the ESM2 protein language model to generate embeddings for our protein library. The embeddings were generated following the procedure outlined on the ESM2 GitHub repository (https://github.com/facebookresearch/esm). For each assembled genome, the inferred sequences obtained through AntiSMASH were processed to generate embeddings using this approach. The resulting vectors were then indexed and stored in an OpenSearch database, along with additional metadata such as strain name and the environment from which the samples were collected.

To perform similarity searches, we implemented a k-nearest neighbors (kNN) algorithm using the embedded vector descriptors stored in the database. For each query, the top 20 closest matches were retrieved based on the similarity of their embeddings. For example, in cases where a known enzyme was associated with a specific class of chemical compounds in the MIBiG database, we used the enzyme sequence as the query to identify and return the top 20 hits that shared similar embeddings.

## REFERENCES


1. Koehn, F. E.; Carter, G. T., The evolving role of natural products in drug discovery. *Nat. Rev. Drug Discovery* **2005,** *4* (3), 206-220.
2. Hobson, C.; Chan, A. N.; Wright, G. D., The Antibiotic Resistome: A Guide for the Discovery of Natural Products as Antimicrobial Agents. *Chem. Rev.* **2021,** *121* (6), 3464-3494.
3. Dong, S.; Guo, X.; Han, F.; He, Z.; Wang, Y., Emerging role of natural products in cancer immunotherapy. *Acta Pharmaceutica Sinica B* **2022,** *12* (3), 1163-1185.
4. Naeem, A.; Hu, P.; Yang, M.; Zhang, J.; Liu, Y.; Zhu, W.; Zheng, Q., Natural Products as Anticancer Agents: Current Status and Future Perspectives. *Molecules* **2022,** *27* (23), 8367.
5. Kenshole, E.; Herisse, M.; Michael, M.; Pidot, S. J., Natural product discovery through microbial genome mining. *Curr. Opin. Chem. Biol.* **2021,** *60*, 47-54.
6. Bauman, K. D.; Butler, K. S.; Moore, B. S.; Chekan, J. R., Genome mining methods to discover bioactive natural products. *Nat. Prod. Rep.* **2021,** *38* (11), 2100-2129.
7. Jarmusch, S. A.; van der Hooft, J. J. J.; Dorrestein, P. C.; Jarmusch, A. K., Advancements in capturing and mining mass spectrometry data are transforming natural products research. *Nat. Prod. Rep.* **2021,** *38* (11), 2066-2082.
8. Blin, K.; Shaw, S.; Augustijn, H. E.; Reitz, Z. L.; Biermann, F.; Alanjary, M.; Fetter, A.; Terlouw, B. R.; Metcalf, W. W.; Helfrich, E. J. N.; van Wezel, G. P.; Medema, M. H.; Weber, T., antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **2023,** *51* (W1), W46-W50.
9. Skinnider, M. A.; Johnston, C. W.; Gunabalasingam, M.; Merwin, N. J.; Kieliszek, A. M.; MacLellan, R. J.; Li, H.; Ranieri, M. R. M.; Webster, A. L. H.; Cao, M. P. T.; Pfeifle, A.; Spencer, N.; To, Q. H.; Wallace, D. P.; Dejong, C. A.; Magarvey, N. A., Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **2020,** *11* (1), 6058.
10. Hannigan, G. D.; Prihoda, D.; Palicka, A.; Soukup, J.; Klempir, O.; Rampula, L.; Durcak, J.; Wurst, M.; Kotowski, J.; Chang, D.; Wang, R.; Piizzi, G.; Temesi, G.; Hazuda, D. J.; Woelk, C. H.; Bitton, D. A., A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **2019,** *47* (18), e110-e110.
11. Rios-Martinez, C.; Bhattacharya, N.; Amini, A. P.; Crawford, L.; Yang, K. K., Deep self-supervised learning for biosynthetic gene cluster detection and product classification. *bioRxiv* **2022**, 2022.07.22.500861.
12. Sindelar, M.; Patti, G. J., Chemical Discovery in the Era of Metabolomics. *J. Am. Chem. Soc.* **2020,** *142* (20), 9097-9105.
13. Goldman, S.; Li, J.; Coley, C. W., Generating Molecular Fragmentation Graphs with Autoregressive Neural Networks. *Anal. Chem.* **2024,** *96* (8), 3419-3428.
14. Wang, F.; Liigand, J.; Tian, S.; Arndt, D.; Greiner, R.; Wishart, D. S., CFM-ID 4.0: More Accurate ESI-MS/MS Spectral Prediction and Compound Identification. *Anal. Chem.* **2021,** *93* (34), 11692-11700.
15. Huber, F.; van der Burg, S.; van der Hooft, J. J. J.; Ridder, L., MS2DeepScore: a novel deep learning similarity measure to compare tandem mass spectra. *J. Cheminform.* **2021,** *13* (1), 84.
16. Meunier, M.; Bréard, D.; Boisard, S.; Blanchard, P.; Litaudon, M.; Awang, K.; Schinkovitz, A.; Derbré, S., Looking for Actives in the Haystack: Merging HRMS2-Based Molecular Networking, Chemometrics, and 13C NMR-Based Dereplication Approaches. *J. Nat. Prod.* **2024,** *87* (10), 2398-2407.

17. Collins, S. L.; Koo, I.; Peters, J. M.; Smith, P. B.; Patterson, A. D., Current Challenges and Recent Developments in Mass Spectrometry–Based Metabolomics. *Annual Review of Analytical Chemistry* **2021,** *14* (Volume 14, 2021), 467-487.
18. da Silva, R. R.; Dorrestein, P. C.; Quinn, R. A., Illuminating the dark matter in metabolomics. *PNAS* **2015,** *112* (41), 12549-12550.
19. Tay, D. W. P.; Tan, L. L.; Heng, E.; Zulkarnain, N.; Ching, K. C.; Wibowo, M.; Chin, E. J.; Tan, Z. Y. Q.; Leong, C. Y.; Ng, V. W. P.; Yang, L. K.; Seow, D. C. S.; Lim, Y. W.; Koh, W.; Koduru, L.; Kanagasundaram, Y.; Ng, S. B.; Lim, Y. H.; Wong, F. T., Exploring a general multi-pronged activation strategy for natural product discovery in Actinomycetes. *Commun. Biol.* **2024,** *7* (1), 50.
20. Verkuil, R.; Kabeli, O.; Du, Y.; Wicky, B. I. M.; Milles, L. F.; Dauparas, J.; Baker, D.; Ovchinnikov, S.; Sercu, T.; Rives, A., Language models generalize beyond natural proteins. *bioRxiv* **2022**, 2022.12.21.521521.
21. Ng, S. B.; Kanagasundaram, Y.; Fan, H.; Arumugam, P.; Eisenhaber, B.; Eisenhaber, F., The 160K Natural Organism Library, a unique resource for natural products research. *Nat. Biotechnol.* **2018,** *36* (7), 570-573.
22. Tay, D. W. P.; Tan, L. L.; Heng, E.; Zulkarnain, N.; Chin, E. J.; Tan, Z. Y. Q.; Leong, C. Y.; Ng, V. W. P.; Yang, L. K.; Seow, D. C. S.; Kanagasundaram, Y.; Ng, S. B.; Lim, Y. H.; Wong, F. T., Tandem mass spectral metabolic profiling of 54 actinobacterial strains and their 459 mutants. *Sci. Data* **2024,** *11* (1), 977.
23. 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., Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **2016,** *34* (8), 828-837.
24. Wei, J. N.; Belanger, D.; Adams, R. P.; Sculley, D., Rapid Prediction of Electron–Ionization Mass Spectrometry Using Neural Networks. *ACS Cent. Sci.* **2019,** *5* (4), 700-708.
25. Tay, D. W. P.; Yeo, N. Z. X.; Adaikkappan, K.; Lim, Y. H.; Ang, S. J., 67 million natural product-like compound database generated via molecular language processing. *Sci. Data* **2023,** *10* (1), 296.

26. Tay, D. W. P.; Ang, S. J.; Lim, Y. H.; Wong, Z. M. Generation of Collision Energy Specific Tandem Mass Spectra from Molecular Structures. SG Patent Application No. 10202400652R, 2024.
27. Tay, D. W. P.; Adaikkappan, K.; Yeo, N. Z. X.; Ang, S. J.; Lim, Y. H. Automated Workflow for Intelligent Structural Elucidation (WISE) from Raw Liquid Chromatography Tandem Mass Spectrometry Data. SG Patent Application No. 10202403057W, 2024.
28. Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT online: Collection of Open Natural Products database *J. Cheminform.* [Online], 2021, p. 2. https://doi.org/10.1186/s13321-020-00478-9 (accessed 2021/01/10).
29. Zhang, C.; Freddolino, L., A large-scale assessment of sequence database search tools for homology-based protein function prediction. *Brief. Bioinformatics* **2024,** *25* (4).
30. Liu, W.; Wang, Z.; You, R.; Xie, C.; Wei, H.; Xiong, Y.; Yang, J.; Zhu, S., PLMSearch: Protein language model powers accurate and fast sequence search for remote homology. *Nat. Commun.* **2024,** *15* (1), 2775.
31. Johnson, S. R.; Peshwa, M.; Sun, Z., Sensitive remote homology search by local alignment of small positional embeddings from protein language models. *eLife* **2024,** *12*, RP91415.
32. Huang, S.; Liu, Y.; Liu, W.-Q.; Neubauer, P.; Li, J., The Nonribosomal Peptide Valinomycin: From Discovery to Bioactivity and Biosynthesis. *Microorganisms* **2021,** *9* (4), 780.
33. Zhang, D.; Ma, Z.; Chen, H.; Lu, Y.; Chen, X., Valinomycin as a potential antiviral agent against coronaviruses: A review. *Biomedical Journal* **2020,** *43* (5), 414-423.
34. Théatre, A.; Cano-Prieto, C.; Bartolini, M.; Laurin, Y.; Deleu, M.; Niehren, J.; Fida, T.; Gerbinet, S.; Alanjary, M.; Medema, M. H.; Léonard, A.; Lins, L.; Arabolaza, A.; Gramajo, H.; Gross, H.; Jacques, P., The Surfactin-Like Lipopeptides From Bacillus spp.: Natural Biodiversity and Synthetic Biology for a Broader Application Range. *Front. Bioeng. Biotechnol.* **2021,** *9*.
35. Pang, X.; Zhao, J.; Fang, X.; Liu, H.; Zhang, Y.; Cen, S.; Yu, L., Surfactin derivatives from Micromonospora sp. CPCC 202787 and their anti-HIV activities. *J. Antibiot.* **2017,** *70* (1), 105-108.
36. Richter, M.; Willey, J. M.; Süßmuth, R.; Jung, G.; Fiedler, H.-P., Streptofactin, a novel biosurfactant with aerial mycelium inducing activity from Streptomyces tendae Tü 901/8c. *FEMS Microbiol. Lett.* **1998,** *163* (2), 165-171.
37. Sutcliffe, J. A., Improving on nature: antibiotics that target the ribosome. *Current Opinion in Microbiology* **2005,** *8* (5), 534-542.
38. MACDONALD, R. H.; BECK, M., Neomycin: a review with particular reference to dermatological usage. *Clinical and Experimental Dermatology* **1983,** *8* (3), 249-258.
39. Abdulaziz, L.; Elhadi, E.; Abdallah, E. A.; Alnoor, F. A.; Yousef, B. A., Antiviral Activity of Approved Antibacterial, Antifungal, Antiprotozoal and Anthelmintic Drugs: Chances for Drug Repurposing for Antiviral Drug Discovery. *Journal of Experimental Pharmacology* **2022,** *14* (null), 97-115.
40. Zhou, Q.; Ning, S.; Luo, Y., Coordinated regulation for nature products discovery and overproduction in Streptomyces. *Synth. Syst. Biotechnol.* **2020,** *5* (2), 49-58.
41. Navarro-Muñoz, J. C.; Selem-Mojica, N.; Mullowney, M. W.; Kautsar, S. A.; Tryon, J. H.; Parkinson, E. I.; De Los Santos, E. L. C.; Yeong, M.; Cruz-Morales, P.; Abubucker, S.; Roeters, A.; Lokhorst, W.; Fernandez-Guerra, A.; Cappelini, L. T. D.; Goering, A. W.; Thomson, R. J.; Metcalf, W. W.; Kelleher, N. L.; Barona-Gomez, F.; Medema, M. H., A computational framework to explore large-scale biosynthetic diversity. *Nat. Chem. Biol.* **2020,** *16* (1), 60-68.

42. Terlouw, B. R.; Blin, K.; Navarro-Muñoz, J. C.; Avalon, N. E.; Chevrette, M. G.; Egbert, S.; Lee, S.; Meijer, D.; Recchia, Michael J. J.; Reitz, Zachary L.; van Santen, Jeffrey A.; Selem-Mojica, N.; Tørring, T.; Zaroubi, L.; Alanjary, M.; Aleti, G.; Aguilar, C.; Al-Salihi, Suhad A. A.; Augustijn, Hannah E.; Avelar-Rivas, J A.; Avitia-Domínguez, Luis A.; Barona-Gómez, F.; Bernaldo-Agüero, J.; Bielinski, V. A.; Biermann, F.; Booth, Thomas J.; Carrion Bravo, Victor J.; Castelo-Branco, R.; Chagas, Fernanda O.; Cruz-Morales, P.; Du, C.; Duncan, Katherine R.; Gavriilidou, A.; Gayrard, D.; Gutiérrez-García, K.; Haslinger, K.; Helfrich, Eric J. N.; van der Hooft, Justin J. J.; Jati, Afif P.; Kalkreuter, E.; Kalyvas, N.; Kang, Kyo B.; Kautsar, S.; Kim, W.; Kunjapur, Aditya M.; Li, Y.-X.; Lin, G.-M.; Loureiro, C.; Louwen, Joris J. R.; Louwen, Nico L. L.; Lund, G.; Parra, J.; Philmus, B.; Pourmohsenin, B.; Pronk, Lotte J. U.; Rego, A.; Rex, Devasahayam Arokia B.; Robinson, S.; Rosas-Becerra, L R.; Roxborough, Eve T.; Schorn, Michelle A.; Scobie, Darren J.; Singh, Kumar S.; Sokolova, N.; Tang, X.; Udwary, D.; Vigneshwari, A.; Vind, K.; Vromans, Sophie P. J. M.; Waschulin, V.; Williams, Sam E.; Winter, Jaclyn M.; Witte, Thomas E.; Xie, H.; Yang, D.; Yu, J.; Zdouc, M.; Zhong, Z.; Collemare, J.; Linington, Roger G.; Weber, T.; Medema, Marnix H., MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.* **2022,** *51* (D1), D603-D610.

43. Nicolescu, T. O., Interpretation of mass spectra. *Mass Spectrometry* **2017**, 23-78.

**ACKNOWLEGEMENTS**

The authors gratefully acknowledge financial support from the National Research Foundation, Singapore (NRF-CRP19-2017-05-00). This work is also supported by the Agency for Science, Technology and Research (A*STAR, Singapore) under C211917003, C211917006, C233017006, and Singapore Integrative Biosystems and Engineering Research Strategic Research & Translational Thrust (SIBER SRTT). We gratefully acknowledge the Biocomputing Centre at the Bioinformatics Institute of Singapore for their crucial role in hosting and managing the genomics datasets that were fundamental to this research. We also thank the compute support team at the A*STAR Computational Resource Centre (A*CRC) for providing the robust technical infrastructure and expertise necessary for data hosting and seamless computational workflows.