npj | antimicrobials & resistance

Article

https://doi.org/10.1038/s44259-026-00206-7

Accelerating natural product discovery with linked MS-genomics and language/transformer-based models

Dillon W. P. Tay[1,7] ✉, Winston Koh[1,2,7] ✉, Shi Jun Ang[3,4], Zicong Marvin Wong[3], Yi Wee Lim[1], Elena Heng[5], Yu Hung Ng[1], Naythan Z. X. Yeo[1], Krishnan Adaikkappan[1], Fong Tian Wong[1,5] ✉ & Yee Hwee Lim[1,6] ✉

Integrated chem-bio characterization of microbial strain libraries can streamline natural product discovery by prioritizing candidate producers. Here, we employ language- and transformer-based models to extract actionable insights from linked mass spectrometry (MS)-genome datasets. Our framework enables ranking of microbial producers to prioritise high-potential candidates for targeted validation. Across three representative case studies, this approach prioritized producers of diverse natural products with 75–100% precision. These findings demonstrate the transformative potential of AI-enabled chem-bio characterization to significantly accelerate natural product discovery and enable access to microbial chemical diversity beyond reference knowledge.

Natural products (NPs) offer exceptional chemical diversity and therapeutic potential[1]. However, their discovery is hindered by inefficiencies in linking genomic information to structural validation[2]. Despite advancements in genomic and analytical technologies such as mass spectrometry (MS), existing approaches often rely heavily on predefined rules and reference libraries[3,4]. Consequently, most structural assignments remain confined to previously characterized metabolites[5], leaving large portions of natural chemical space underexplored. To address these challenges, we introduce a multi-modal prioritization framework that integrates complementary genomic and metabolomic inference layers to decouple NP discovery from reference libraries, enabling rapid navigation of uncharted microbial chemical space (Fig. 1).

At the genomics layer, protein language models (PLMs) are used to represent individual proteins as high-dimensional embeddings that capture structural and functional relationships beyond primary sequence homology. This enables scalable similarity searches that can detect remote functional analogues and assess biosynthetic potential even in fragmented or incomplete genome assemblies. As a result, the genomic inference layer provides a protein-level view of latent biosynthetic capability (Fig. 1, bottom-middle box on genomic mining), independent of whether a pathway is expressed under the tested conditions.

In parallel, our metabolomics pipeline employs our in-house developed Workflow for Intelligent Structural Elucidation (WISE) process that combines generative chemistry and transformer-based models to generate putative molecular structures from mass spectrometry (MS) data. By leveraging generative molecular language processing models[6] trained on large collections of natural products[7], we extend the search space beyond existing reference libraries, enabling hypothesis generation for previously uncharacterized compounds (Fig. 1, top-middle box on compound identification). This approach provides a scalable, structure-centric interpretation of MS data without reliance on spectral reference databases.

When integrated, the genomic and metabolomic inference layers provide complementary genotype-phenotype evidence that enables more confident prioritization of candidate producers to reduce experimental burden. Unlike existing metabologenomics approaches[8–10], our framework is uniquely resilient to fragmented genomic assemblies and can capture information from previously unassigned spectral features. By decoupling discovery from the need for high-quality assemblies or exhaustive spectral libraries, we enable high-confidence prioritization even in data-sparse environments.

[1]Institute of Sustainability for Chemicals, Energy and Environment (ISCE2), Agency for Science Technology and Research (A*STAR), 8 Biomedical Grove, #07-01 Neuros Building, 138665 Singapore, Republic of Singapore. [2]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, #07-01 Matrix Building, 138671 Singapore, Republic of Singapore. [3] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, 138632 Singapore, Republic of Singapore. [4]Department of Chemistry, National University of Singapore, 3 Science Drive 3, 117543 Singapore, Republic of Singapore. [5]Institute of Molecular and Cell Biology (IMCB), Agency for Science, Technology and Research (A*STAR), 61 Biopolis Drive, #07-06, Proteos, 138673 Singapore, Republic of Singapore. [6]Synthetic Biology Translational Research Program, Yong Loo Lin School of Medicine, National University of Singapore, 10 Medical Drive, 117597 Singapore, Republic of Singapore. [7]These authors contributed equally: Dillon W. P. Tay, Winston Koh. ✉e-mail: dillon_tay@a-star.edu.sg; Winston_Koh@a-star.edu.sg; wong_fong_tian@a-star.edu.sg; lim_yee_hwee@a-star.edu.sg

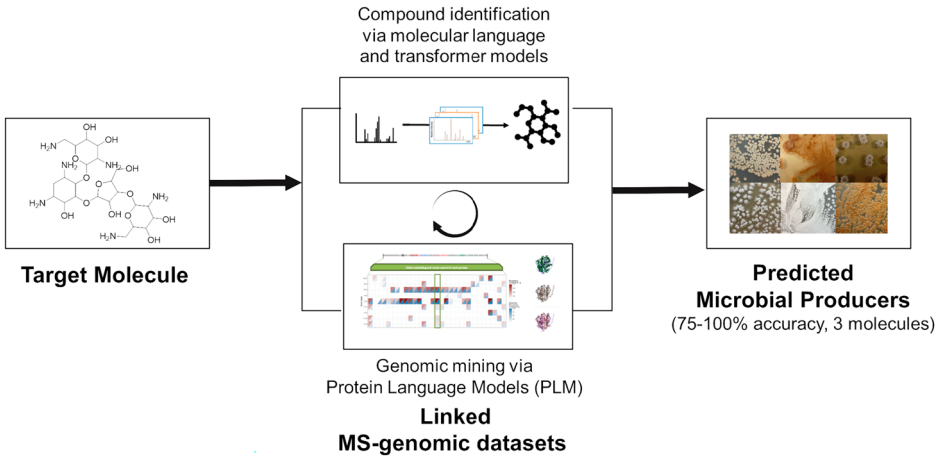

Fig. 1 | Concept of a linked genomics and mass spectrometry (MS) multi-modal dataset and predictive models for streamlined discovery of natural products in microbial strain libraries. Putative producers of compounds of interest are prioritized through cross-validation of the top ranked strains via Protein Language Model (PLM) embeddings and via our in-house developed Workflow for Intelligent Structural Elucidation (WISE).

Results

Genomic inference via protein language model (PLM) embeddings

Protein language models (PLMs) such as ESM-2[11] provide an efficient way to encode genomic sequences into meaningful embeddings. Recent progress in PLMs has also enhanced in silico biophysical analyses, particularly for structure-related tasks including secondary structure[11] and contact prediction[12]. Embeddings from these models can extract remote homology information that may be obscured in primary sequences, enabling a more sophisticated understanding of protein function[13,14]. 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[13,14]. These high-resolution embeddings can subsequently facilitate rapid similarity searches through vector databases such as Faiss[15], allowing high-confidence, scalable, and efficient genomic analyses.

Building on these advancements, we concentrate on higher-resolution protein-level embeddings instead of those derived from entire biosynthetic gene clusters (BGCs). This strategy circumvents the necessity for high-quality genome assemblies, as protein sequences can be annotated even from fragmented or incomplete genomes, as demonstrated in our current genomic dataset (Table S1). Existing methods such as NPLinker[16] and metabologenomic gene cluster family (GCF) networking[8] operate at the level of annotated BGCs or GCFs, relying on tools like antiSMASH[17] and BiG-SCAPE[18] to define cluster boundaries before inferring BGC-metabolite links through co-occurrence statistics across large strain panels. These approaches require both accurate BGC annotation and sufficient panel diversity to achieve reliable correlation. In contrast, by leveraging PLMs, we operate at the level of individual biosynthetic proteins rather than pre-annotated clusters. This distinction has practical consequences: in fragmented genome assemblies, where contigs may split a BGC across multiple scaffolds, conventional annotation pipelines may fail to delineate the cluster entirely. In our case, individual protein embeddings can still be derived and compared against well-characterized reference BGCs in databases such as MIBiG[19]. The genomic signal is therefore preserved even when assembly quality is limiting.

However, we also acknowledge that many biosynthetic enzyme families (e.g. oxidoreductases and transferases) are widely distributed across genomes and often participate in primary metabolism. As such, similarity at the level of a single protein may not be sufficient to infer biosynthetic relevance. To mitigate this, our framework ranks strains based on the coordinated similarity of multiple proteins from a reference BGC rather than any single pairwise comparison. Specifically, we defined a composite score for each BGC-strain pair as the product of the group mean Euclidean similarity and the number of individual protein hits exceeding a minimum similarity threshold of 0.75. A high-confidence composite score threshold of 2.5 was also selected on the basis that it corresponds to the 98th percentile of all BGC-strain searches, thereby isolating only the most statistically exceptional candidates from the extreme right tail of the search score distribution. This data-driven threshold ensures that prioritised candidates represent genuine outliers, which are those combining high mean embedding similarity with multiple independent corroborating hits. By coupling average embedding similarity with the count of high-confidence hits, this formulation intrinsically down-weights strains that match only one or two ubiquitous enzyme families and elevates those exhibiting concordant similarity across several proteins within a pathway, thereby providing a quantitative proxy for biosynthetic context at the protein level. Strains with only isolated matches to common enzyme families tend to rank lower than those showing coordinated similarity across several proteins within the pathway, providing a coarse form of biosynthetic context. We further note that strains with large genomes and extensive secondary metabolism may score highly in genomic analyses alone. To address this potential bias, the genomic signal is integrated with metabolomic evidence through our integrated multi-modal framework described above. In this setting, strains must not only exhibit genomic similarity but also produce chemical features consistent with the target compound to be prioritized. While this strategy does not fully resolve the challenges of multifunctional enzymes or pathway context, it provides a practical means of reducing false positives, and future extensions may incorporate more explicit biosynthetic logic at the pathway level.

Metabolomic inference via the workflow for intelligent structural elucidation (WISE)

Our in-house developed Workflow for Intelligent Structural Elucidation (WISE) leverages generative molecular language processing models and transformer-based structure-to-MS/MS predictive models for putative structural annotation of MS data (Fig. 2A).

WISE first performs deconvolution of LC-MS/MS data to isolate individual metabolite spectra from complex mixtures. The characteristics of individual m/z traces are evaluated across retention time to remove background signals and resolve overlapping compound spectra (Figure S1). To extend the search space beyond existing reference libraries, we leverage molecular language processing[6], a generative approach that taps on Simplified Molecular Input Line Entry System (SMILES) representations as a 'language' of molecular structures to produce large libraries of natural product-like candidates. Predictive transformer-based models[20] are then used to simulate reference MS/MS spectra for these candidates. By fine-tuning the predictive models to account for collision energy, we improve on structural prediction accuracy over state-of-the-art methods[21–23] (Table S2). Finally, a WISE score combining MS/MS, isotopic distribution, and monoisotopic mass similarities is then used to rank candidate structures for putative structural assignments.

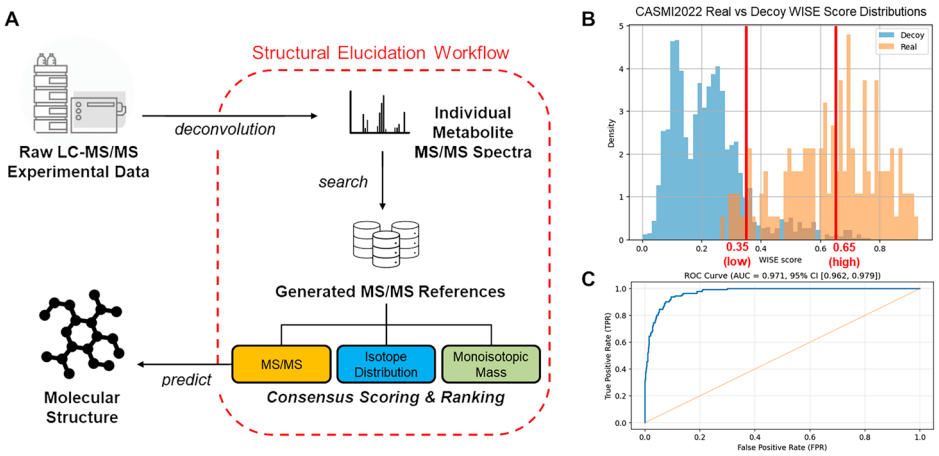

Fig. 2 | Workflow for Intelligent Structural Elucidation (WISE). A Overview of the cheminformatics analysis workflow (WISE) to annotate individual molecules from raw liquid chromatography-tandem mass spectral (LC-MS/MS) data of fermentation extracts. **B** Distribution of WISE scores for correct structures (Real) versus isobaric decoy structures (within 0.005 m/z) with different molecular formulae (Decoy) from the Critical Assessment for Small Molecule Identification (CASMI2022) dataset. Low and high WISE score thresholds of 0.35 and 0.65 respectively are annotated for reference. **C** Receiver Operating Characteristic (ROC) curve demonstrating the discrimination of correct structures from decoys with differing molecular formulae within a 0.005 m/z window. AUC = Area Under Curve. CI = Confidence Interval. Detailed data in Table S3.

To assess the reliability of using WISE scores as a confidence metric, we evaluated its performance on the Critical Assessment for Small Molecule Assessment (CASMI2022) benchmark dataset. When distinguishing between isobaric candidate structures, WISE scores identified the correct structure from decoys with different molecular formulae within a 0.005 m/z window with 94% precision (Fig. 2B and Table S3), achieving an Area Under Receiver Operating Characteristic (AUROC) of 0.971 (Fig. 2C). Its performance reduced to 70% precision and an AUROC of 0.737 when structural isomers of identical molecular formulae were included (Figure S2 and Table S3). This was expected as two of the three components of the WISE score (i.e. monoisotopic mass and isotopic distribution) are identical for such candidates, leaving MS/MS similarity as the sole discriminating metric. Despite this limitation however, WISE scores were able to distinguish specific Bemis-Murcko[24] scaffolds (i.e. flavone, chromone, xanthone) from other scaffolds with identical molecular formulae with an AUROC of 0.927 (Figure S3, Tables S4 and S5). This highlights its utility for practical applications such as identifying pharmacophore producers or shortlisting extracts for more resource-intensive NMR characterization.

Non-parametric kernel density estimation[25] (KDE) of WISE score distributions (Figs. 2B and S2A, Tables S3 and S6) established two functional thresholds for structural annotation. Scores above 0.65 correspond to high confidence assignments, with sufficient discriminative power to resolve structural isomers sharing identical molecular formulae. In contrast, scores below 0.35 define a low confidence region characterized by high decoy density and an increased likelihood of false positives. The significance of these thresholds is illustrated in Fig. 2B. The 0.35 low score threshold serves as a filter for 'obvious' errors. By exceeding the mean WISE score for isobaric decoys (0.22 ± 0.12) by approximately one standard deviation, the 0.35 cutoff successfully excludes the bulk of decoys with incorrect molecular formulae. To address the broader and more challenging decoy population that includes structural isomers (mean WISE score = 0.48 ± 0.18, Figure S2A), a more stringent threshold is required. The 0.65 cutoff, also positioned approximately one standard deviation above this broader decoy distribution, provides a more robust safeguard against high scoring false positives.

A multi-modal prioritization framework to accelerate natural product discovery

When integrated, the two streams of genomic and metabolomic data provide complementary modalities of evidence. The PLM-based analysis evaluates the genomic potential of a strain to produce a given class of molecules, while the MS-based workflow assesses the realized chemical phenotype observed under specific cultivation conditions. As biosynthetic gene clusters may be silent, conditionally expressed, or structurally divergent from known references, combining genomic potential with chemical evidence helps prioritize higher confidence strain-compound hypotheses where both signals co-occur, reducing false positives compared to either modality alone. Together, the cross-validation between these two independent modalities within a single strain complements existing co-occurrence approaches[8–10] by reducing the dependency on panel scale and genome completeness, broadening the settings in which confident BGC-metabolite prioritization can be achieved. Strains and extracts that exhibit concordant signals across both modalities are prioritized for downstream validation, focusing experimental effort on the most promising candidates.

To illustrate this, we applied the framework to a dataset comprising 2138 LC-MS/MS analyses[26] of fermentation extracts generated from a diverse collection of 54 actinobacterial strains and their mutants[27] (Table S9) from the A*STAR National Organism Library[28]. We evaluated three representative antimicrobial case studies (Table S7): valinomycin (a cyclic depsipeptide ionophore, Figures S4–S6), surfactin (a lipopeptide biosurfactant with antimicrobial activity, Figures S7–S8), and neomycin B (an aminoglycoside antibiotic, Figures S9–S11). These compounds span distinct biosynthetic classes and structural architectures, enabling assessment of the framework's generalizability across diverse modes of antimicrobial chemistry.

Across the entire dataset, the WISE workflow yielded 130,728 unique putative structures when considering the top 10 candidates for every MS signal. Although the expanded coverage of WISE reflects its ability to leverage generative search space to explore beyond curated reference libraries, it also underscores the challenge of structural over-annotation, particularly among closely related structural isomers. Consequently, these predicted compounds are treated not as definitive identifications, but as a library of putative structures designed for downstream cross-validation with genomic data. By shifting from the similarity-based grouping used in GNPS to an individual spectral-level identification, WISE significantly enhances the resolution of chemical exploration, providing the necessary breadth to capture novel metabolites that lack existing reference standards.

Case study application: neomycin B, an aminoglycoside antibiotic

To validate the utility of our multi-modal prioritization framework, we applied it to search for microbial producers of Neomycin B (Fig. 3). The BGC encoding neomycin biosynthesis (BGC0000709[29], Fig. 3A) was examined using the ESM2 protein language model to embed protein components. These embeddings were compared against a vector database derived from the genomic dataset using nearest neighbour algorithms to identify top matches (Fig. 3B). Pairwise sequence alignments and structural alignments of predicted structures of the enzyme within the BGC (Fe-S oxidoreductase, CAF33317) also validated the model's effectiveness for capturing sequence and structural relationships (Fig. 3C).

In parallel, LC-MS/MS data of the strains' fermentation extracts were also processed via WISE (Fig. 3D). High potential candidate producers of Neomycin B (Fig. 3E) ranked by both genomic (PLM) and metabolomic (WISE) modalities, with a minimum of 0.35 WISE score and 0.75 PLM score

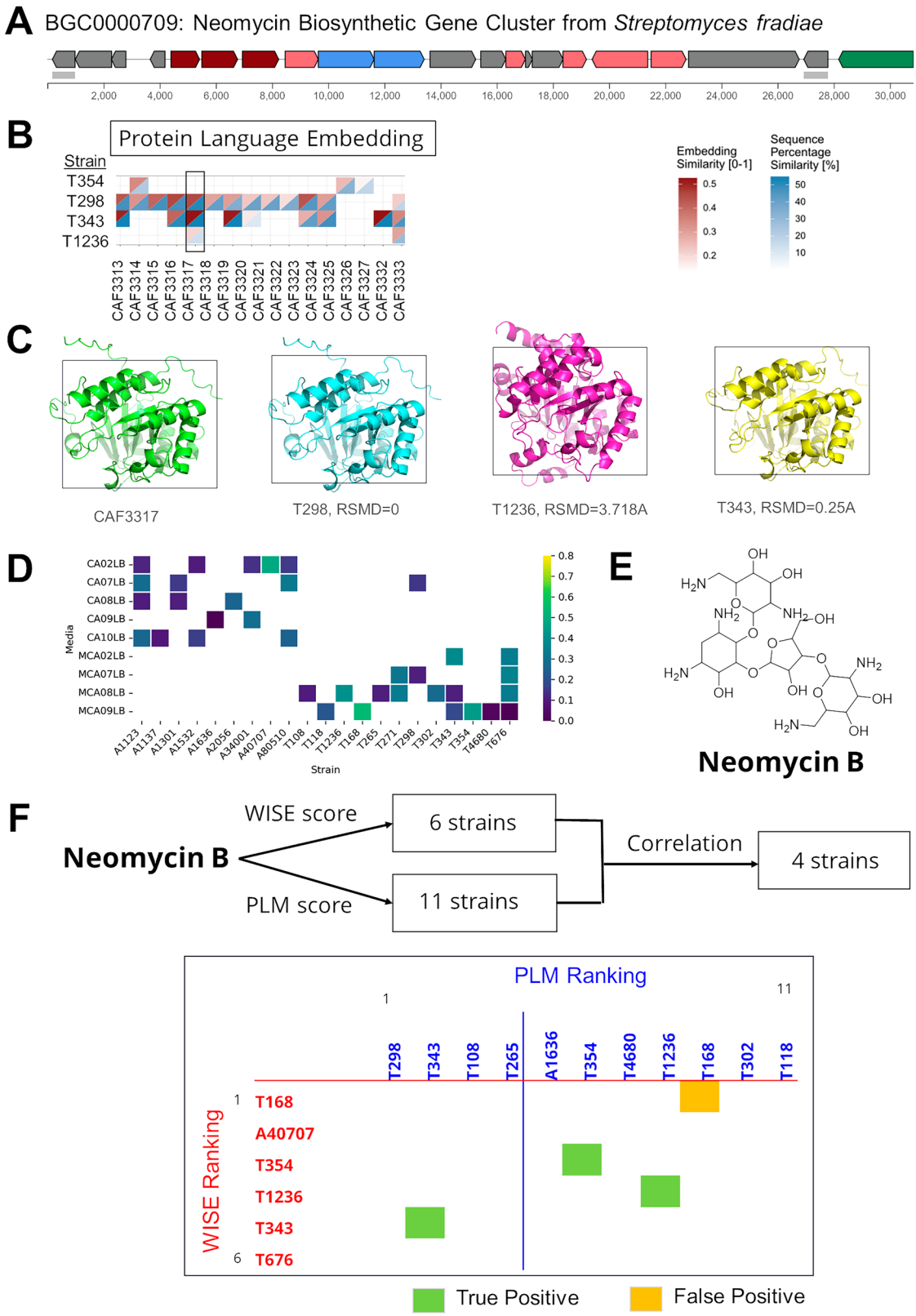

Fig. 3 | Case Study: Neomycin B. A Elucidated biosynthetic gene cluster BGC0000709, responsible for neomycin production. **B** Similarity (%) of each embedded protein and sequence similarity (%) of the top-scoring protein matches. **C** Predicted protein structures of the top matches for Fe-S oxidoreductase (CAF33317) from neomycin BGC. **D** Heatmap of WISE scores for strain-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. Activated strains (if any) are referred to by their parent strain instead of individually (details of the activated strains are in Table S12). **E** Structure of neomycin B. **F** Workflow of the co-occurrence strategy and matrix ranking table with 0.35 WISE score and 0.75 PLM score cutoff. Strains that have been cross validated are highlighted in green or yellow, indicating whether these compounds were present (green) or absent (yellow). Lines indicating higher confidence thresholds of 0.65 WISE score (in red) and 2.5 PLM score (in blue) are also included.

were identified through this cross-validation matrix (Fig. 3F). 4 unique strains were predicted to be Neomycin B producers based on the co-occurrence of PLM and WISE scores, of which three were validated (75% precision) through authentic product MS/MS spectral matching (Table S7, Figures S10–S12). Notably, two of the validated species were previously uncharacterized *Streptomyces* producers of neomycin B (Table S8). Secondary analysis of the T354 draft genome using antiSMASH[17] identified 66 putative BGC regions, dominated by nonribosomal peptide synthetase (NRPS; $n = 18$), type I polyketide synthase (T1PKS; $n = 14$), terpene ($n = 8$), and RiPP-like ($n = 7$) clusters. Notably, antiSMASH failed to detect the neomycin biosynthetic gene cluster (MIBiG accession BGC0000709) or any complete aminoglycoside BGC in the T354 genome. Comparative cluster analysis returned only weak matches (35-48% identity) to transport-associated proteins linked to the neamine cluster (BGC0000708), a related but distinct aminoglycoside pathway. This negative result highlights a known limitation of rule-based genome mining approaches such as antiSMASH which rely on predefined detection rules and reference cluster architectures that may not capture fragmented biosynthetic pathways, particularly in draft genome assemblies where BGCs may be split across multiple contigs. Hence, we interpreted the absence of a neomycin BGC call by antiSMASH not as definitive evidence that T354 lacks the capacity for neomycin production, but rather, as being reflective of the constraints of current computational BGC detection methods when applied to incomplete genome assemblies.

Generalizability and performance validation

By employing the ranking score matrix composed of PLM embeddings and WISE scores with cutoffs, producers of two additional compounds, valinomycin (Figures S12 and S5–S7, Table S10) and surfactin B (Figures S13 and S8–S9, Table S11), were also prioritized with precisions of 86% and 100% respectively. Overall, these performance metrics arise from focused application case studies and therefore should be interpreted as illustrative examples of the framework's practical utility rather than as estimates of global performance across the full dataset. Across the three case studies, the false discovery rate (FDR) across validated predictions came in at a rate of 13% (3 out of 23 cases, Figs. 3, S13, and S14). This level of error supports its intended role as a hypothesis-generation and prioritization tool, where high-confidence candidates are advanced for targeted experimental validation. Notably, these case studies were performed without pre-validation against external models, experimental MS/MS reference data, or high-quality genome annotations, reflecting conditions that frequently arise in real-world natural product discovery campaigns. The framework was also established before applying it to the case study dataset, which served solely as an independent evaluation set. To further assess whether the predictive signal arose from genuine biological associations rather than stochastic correlations, we performed a negative control analysis by randomly permuting the WISE and PLM scores across the 54 strains for the three case study molecules (Figure S14). Under this randomized scenario, precision declined substantially to 10–39%, compared to 75–100% observed with the true score associations. Although this analysis does not provide a global benchmark of predictive accuracy, it demonstrates that the prioritization performance observed in the case studies is driven by a structured biological signal rather than random pairing, supporting the validity of the integrated scoring framework for guiding experimental discovery efforts.

Discussion

Collectively, the multi-modal integration framework significantly addresses several longstanding challenges in natural product research, such as untargeted metabolite inference and identification of new producers in a scalable and accelerated fashion. Individually, advanced techniques in molecular language processing, transformer-based structural predictions, and protein-level embeddings can rapidly identify patterns in their respective datasets. We are cognizant that challenges such as data quality and model interpretability can limit this approach. Predictions based solely on PLM or WISE scores can significantly overestimate potential producers, resulting in false positives that would have otherwise been eliminated by their cross-validation (Table S13). Therefore, it is critical to ensure alignment between biological potential and metabolic outputs. Utilizing this multi-modal prioritization strategy with defined confidence thresholds helps achieve high precision (75–100%). Through integrating both metabolomic and genomic insights, we effectively bridge the critical gap between data generation and actionable biological insights, substantially accelerating discovery of producer strains of target compounds. Like with all prediction methodologies, there remain false positives (2/14 for valinomycin and 1/4 for neomycin B) despite consensus in PLM and WISE score alignment, possibly due to data quality and inherent bias in the processing algorithms. The complexity of interpreting high-dimensional data and the potential for error in vector-based searches underscore the need for continued refinement of our algorithms and methodologies. Nonetheless, the ability to jointly interrogate metabolomic and genomic data at scale creates exciting opportunities for prospective discovery of novel natural products and previously unrecognized producer strains. By expanding the searchable chemical and biosynthetic space beyond curated reference databases, our approach has the potential to uncover previously inaccessible regions of natural product diversity, thereby broadening the boundaries of known chemical and microbial biodiversity.

Methods

Genomics sequencing

Genomic libraries were prepared for each strain using 10 ng of genomic DNA and the plexWell™ 96 kit, according to the manufacturer's instructions. Sequencing was performed on an Illumina HiSeq 4000 platform using a 2 × 151 bp paired-end protocol. Raw Illumina sequencing reads were first evaluated using FastQC to assess base quality, adapter contamination, and sequence duplication. Adapter trimming and quality filtering were performed using Trimmomatic[30] with a minimum read length of 36 bp. High-quality reads were then assembled de novo using SPAdes with parameters optimized for microbial genome reconstruction. Assembly quality was quantified using N50 and total assembly size, across the dataset. Protein-coding sequences were identified from assembled contigs using Prodigal[31]. The predicted protein sequences within these regions were then extracted to serve as candidates for downstream computational analysis, including embedding-based similarity search.

Similarity searches using kNN in protein language model embeddings

To identify functional analogs across the sequenced strains, we applied ESM2, a transformer-based protein language model, to generate embeddings for the predicted proteins. Embeddings were computed following the ESM2 protocol (https://github.com/facebookresearch/esm) and indexed using an OpenSearch vector database, paired with metadata including strain identifiers and environmental origin. Similarity searches were conducted using a k-nearest neighbors (kNN) algorithm based on cosine distance in the high-dimensional embedding space. For each query protein—such as enzymes linked to specific compound classes in the MIBiG reference database—we retrieved the top-ranked candidates by embedding proximity. This approach enabled rapid prioritization of homologous or convergently evolved enzymes potentially involved in natural product biosynthesis, even in the absence of high sequence identity, demonstrating the utility of language model-based representations in functional protein mining. Using a PLM approach on fragmented genomes, 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.

antiSMASH analysis

The T354 draft genome assembly (1787 contigs; 15.5 Mb) was subjected to secondary analysis using antiSMASH v8.0.4[17] with default settings. Gene prediction was performed using Prodigal[31], and candidate biosynthetic gene

clusters (BGCs) were annotated using the antiSMASH pipeline, including KnownClusterBlast searches against the MIBiG[19] repository for similarity to previously characterized clusters.

Structure-to-MS/MS prediction

Tandem mass spectra (MS/MS) of molecular structures were predicted using a transformer-based model adapted from ICEBERG[20] that takes as inputs: (1) molecular structure in SMILES format, (2) adduct type, and (3) collision energy in eV, to predict their corresponding MS/MS spectral reference. Spectral references were generated for structures from both the COCONUT[7] and generated 67 million natural product[6] 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)

The workflow begins by performing deconvolution analyses of MS1 signals at the individual m/z level that separates MS1 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, indicating a common scaffold or fragment from compound analogs eluting at different distinct times, and (3) 'Compound-like': signals which follow a regular peak shape distribution, in line with compound elution from the column (Figure S1). The MS/MS spectra of the deconvoluted compound-like signals are compared against the generated MS/MS reference database across three metrics of (1) MS/MS spectra similarity, (2) monoisotopic mass difference, and (3) isotopic distribution similarity, to give a composite WISE score. For each of the deconvoluted compound-like signals, the corresponding precursor *m/z* is first used to shortlist and filter the augmented MS/MS spectral reference database for entries that are within ±0.005 *m/z*. WISE scores are then calculated between the identified compound-like signal and every shortlisted reference according to Eq. (1):

$$WISE\ Score = 0.5 \cdot (ms\ match) + 0.2 \cdot (iso\ match) + 0.3 \cdot (mass\ match) \quad (1)$$

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 theoretical monoisotopic mass

Formulas for the individual match scores are detailed in Eqs. (2) to (4):

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

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 molecular structure[32].

$$iso\ match = \max\left(\left(\frac{\sum_i norm\ \exp_i \cdot norm calc_i}{\sqrt{\left(\sum_i norm\ \exp_i^2\right) \cdot \left(\sum_i norm calc_i^2\right)}} - 0.999\right) \cdot 1000, 0\right) \quad (3)$$

$Norm\_exp_i$ 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 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%.

$$mass\ match = \frac{\max(0.005 - abs(mass\_\exp - mass\_calc), 0)}{0.005} \quad (4)$$

Mass_exp is the precursor *m/z* of the experimental compound observed and mass_calc is the calculated theoretical precursor *m/z* of the reference compound. 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).

WISE score threshold analysis

To determine a data-driven decision threshold for WISE scores, we compared the empirical score distributions of true ('real') matches and decoy matches using non-parametric kernel density estimation[25] (KDE). Real WISE scores and decoy WISE scores were first aggregated across all entries into two one-dimensional score vectors. The pooled real and decoy score distributions were visualized as normalized histograms and further modelled using Gaussian KDEs. KDEs were fitted independently to the real and decoy score vectors using the default bandwidth selection implemented in scipy.stats.gaussian_kde (SciPy v1.10.1). To identify classification thresholds, the score domain was defined as the interval spanning the minimum and maximum values observed across both pooled distributions. This interval was discretized into 1000 evenly spaced candidate score values (i.e. 0.001 WISE score intervals). At each candidate WISE score, the estimated densities of the real and decoy distributions were evaluated, and the absolute density difference was computed. The threshold was taken as the candidate score at which this absolute difference was minimized (i.e. the numerical approximation to the intersection of the two KDE curves).

Metabolomic characterization of the strain library

The reported 2138 LC-MS/MS dataset[27] is 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 spectra identified were taken to form a pool of putative metabolite annotations from which searches were done to identify producers of valinomycin, surfactin B, and neomycin B.

Case study validation

The presence of valinomycin and surfactin B in microbial strain fermentation extracts were validated via isolation and structural confirmation by NMR[27], as well as independent spectral matching of experimental LC-MS/MS data against the Global Natural Products Social Molecular Networking (GNPS)[5] spectral reference libraries[26]. 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 ±50 Da 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 reference MS/MS obtained from analyzing a commercial standard of Neomycin B, as well as independent comparison with the GNPS reference of Neomycin B (CCMSLIB00010114511). Specifically, the WISE processed experimental LC-MS/MS dataset was filtered to those annotated as $[M+Na]^+$ adducts, had precursor *m/z* that was ±0.005 *m/z* of

637.3015 (the calculated theoretical monoisotopic mass of the $[M+Na]^+$ adduct of Neomycin B), and a retention time between 6.94–7.14 min (the range of experimental retention times observed for Neomycin B). The shortlist was further filtered to keep only experimental spectra with at least 4 matched peaks (that were at least 5% of base peak abundance and ±5 Da away from each other) to both in-house commercial and GNPS reference spectra.

Negative control analysis by score permutation

For each of the three case studies (valinomycin, surfactin B, and neomycin B), the WISE and PLMs scores assigned were randomly shuffled across the 54 strains while preserving the original marginal score distributions. Specifically, a common random permutation of strain indices was generated for each iteration, and both the WISE and PLM scores were reassigned according to this permutation, thereby disrupting the true strain-score associations while maintaining the underlying range and structure of the scores. For each shuffled dataset, classification performance was then evaluated using the same fixed decision rule applied in the main analysis: strains were predicted as positives only if both the WISE score and PLM score exceeded predefined thresholds (WISE ≥ 0.35 and PLM ≥ 0.75). This procedure was repeated across 100 independent random shuffles, using distinct random seeds sampled without replacement.

Data availability

The LC-MS/MS dataset of fermentation extracts from 54 actinobacterial strains and their mutants used in this work has been deposited and is publicly accessible via MassIVE with the accession number MSV000092237 (https://doi.org/10.25345/C53X83W53). Reference MS/MS spectra of Valinomycin (CCMSLIB00005721598), Surfactin B (CCMSLIB00005727861), and Neomycin B (CCMSLIB00010114511) are available from the Global Natural Products Social Molecular Networking (GNPS)[5] spectral reference libraries.

Code availability

The code used to train the molecular language processing model as well as the trained model used to generate natural product-like molecules is available from Github at https://github.com/SIBERanalytics/Natural-Product-Generator [6]. The code used to train the structure-to-MS/MS model (ICEBERG)[20] as well as the pre-trained model is available from Github at https://github.com/samgoldman97/ms-pred. The data processing Workflow for Intelligent Structural Elucidation (WISE) is partially based on Mestrenova v15.0.1-35756 which is commercially available from Mestrelab Research S. L. (www.mestrelab.com). The remaining code in WISE for MS/MS-to-structure identification and to recreate the application case studies shown here is available from Github at https://github.com/SIBERanalytics/MS2-to-Structure. The code used for generating protein sequence embeddings using ESM-2, indexing them in a vector database for fast similarity searches via approximate nearest neighbor (ANN) methods is available from Github at https://github.com/lianchye/pLM_analysis.git.

Received: 28 June 2025; Accepted: 13 April 2026;
Published online: 30 April 2026

References

1. Butler, M. S. The role of natural product chemistry in drug discovery. *J. Nat. Prod.* **67**, 2141–2153 (2004).
2. Jensen, P. R., Chavarria, K. L., Fenical, W., Moore, B. S. & Ziemert, N. Challenges and triumphs to genomics-based natural product discovery. *J. Ind. Microbiol. Biotechnol.* **41**, 203–209 (2014).
3. Poynton, E. F. et al. The Natural Products Atlas 3.0: extending the database of microbially derived natural products. *Nucleic Acids Res.* **53**, D691–D699 (2024).
4. Johnson, S. R. & Lange, B. M. Open-access metabolomics databases for natural product research: present capabilities and future potential. *Front. Bioeng. Biotechnol*. **3**, https://doi.org/10.3389/fbioe.2015.00022 (2015).
5. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
6. 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* **10**, 296 (2023).
7. Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A. & Steinbeck, C. COCONUT online: collection of open natural products database. *J. Cheminform.* **13**, 2 (2021).
8. Caesar, L. K. et al. Correlative metabologenomics of 110 fungi reveals metabolite–gene cluster pairs. *Nat. Chem. Biol.* **19**, 846–854 (2023).
9. Parkinson, E. I. et al. Discovery of the tyrobetaine natural products and their biosynthetic gene cluster via metabologenomics. *ACS Chem. Biol.* **13**, 1029–1037 (2018).
10. Goering, A. W. et al. Metabologenomics: correlation of microbial gene clusters with metabolites drives discovery of a nonribosomal peptide with an unusual amino acid monomer. *ACS Cent. Sci.* **2**, 99–108 (2016).
11. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
12. Chen, J.-Y. et al. Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review. *Front. Bioeng. Biotechnol*. **13**, https://doi.org/10.3389/fbioe.2025.1506508 (2025).
13. Liu, W. et al. PLMSearch: Protein language model powers accurate and fast sequence search for remote homology. *Nat. Commun.* **15**, 2775 (2024).
14. Johnson, S. R., Peshwa, M. & Sun, Z. Sensitive remote homology search by local alignment of small positional embeddings from protein language models. *eLife* **12**, RP91415 (2024).
15. Douze M. et al. The Faiss Library. In *IEEE Transactions on Big Data*, **12**, 346–361 (2026).
16. Hjörleifsson Eldjárn, G. et al. Ranking microbial metabolomic and genomic links in the NPLinker framework using complementary scoring functions. *PLoS Comput. Biol.* **17**, e1008920 (2021).
17. Blin, K. et al. antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. *Nucleic Acids Res.* **53**, W32–W38 (2025).
18. Draisma, A. et al. BiG-SCAPE 2.0 and BiG-SLiCE 2.0: scalable, accurate and interactive sequence clustering of metabolic gene clusters. *Nat. Commun.* **17**, 2000 (2026).
19. Zdouc, M. M. et al. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. *Nucleic Acids Res.* **53**, D678–D690 (2024).
20. Goldman, S., Li, J. & Coley, C. W. Generating molecular fragmentation graphs with autoregressive neural networks. *Anal. Chem.* **96**, 3419–3428 (2024).
21. Wang, F. et al. CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. *Anal. Chem.* **93**, 11692–11700 (2021).
22. Wei, J. N., Belanger, D., Adams, R. P. & Sculley, D. Rapid prediction of electron–ionization mass spectrometry using neural networks. *ACS Cent. Sci.* **5**, 700–708 (2019).
23. Tsugawa, H. et al. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Anal. Chem.* **88**, 7946–7958 (2016).
24. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
25. Marzio, M. D. & Taylor, C. C. Kernel density classification and boosting: an L2 analysis. *Stat. Comput.* **15**, 113–123 (2005).
26. Tay, D. W. P. et al. Tandem mass spectral metabolic profiling of 54 actinobacterial strains and their 459 mutants. *Sci. Data* **11**, 977 (2024).

27. Tay, D. W. P. et al. Exploring a general multi-pronged activation strategy for natural product discovery in Actinomycetes. *Commun. Biol.* **7**, 50 (2024).
28. Ng, S. B. et al. The 160K Natural Organism Library, a unique resource for natural products research. *Nat. Biotechnol.* **36**, 570–573 (2018).
29. Huang, F. et al. The neomycin biosynthetic gene cluster of Streptomyces fradiae NCIMB 8233: characterisation of an aminotransferase involved in the formation of 2-deoxystreptamine. *Org. Biomol. Chem.* **3**, 1410–1418 (2005).
30. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
31. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **11**, 119 (2010).
32. Nicolescu, T. O. Interpretation of mass spectra. In *Mass Spectrometry* (ed. Aliofkhazraei, M.) 23–78 https://doi.org/10.5772/intechopen.68595 (IntechOpen, 2017).

Acknowledgements

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, A*STAR). We gratefully acknowledge the Biocomputing Centre at the A*STAR Bioinformatics Institute 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.

Author contributions

D.W.P.T. contributed to the methodology, algorithm development, data analysis, visualization, and manuscript editing. W.K. was involved in conceptualization, algorithm development, data analysis, manuscript editing, and supervision of the work. S.J.A., Z.M.W., N.Z.X.Y., K.A. and Y.H.N. contributed to algorithm development and data analysis. Y.W.L. and E.H. performed experimental validation and investigations. F.T.W. and Y.H.L. supervised the work, conceptualized the study, and edited the manuscript. Y.H.L. wrote the original draft.

Competing interests

Two provisional patent applications have been filed by some of the authors wherein some of the data are disclosed in this manuscript. The authors (Y.H.L., D.T., S.J.A. and Z.M.W.) are inventors to SG Patent 10202400652R, and the authors (Y.H.L., D.T., S.J.A., N.Z.X.Y. and K.A.) are inventors to SG Patent 10202403057W.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44259-026-00206-7.

Correspondence and requests for materials should be addressed to Dillon W. P. Tay, Winston Koh, Fong Tian Wong or Yee Hwee Lim.