Acta Pharmaceutica Sinica B 2026;16(8):5034−5052

Chinese Pharmaceutical Association
Institute of Materia Medica, Chinese Academy of Medical Sciences

Acta Pharmaceutica Sinica B

www.elsevier.com/locate/apsb
www.sciencedirect.com

TOOLS

DeepHalo: Deep learning-powered exploration of halogenated metabolites uncovering antibacterial depsipeptides

Shanshan Chang[a,†], Xin Qi[a,†], Mengyuan Wang[a,†], Xinyue Huang[a], Ning He[a], Mingxu Chen[a], Qing Lv[a], Jiahan Wang[a], Yu Du[a], Shuchen Wang[b], Yihong Li[a], Quanxiu Gao[a], Xinran Chen[a], Xingxing Li[a,*], Bin Hong[a,*], Yunying Xie[a,*]

[a]*CAMS Key Laboratory of Synthetic Biology for Drug Innovation, Institute of Medicinal Biotechnology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100050, China*
[b]*The School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 100049, China*

Received 30 August 2025; received in revised form 22 December 2025; accepted 30 December 2025

KEY WORDS

Halogenated natural products;
Metabolomics;
Deep learning;
Halogenases;
Antibacterial activity;
Vancomycin-resistant *Enterococcus faecium* (VRE);
High-resolution mass spectrometry;
Metabolites

Abstract In the omics era, confident high-throughput analytical tools are crucial for the efficient identification of metabolites. Here, we present DeepHalo, a deep learning-integrated and hierarchically optimized workflow designed for high-throughput exploration of halogenated metabolites from high-resolution mass spectrometry-based metabolomics. DeepHalo leverages deep learning models combined with a comprehensive scoring to enhance the reliability of halogen predictions. It integrates PyOpenMS for fast isotope pattern detection and incorporates a halogen-based dereplication algorithm with GNPS molecular networking to efficiently exploit and annotate halogenates from complex biological matrices. To validate its performance, DeepHalo was applied to explore halogenated metabolites from 1296 microbial culture crudes, leading to the discovery of six families of structurally diverse halogenated molecules. This included a new class of cyclic depsipeptides, aglomycins A−E, featuring rare 3-chloroanthranilic acid and/or epoxyvaline blocks. Additionally, a plausible biosynthetic pathway of aglomycins was proposed through bioinformatics analyses and targeted gene knockout experiments. Bioassays revealed that aglomycin A exhibits synergistic antibacterial activity with linezolid against vancomycin-resistant

*Corresponding authors.
E-mail addresses: lixingxing@imb.pumc.edu.cn (Xingxing Li), hongbin@imb.pumc.edu.cn (Bin Hong), xieyy@imb.pumc.edu.cn (Yunying Xie).
[†]These authors made equal contributions to this work.
Peer review under the responsibility of Chinese Pharmaceutical Association and Institute of Materia Medica, Chinese Academy of Medical Sciences.

https://doi.org/10.1016/j.apsb.2026.02.010

Enterococcus faecium (VRE) both *in vitro* and *in vivo*. We envision that DeepHalo, a user-friendly standalone executable freely available at https://github.com/xieyying/deephalo/releases/tag/DeepHalo_V1.0.0, will become a powerful tool for accelerating the discovery of halogenated "dark matter".

1. Introduction

Halogenated natural products represent an important class of pharmacologically relevant compounds, as the incorporation of halogen atoms often significantly enhances their biological activity[1]. Additionally, halogens can also improve a substance's metabolic stability and membrane permeability[1,2]. These combined advantages have made halogen substituents a common feature in over 25% of approved drugs[2], highlighting their immense potential in drug development.

Nature, particularly microorganisms, harbors a wealth of halogenated metabolites awaiting exploration. Biogenically, halogens, primarily chlorine and bromine, are incorporated into substrates by diverse enzymes known as halogenases through electrophilic, radical, or, less commonly, nucleophilic mechanisms[3,4]. Flavin-dependent halogenases (FDHs) and iron (II)-α-ketoglutarate-dependent halogenases (KDHs) are the most prevalent types, while new classes of halogenases are continually being uncovered[3-5]. Genomic studies have revealed that microorganisms possess a vast repertoire of halogenase-encoding genes[6-8]. However, a significant gap exists between the prevalence of halogenase candidates and the number of identified halogenated metabolites[6,7], suggesting that many such natural products remain undiscovered. Traditional natural product discovery approaches are hindered by high rediscovery rates and the frequent silencing of biosynthetic gene clusters (BGCs), necessitating more efficient strategies. While genomics has advanced the exploration of halogenated compounds in the omics era[9], its reliance on known halogenases limits the discovery of structurally diverse metabolites. In contrast, metabolomics provides a powerful alternative by directly profiling halogenated molecules, circumventing enzyme-specific constraints and enabling the identification of a broader range of halogenated natural products.

High-resolution mass spectrometry (HRMS)-based metabolomics is a powerful tool for exploring halogenated metabolites. Statistical analysis reveals that chlorinated and brominated derivatives collectively account for over 98.5% of characterized halogenated natural products[10]. The unique isotopic distributions of chlorine ($^{35}Cl/^{37}Cl$) and bromine ($^{79}Br/^{81}Br$) atoms provide a valuable chemical signature for their identification based on HRMS. Although specialized tools, such as MeHaloCoA[11], dynamic cluster analysis (DCAnalysis)[12], HaloSeeker[13], and ChloroDBPFinder[14], have been developed for the targeted detection of halogenated compounds from MS data, it remains far from achieving high-throughput and efficient discovery of these metabolites. The challenges primarily arise from three major obstacles. First, the confidence of halogen prediction has not yet met the requirements of high-throughput screening. Each culture extract typically yields hundreds to thousands of features[15], requiring an analysis method with an exceptionally low false positive rate (ideally below 0.1%) to mitigate subsequent laborious investigations on erroneous detections. Technically, the inaccuracy mainly stems from inherent limitations in current isotope-based prediction models[11-14,16], and is further exacerbated by errors in the extracted isotope patterns, a critical flaw overlooked by all existing algorithms. Second, the speed of analysis is also a significant issue. Many existing approaches require several to dozens of minutes per sample[11-14] and depend on heavily semi-automated[11,13] or manual interventions[12] (Supporting Information Table S1), rendering them impractical for analysis of large-scale MS data. Finally, efficient dereplication is also critical to distinguish genuinely new halogenated natural compounds from known entities in complex chemical backgrounds. Unfortunately, most current methods fall short in this regard[11,12], impeding the prioritization of truly novel compounds. Therefore, there is an urgent need for a more accurate, rapid, and automated computational approach to overcome these limitations and enable the efficient discovery of novel halogenated metabolites.

Herein, we present DeepHalo[17], a deep learning-integrated and fully automated workflow that seamlessly integrates the PyOpenMS[18] and GNPS[19] platforms, enabling the efficient and high-throughput discovery of novel halogenated natural products from complex biological matrices (Fig. 1A). DeepHalo is designed to overcome the major limitations of existing tools through several key innovations: (1) a deep learning-based Element Prediction Model (EPM) that achieves state-of-the-art accuracy in halogen detection by leveraging isotopic patterns; (2) an optional isotope pattern validation step that corrects for extraction errors, a critical flaw overlooked by other methods; (3) a hierarchical halogen confidence score (H-score) enabling adjustable, high-confidence filtering; (4) integrated MS^1 dereplication and MS^2 molecular networking for efficient annotation; and (5) computational optimizations granting orders-of-magnitude faster analysis speed, making it fully applicable to large-scale datasets. To demonstrate its capabilities, we applied DeepHalo to explore halogenated metabolites from 1296 microbial culture crudes (Fig. 1B), resulting in the identification of a new class of halogenated cyclic peptides, aglomycins, along with five families of known halogenated molecules. Subsequent bioinformatics analyses and targeted gene knockout experiments disclosed a plausible biosynthetic pathway of aglomycins, while bioactivity assays revealed that aglomycin A exhibits potent antibacterial activity against clinical multidrug-resistant *Enterococcus* strains and demonstrates synergistic effects with linezolid both *in vitro* and *in vivo*.

2. Materials and methods

2.1. Datasets and preprocessing

2.1.1. Datasets for model building

For the training and evaluation of the models of Element Prediction (EPM) and Anormal isotope pattern Detection (ADM), we

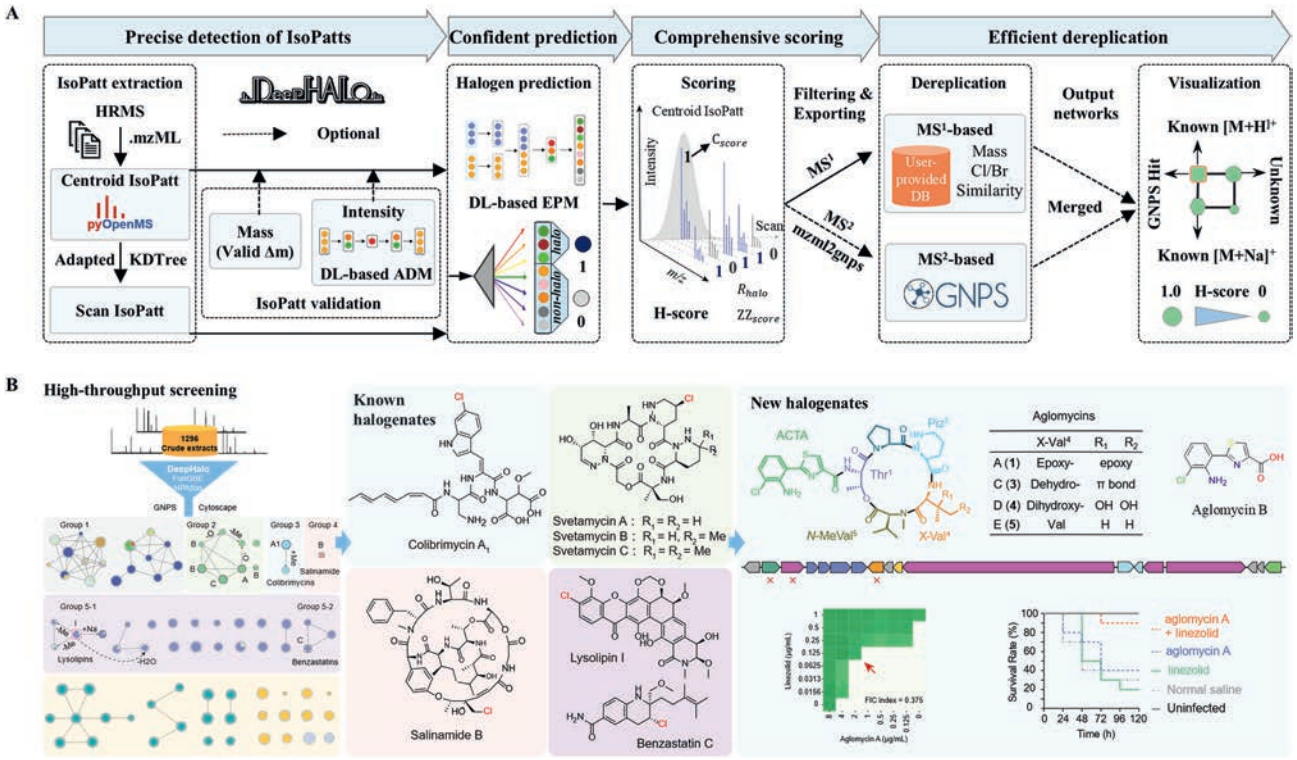

Figure 1 DeepHalo-assisted discovery of antibacterial depsipeptides. (A) Schematic diagram of DeepHalo workflow. (B) DeepHalo was applied for high-throughput screening of halogenated metabolites from 1296 microbial culture crudes, leading to the identification of five classes of structurally diverse halogenates, along with a new group of halogenated cyclic peptides, aglomycins.

generated over one million simulated isotopic spectra from six compound databases[10,20-24]. The datasets were carefully filtered, augmented (*e.g.*, dimerization, synthetic molecules), and split into subsets (IsoBase, IsoHN, and IsoBase_Val_SN_Eva) with added noise to support model training, validation, and fair benchmarking.

2.1.2. Real-world dataset for EPM model evaluation

To rigorously evaluate the EPM under complex real-world conditions, we utilized the CASMI_Myxo_Plus dataset (Supporting Information Table SD1). This dataset integrates high-resolution isotope patterns from the *myxo* dataset[25](88 patterns acquired on a Bruker MaXis 2G qTOF spectrometer) and the CASMI 2016 dataset[26,27] (366 positive and 166 negative patterns measured on a Q Exactive Plus Orbitrap), totaling 619 isotope patterns (mass range: 67.042−2213.962 Da) featuring chlorine, boron, and bromine-containing compounds (with the only selenium compound excluded due to large mass error). To further challenge the model, we augmented the dataset with four edge cases: two iron adducts, one overloaded sample, and one overlapped dehydro isomers.

2.1.3. Simulated dataset for DeepHalo workflow benchmarking

Gold standard datasets with verified ground truths are crucial for benchmarking algorithms. However, real-world LC−MS datasets often fall short of this ideal due to contaminants and the inherent uncertainty in instrument noise and accuracy. To circumvent these issues, we employed simulated metabolomic data with precise ground truth. Using SMITER_modified, a customized version of the LC−MS simulator from the original SMITER tool[28], we generated a series of crowded LC−MS chromatograms with varying resolutions and accuracies, based on a published collection of 1820 compounds[29] (Supporting Information) to form SM1820-Base. From these simulations, ten replicate LC−MS data with a mass accuracy of 1 ppm and a resolution of 2×10^4 were selected and combined to form the SM1820-R20K dataset, which was used to benchmark DeepHalo against other methods.

2.1.4. Measured dataset for DeepHalo workflow benchmarking

To evaluate the effectiveness of DeepHalo in identifying chlorine- or bromine-containing compounds in real-world complex samples, a series of LC−HRMS/MS analyses were acquired using UPLC−QTof (Supporting Information). A standard sample containing 11 purified compounds (5 with one or two chlorine atoms and 2 with a bromine atom), with masses ranging from 238.0509 to 1447.4300 Da, was analyzed at low, medium, and high concentrations using UPLC−HRMS/MS. Each concentration was analyzed with one to three replicates, resulting in 14 LC−HRMS runs, collectively referred to as the IN-house STandards dataset (INST). Additionally, to assess DeepHalo's capability in detecting halogenated compounds within complex samples, spiked sample analyses were performed. Varying amounts of the standard samples were mixed with F1 and M3 culture media (Supporting Information) and analyzed *via* UPLC−HRMS/MS, generating the IN-house STandards Spiked in Culture Media dataset (INST_SCM). All resulting data were converted to. mzML format, making them suitable for DeepHalo analysis.

Additionally, we also evaluated the efficiency of DeepHalo by applying it to the dataset of the Critical Assessment of Small Molecule Identification 2022 contest (CASMI 2022)[30]. This

dataset consists of 145 Orbitrap LC−MS/MS data in both ESI(+) and ESI(−) modes for 500 unique compounds, with 37 precursor adducts bearing chlorine or bromine.

2.2. Construction of DeepHalo

The DeepHalo framework was implemented in Python using TensorFlow, Keras, PyOpenMS, and KD-tree. It consists of four main components: a dataset processing module, a deep learning module, an LC−MS data analysis module, and a post-processing module.

The dataset processing module handles raw input data and generates training features by calculating isotope patterns from molecular formulas based on the above datasets and preprocessing methods. For each compound, the first six isotope peaks (p0−p5) are extracted. The relative intensities of these peaks, normalized to their maximum value, along with the normalized monoisotopic mass (m0) and the mass differences Δm1 (p1−p0) and Δm2 (p2−p1), are computed and prepared as input features for model training.

The deep learning module constructs two deep neural network (DNN) models: one for element prediction (EPM) and the other for anomalous isotope pattern detection (ADM). DeepHalo employs isotope patterns to identify the presence of chlorine and bromine in molecules. To achieve this, we developed the Isotope Pattern Neural Network (IsoNN) architecture and trained an Element Prediction Model (EPM) using approximately 1.09 million simulated isotope patterns. IsoNN is a dual-branched deep neural network that processes two distinct types of input data: the relative intensities and masses of isotope peaks (including monoisotopic mass and mass differences Δm1/Δm2). The intensity branch incorporates Gaussian noise injection for regularization, followed by dense layer transformations, while the mass feature branch applies a nonlinear scaling operation, Gaussian noise injection, and dense processing. The processed features from both branches are concatenated and passed through a feed-forward network, culminating in classification probabilities *via* a softmax output layer. Model performance was evaluated on both simulated (IsoBase_Val_SN_Eva) and real (CASMI_Myxo_Plus) datasets using standard metrics such as Precision, Recall, F1-score, and False Positive Rate (FPR), with comparisons to alternative methods. The training process and hyperparameter optimization details are provided in the Supporting Information

Following the construction of the EPM, an ADM was developed for isotope peaks validation based on their intensities. Unlike the EPM, the ADM employs a deep autoencoder neural network trained on the IsoHN dataset. It uses the feature embeddings (output from the first dense layer of the relative intensity branch in the EPM) as input, leveraging its pre-trained representations to enhance anomaly sensitivity. The network uses ReLU activation functions, two hidden layers (dimensions: [64,16]), optimized by minimizing MSE loss *via* Adam (learning rate: 0.0003). After training, the reconstruction error was computed on the IsoHN dataset. The anomaly threshold was determined by locating the inflection point of the reconstruction error (Q_τ) versus percentile (τ) curve, defined as Eqs. (1) and (2):

$$t_{\text{threshold}} = Q_{\tau^*} \tag{1}$$

$$\tau^* = \underset{\tau}{argmin}\left|\frac{\mathrm{d}^2 Q_\tau}{\mathrm{d}\tau^2}\right| \tag{2}$$

where Q_τ is the τ-th percentile of reconstruction errors, and the second derivative is approximated *via* discrete differencing.

The LC−MS data analysis module is the core component of DeepHalo and performs four key functions: (1) isotope pattern detection and validation from raw LC−MS data, (2) halogen presence prediction, (3) comprehensive scoring, and (4) MS^1/MS^2 data output. For isotope pattern extraction, we developed a robust PyOpenMS-based solution, selected for its superior performance and Python compatibility. The customized workflow enables dual-level isotopic pattern extraction, simultaneously characterizing overall distributions at the centroid feature level while capturing individual patterns at the scan level. To further improve the confidence of feature detection, we established a two-dimensional validation process for centroid isotope patterns. The first validation stage examines mass differences using strict criteria derived from the IsoHN dataset (0.05%−95.5% quantile ranges for Δm1 and Δm2). Specifically, starting with the first isotope peak in a detected cluster, three consecutive peaks that meet the mass difference conditions are identified as the first three isotope peaks. The second validation stage employs the trained ADM to filter isotopes based on their relative intensity patterns. Validated isotopic features are then classified using the optimized EPM. We further introduce a halogen confidence score (H-score) that integrates evidence from both centroid and scan levels to comprehensively evaluate halogen presence. The H-score is defined as Eq. (3):

$$H_{\text{score}} = \frac{1}{3}(\text{ZZ}_{\text{score}} + R_{\text{halo}} + C_{\text{score}}) \tag{3}$$

where ZZ_{score} represents the consistency of halo-positive across consecutive scans, calculated using a modified ZigZag index[31] as Eq. (4):

$$\text{ZZ}_{\text{score}} = \left(1 - \frac{\sum_{i=1}^{N-2}(2S_i - S_{i-1} - S_{i+1})^2}{4N-8}\right) R_{\text{halo}} \tag{4}$$

R_{halo} represents the proportion of halogen-positive scans: $R_{\text{halo}} = \frac{\sum_{i=1}^{N} S_i}{N}$

C_{score} is a binary indicator (1/0) based on the prediction of the centroid isotope pattern of the feature. S_i represents the EPM prediction result for scan i (1 for halogen presence, 0 for absence), and N is the total number of scans in a feature. Subsequently, the H-score serves as a critical filter, with only features meeting the threshold proceeding to MS^1 data export (CSV format). For LC−MS data acquired in DDA mode, the corresponding MS^2 spectra of filtered features can be optionally exported *via* a customized mzml2gnps package.

To establish an optimal threshold for this filter, we evaluated the impact of the H-score on recall, precision, and F1-score using simulated (SM1820-R20K) and experimental datasets (INSD, INSD_SCM, CASMI 2022). Given that both sensitivity and confidence are equally critical in large-scale screening, we employed the F1-score (the harmonic mean of recall and precision) as our primary metric, defining the optimal threshold as the H-score yielding the highest F1-score.

Finally, the post-processing module is responsible for dereplication and writing DeepHalo results to GNPS output files. If a user-supplied chemistry database is available, the pipeline performs MS^1-based dereplication. Candidate compounds are

identified by matching precursor ions ($[M+H]^+$/$[M+Na]^+$) within a user-defined mass accuracy, detecting halogen signatures (Cl/Br count $\geq$1), and comparing isotopic patterns using cosine similarity between experimental and theoretical intensities for further filtering with a user-defined threshold (default: 0.96). The cosine similarity is defined as Eq. (5):

$$\text{cosine\_similarity} = \frac{I_{\text{exp}} \cdot I_{\text{thre}}}{\|I_{\text{exp}}\| \cdot \|I_{\text{thre}}\|} \quad (5)$$

where I_{exp} and I_{thre} represent the experimental and theoretical isotopic peak intensities, respectively. Additionally, if molecular networking is performed *via* GNPS platform[19], the pipeline also writes the prediction and MS^1-based dereplication results into the network file output by GNPS, which subsequently can be visualized using Cytoscape[32].

2.3. Performance evaluation of DeepHalo workflow

We systematically evaluated the DeepHalo workflow using both simulated and real-world datasets. For the simulated data (SM1820-Base and SM1820-R20K) with complete ground truth, we defined true positives (TP) as correctly identified halogenates, false positives (FP) as both misclassified non-halogenates and artificial features predicted as positive, false negatives (FN) as undetected or misclassified halogenates, and true negatives (TN) as correctly classified non-halogenates (excluding false features). For three real-world LC−HRMS datasets, metrics were computed on curated subsets with confirmed ground truth ($n = 132$ for INST, $n = 264$ for INST_SCM, and $n = 500$ for CASMI 2022), where TP represented correctly identified validated halogenated features, FP denoted misclassified validated non-halogenates, FN included undetected or misclassified validated halogenates, and TN comprised properly classified validated non-halogenates. From these values, we computed recall, precision, F1-score, and false positive rate (FPR) using the standard formulas provided in the scikit-learn documentation. ChloroDBPFinder was not included in the SM1820-R20K analysis due to its inability to automatically process LC−MS data without tandem mass spectrometry. Additionally, due to the extremely long running time of competing methods (>12 h), only DeepHalo was evaluated on the CASMI 2022 dataset ($n = 145$ LC−HRMS data). The evaluations were performed on a Windows 10 system equipped with a 12th Gen Intel Core i7-12700F processor, running at 2.10 GHz with 12 cores, and 64 GB of RAM, accelerated by NVIDIA RTX A4500.

2.4. High-throughput mining of halogenated metabolites

Fifty-three *Streptomyces* and one *Micromonospora* strains (Supporting Information Table S2) were each cultured in 24 different media (Supporting Information Table S3), preprocessed, and analyzed using UPLC−HRMS/MS as previously described[33]. Including 60 blank culture media, a total of 1356 UPLC−HRMS/MS data were collected. The raw data were converted to. mzML format using MSconvert and automatically analyzed with DeepHalo's 'detect' function. The minimum peak intensity was set to 1000, signal-to-noise ratio to 6, minimum intensity to $1e^5$, and H-score threshold to 0.4, with feature validation through mass difference and isotopic peak intensities (Supporting Information Table S4). The analysis was performed using the same computer configuration as previously employed for DeepHalo evaluation, with the complete process requiring about 4.9 h. The output MS^2 data from DeepHalo were uploaded to the GNPS server for analysis using classic molecular networking with a precursor and fragment *m/z* tolerance of 0.02, a cosine score of 0.6, and a minimum of 5 fragments, with other parameters set to default. After that, the dereplication function of DeepHalo was used to dereplicate based on NPAtlas (bacterial-derived data only) and FunGBE[34] database, then write the predictions and dereplication results to the GNPS output network file. The final results were visualized using Cytoscape V3.8[32].

2.5. Investigation of halogenase distribution

The draft genomes of 53 *Streptomyces* strains were sequenced on an Illumina HiSeq platform (Illumina, San Diego, CA, USA) and assembled with SPAdes v3.13.1. Candidate halogenases were identified using HMMER 3.1b2. For FDHs, the queries included Pfam profiles PF04820 (Trp halogenase)[35] and PF22045 (chloramphenicol halogenase, halogenating an alkyl group)[36] together with sequences of characterized FDH with diverse substrates (PrnA: Trp, C-7; RebH: Trp, C-7; PyrH: Trp, C-5; SttH: Trp, C-6; Thal: Trp, C-6; PltA: pyrroly-PCP; Bmp2: pyrroly-PCP; Mpy16: pyrroly-PCP; NapH2: naphthoquinone precursor; PltM: diverse phenolic compounds; SgcC3: tyrosyl-PCP; BhaA: β-hydroxy tyrosyl-PCP; VhaA: hexapeptide-peptidyl-PCP; ClmS: acetyl-CoA; CalO3: 2-methoxyorsellinate-PCP; VirX1: iodinase, diverse substrates)[4], while for KDHs, the queries comprised Pfam profiles PF22814 (carrier-protein-independent halogenase WelO5)[37] and PF23169 (halogenase D)[38] along with reference KDH sequences (KthP: Piz-PCP, CytC3: aminobutyryl-PCP, SyrB2: L-Thr-PCP; KtzD: L-isoleucine-PCP; BesD: L-lysine, C-4; HalB: L-lysine, C-5; and HalD: L-ornithine)[4]. A strict E-value cutoff of $1e^{-20}$ was applied. The putative halogenases, together with known ones, were then analyzed for sequence similarity, and SS networks (SSNs) were constructed using EFI-EST[39]. The resulting SSNs were visualized with Cytoscape[32]. The amino acid sequences of the corresponding proteins for 53 in-house Streptomyces strains have been deposited in GenBank, and the accession numbers are provided in Supporting Information Table SD2.

All 2110 *Streptomyces* genomic sequences (as of January 14, 2021) were downloaded from the National Center for Biotechnology Information. Using the same method described above, potential halogenases were identified in these genomes, and their corresponding accessions have been archived in Supporting Information Table SD3.

2.6. Acquisition of aglomycin A

Three strains with potential halogenate production, based on DeepHalo analysis, were further cultured in 500 mL Erlenmeyer flasks and analyzed using UPLC−LCMS/MS and DeepHalo to confirm the presence of the target halogenates. While the selected three strains all produced the target compounds, only *Streptomyces* sp. 11-23 yielded enough for further study. Previous studies have shown that *N*-acetylglucosamine (GlcNAc) can enhance secondary metabolite yields[40], and that NaCl can boost the production of certain halogenated compounds[41,42]. Thus, various concentrations of GlcNAc and NaCl were supplemented to M3 media to test their effects on the production of major components, and the optimal media was used as culture media. The strain 11-23 was cultured on ISP2 plates at 28 °C for 7 days. For seed culture preparation, a mycelial agar plug (1 cm^2) was inoculated into each of five 500 mL Erlenmeyer flasks containing 100 mL of ISP2

medium and incubated at 28 °C on a rotary shaker at 220 rpm for 2 days. For large-scale fermentation, 5 mL of the seed culture was added to each of two hundred 500 mL Erlenmeyer flasks containing 100 mL of culture medium and cultivated at room temperature for 9 days.

The fermentation broth (19 L) was collected and adsorbed onto 1.5 L of microporous adsorption resin X1180. The column was sequentially rinsed with deionized water, 30%, 50%, 70%, and 100% ethanol−water (*v*/*v*) solutions. Each gradient was eluted until the effluent was colorless, resulting in fractions Fr. A, Fr.B, Fr.C, and Fr.D. Fr. D (260 mg) was further fractionated using reversed-phase C_{18} flash column chromatography with a stepwise gradient of MeCN−H_2O, yielding five subfractions (Fr.D1−D5). Subfraction D3 (46 mg) was subjected to semipreparative HPLC using a YMC-Pack C_8 column, yielding aglomycin A (**1**) (t_R 41.2 min; 7.5 mg).

2.7. Identification of aglomycin biosynthesis gene cluster

The complete genome of strain 11-23 was sequenced using PacBio HiFi technology, assembled with Hifiasm, circularized using Circlator v1.5.5, and polished with Illumina data using Pilon v1.22. Secondary metabolite biosynthetic gene clusters (BGCs) were analyzed with AntiSMASH[43]. The complete aglomycin BGC has been deposited into GenBank under accession number PV571899. The identities and similarities of the encoded proteins to their homologs were examined using BLASTP.

To confirm the BGC, genes *aglA*, *aglB*, and *aglG* were inactivated by homologous recombination. Upstream and downstream homologous fragments were amplified by PCR, sequenced, and cloned into the temperature-sensitive plasmid pKC1139 to create in-frame deletion plasmids. These plasmids were introduced into *Streptomyces* sp. 11-23 by conjugation with *E. coli* ET12567/pUZ8002. After conjugation, conjugants were selected on MS medium with aztreonam and apramycin. Single-crossover recombinants were cultivated at 37 °C, followed by seven rounds of propagation without antibiotics at 28 °C to obtain double-crossover mutants Δ*aglA*, Δ*aglB*, and Δ*aglG*, confirmed by PCR. For gene complementation, coding regions of *aglB* and *aglG* were amplified and cloned into pL646 to obtain complement plasmids, which were introduced into Δ*aglB* and Δ*aglG* strains by conjugation. Correct recombinants were confirmed by PCR. These strains were selected by their apramycin-resistant phenotype and verified by PCR. The metabolites of engineered strains were analyzed using the same method as high-throughput mining of halogenates as described above.

2.8. MIC determination

The MIC values of aglomycins were determined using the broth microdilution method[44]. *Enterococcus faecium* ATCC 35667, *Staphylococcus aureus* ATCC 29213, *Enterococcus faecalis* (clinical isolates), *Escherichia coli* ATCC 25922, *Klebsiella pneumoniae* ATCC BAA2470, *Acinetobacter baumannii* ATCC 19606, *Pseudomonas aeruginosa* ATCC 27853 and 9 clinical isolates of *Enterococci* were cultured overnight at 37 °C on TSA plates. The bacterial suspensions were diluted to a concentration of 5×10^5 CFU/mL using fresh MHB medium, and 100 μL of each suspension was transferred into the wells of a 96-well microplate. Aglomycins, dissolved in DMSO at a concentration of 6.4 μg/μL, were added at 2 μL per well in the first column containing the bacterial suspension. Serial two-fold dilutions were then performed to achieve a concentration range for aglomycins from 0.0625 to 64 μg/mL. The plates were incubated at 30 °C for 16−18 h, and the MIC was defined as the lowest concentration of aglomycins at which no visible bacterial growth was observed.

2.9. Cytotoxicity assay

The cytotoxic effects of compounds were determined on various cell lines, including Huh7 (Human hepatocellular carcinoma), HEK293T (human embryonic kidney) and Vero E6 (monkey kidney). Cells were seeded into 96-well plates at a cell density of 10,000 cells per well for Huh7, HEK293T, and Vero E6 cells. After 24 h, the test compounds, with an initial concentration of 64 μg/mL, were serially diluted twofold for eight dilutions and co-incubated with cells for 24 h. The cytotoxic effects of compounds on various cells were measured by adding 10 μL cell counting kit-8 (CCK-8) solution (Vazyme Biotech) and incubating for 1 h. The absorbance value was determined at 450 nm in a Victor X5 Plate Reader (PerkinElmer, Waltham, MA, USA). Cell viability was expressed as a percentage of vehicle control, and 50% cytotoxic concentration (CC_{50}) was calculated using GraphPad Prism 8.0.

2.10. Checkerboard assay for determining synergistic activity

E. faecium-VRE-1-*van*A was cultured overnight at 37 °C on TSA plates. The bacterial suspension was diluted to a concentration of 5×10^5 CFU/mL using fresh MHB medium, and 100 μL of the suspension was added to each well. A checkerboard dilution method was employed to calculate the required final concentration for each well, and the appropriate compounds were added. After thorough mixing, the plates were incubated at 30 °C for 18 h then analyzed at 600 nm using a Tecan Infinite Spark microplate reader (Tecan, Switzerland). A blank control (MH medium without bacteria) and a positive control (MH medium with bacteria but without antibacterial compounds) were included. Fractional inhibitory concentration (FIC) indices were used to assess the synergistic effects of two antibacterial compounds[45]. The FIC index value was calculated using a previously defined Eq. (6) [46]:

$$\text{FIC index} = \frac{\text{MIC}_{\text{a in combination}}}{\text{MIC}_{\text{a alone}}} + \frac{\text{MIC}_{\text{b in combination}}}{\text{MIC}_{\text{b alone}}} \quad (6)$$

where $\text{MIC}_{\text{a in combination}}$ and $\text{MIC}_{\text{b in combination}}$ represent the MICs of each antibiotic when used in combination (within a single well), $\text{MIC}_{\text{a alone}}$ and $\text{MIC}_{\text{b alone}}$ denote the MICs of each drug individually. The inhibition growth rate was calculated using Eq. (7):

$$\text{Inhibition growth rate}\ (\%) = \left(\frac{\text{OD}_{600}^{\text{positive}} - \text{OD}_{600}^{\text{sample}}}{\text{OD}_{600}^{\text{positive}} - \text{OD}_{600}^{\text{blank}}}\right) \times 100 \quad (7)$$

where OD_{600} is the optical density at 600 nm. The MIC was defined as the minimum concentration required to inhibit 95% of bacteria. The synergistic effect is indicated by an FIC index below 0.5[47].

2.11. Time-dependent killing assay

We assessed the antimicrobial mode using the time-dependent killing assay. *E. faecium* ATCC 35667 was cultured in TSB medium at 220 rpm and 37 °C to reach the exponential growth phase, then was adjusted to 5×10^5 CFU/mL in MHB medium, with 200 μL added to each well of a 96-well microplate. The stock

solution of each compound in DMSO (0.8 or 3.2 mg/mL for aglomycin A and 0.1 or 0.4 mg/mL for linezolid) was added either individually or in combination to the bacterial suspension, achieving final concentrations equivalent to 0.5−8 × the MIC of the respective compound. A blank control was established by adding 2 μL of DMSO to the bacterial suspension. The 96-well plate was placed in a 30 °C incubator for static incubation, with 20 μL of each sample taken every 2 h and diluted 10,000-fold for colony counting.

2.12. *In vivo antimicrobial efficacy assay*

We evaluated the *in vivo* antimicrobial activity using a *Galleria mellonella* model. First, *E. faecium* ATCC 35667 was inoculated in LB medium and incubated at 37 °C for 24 h. The culture was centrifuged and washed with saline to collect the bacterial cells. Then the bacterial suspension was adjusted to 8×10^9 CFU/mL using saline. For each group, 20 healthy, milky-white last-instar larvae of *G. mellonella* were selected. A 5 μL aliquot of the bacterial suspension was injected into the hemolymph of each larva using a Hamilton syringe (7654-01, Hamilton) equipped with a 33G RN needle (10 mm, Hamilton). The larvae were then incubated in a dark environment at 37 °C and 50% humidity for 1 h. Subsequently, aglomycin A, linezolid, or a control solution of physiological saline containing 9.3% DMSO were separately injected into the infected larvae. Uninfected larvae were also included as a blank control group. The larvae were maintained in an incubator, and their survival was monitored every 24 h. Larvae that showed no response to needle stimulation were considered dead. The experiment lasted five days, and survival rates were recorded. Kaplan−Meier survival curves were then plotted. Cox regression analysis was performed using R software (version 4.2.2), along with MSTATA software (https://www.mstata.com/).

2.13. *Data and code availability*

The simulated isotope patterns were generated from six open compound databases: COCONUT (https://coconut.naturalproducts.net/, Version January 2022), NPAtlas (https://www.npatlas.org/, V2022_09), ChEBI (https://www.ebi.ac.uk/chebi/, accessed 1 May 2023), HMDB 5.0 (https://hmdb.ca/downloads, accessed 30 May 2023), DSSTox (https://figshare.com/search?q=CFM-ID_metadata_DTXCID, accessed 25 May 2023), and ZINC20 (https://zinc.docking.org/). The resulting datasets (IsoBase, IsoHN, and IsoBase_Val_SN_Eva) used for training and validation are archived at https://doi.org/10.5281/zenodo.17336407. Real isotope patterns were obtained from open datasets including the myxo dataset (https://bio.informatik.uni-jena.de/software/sirius/) and the CASMI 2016 dataset (http://casmi-contest.org/2016/). The real LC−HRMS datasets for DeepHalo evaluation are from CASMI 2022 (https://fiehnlab.ucdavis.edu/casmi) and in-house datasets INST (MassIVE MSV000098203) and INST_SCM (MassIVE MSV000098204). The dataset used for high-throughput mining of halogenated metabolites is available *via* MassIVE MSV000099431.

The DeepHalo is available at https://github.com/xieyying/deephalo and https://pan.baidu.com/s/1KV5IFEBx-4BX-CDsd9dCsA?pwd=deep; The mzml2gnps have been integrated into DeepHalo and also can be found at https://github.com/xieyying/mzml2gnps; The modified SMITER can be found at https://github.com/xieyying/SMITER_modified.

3. Results

3.1. *Establishment and comparative evaluation of DeepHalo*

DeepHalo was designed for automatic and efficient high-throughput mining of new halogenates from complex samples. To achieve this, we integrated four main functions into DeepHalo: precise detection of isotope patterns, confident halogen prediction, comprehensive scoring, and efficient dereplication (Fig. 1A).

The core function of DeepHalo is to accurately predict the presence of halogens (Cl/Br) based on the isotope patterns of precursor ions (Fig. 2). Toward this goal, DeepHalo employs a deep learning-based Element Prediction Model (EPM), trained on 1.09 million isotope patterns (Supporting Information Fig. S1). To ensure the reliability of EPM, we implemented a series of optimization strategies. First, we generated synthetic isotope patterns to simulate co-eluting dehydrogen isomers (Fig. S1) to enhance the model's ability to filter out this common type of false signals. Second, we incorporated uncommon elements that frequently appear in natural products (Fig. 2A and Supporting Information Fig. S2), not only to reduce false positive predictions for halogens but also to extend the model's capability to detect other uncommon elements. Third, we developed a dual-branch Isotope Neural Network (IsoNN) that processes mass and intensity features of isotope peaks in parallel, to effectively address scale differences and allow for independent noise injection and scaling adjustments (Fig. 2B). Finally, by combining manual tuning with Bayesian optimization, we efficiently optimized the model in just six iterations (Supporting Information Fig. S3). The optimized EPM took relative intensities of the first five isotope peaks, along with the mass differences, Δm1 and Δm2, as input features, and output an 8-dimensional vector representing the predicted probabilities for the eight target categories (Fig. 2B). These systematic enhancements enabled the EPM to achieve state-of-the-art performance in halogen prediction, as demonstrated by comparative evaluations on both simulated and real isotopic datasets (Table 1). Notably, our EPM was the only method to achieve 100% precision and recall, with zero FPR, for the binary classification of halogenated versus non-halogenated compounds on the tested real-world dataset. Even for the more detailed 8-category classification, EPM achieved 99.8% accuracy (622/623 correct), with only one Cl-type molecule misclassified as a Br-type (Fig. 2D). Additionally, the comparative evaluation revealed a clear performance hierarchy (Table 1): deep learning-based models (our EPM and SIRIUS) significantly outperformed traditional machine learning-based tools (ChloroDBPFinder and DCAnalysis) and rule-based methods (HaloSeeker and MeHaloCoA), highlighting the superior capability of deep learning in resolving the complex, element-specific isotope patterns.

The performance was evaluated for the binary classification of halogenated *versus* non-halogenated compounds. Our DNN-based element prediction model (EPM) exhibited superior performance across both simulated and real datasets, significantly outperforming other tools (SIRIUS 4.0[16], HaloSeeker v2.0.3.3[13], MeHaloCoA v0.99.0[11], ChloroDBPFinder v0.1.0[14], and DCAnalysis v1.10[12]).

On the other hand, accurately detecting isotopic features from untargeted mass spectrometry data is a complex yet crucial step in the analytical workflow[48]. In the DeepHalo workflow, we employed PyOpenMS[18] for feature detection and then applied a robust two-dimensional validation process that independently verifies both the mass and intensity dimensions. First, isotope

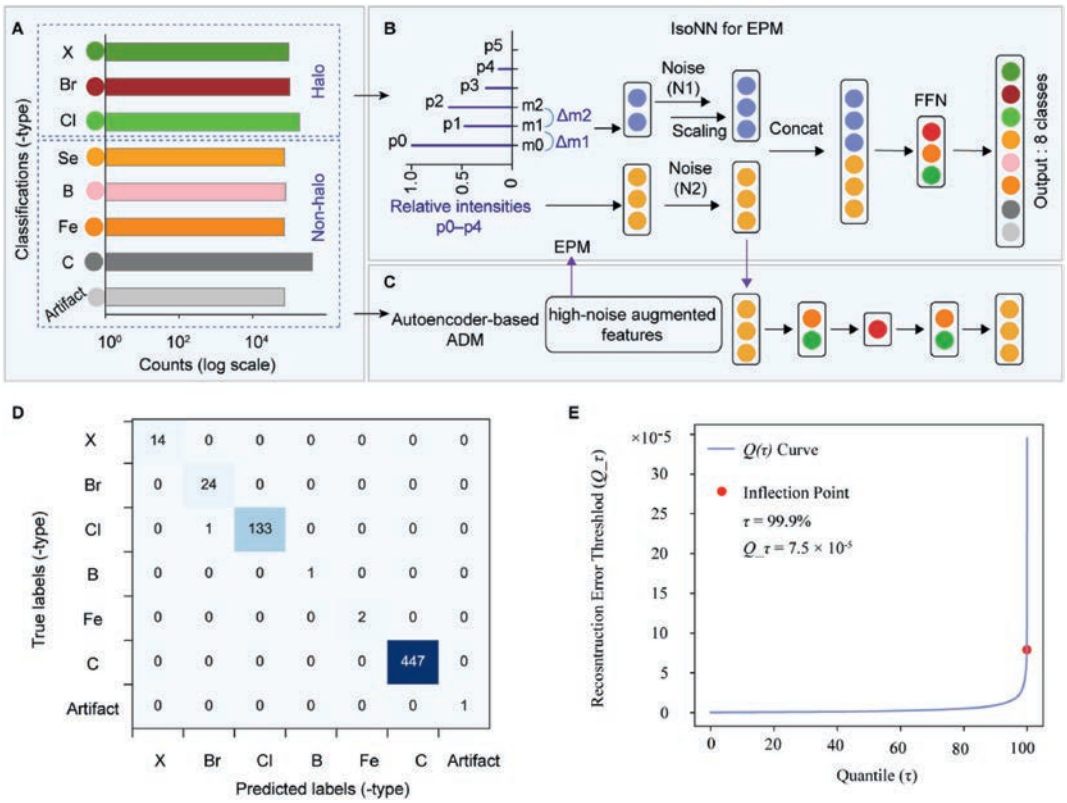

Figure 2 Establishment of deep learning models in the DeepHalo pipeline. (A) Class distributions of isotope patterns used for training models (all types for the Element Prediction Model (EPM); X/Cl/Br/C-types for Anomalous Isotope Patterns Detection Model (ADM). Cl-type: Cl/Cl2; Br-type: Br/Cl3; X-type: mixed/polyhalogenated; B-type: boron-containing; Se-type: selenium-containing, and Fe-type: iron-containing; C-type: others; Artifact-type: simulated isotope patterns for co-eluting dehydrogen isomers, (B, C) Network architectures: IsoNN for EPM (B) and autoencoder for ADM. (D) Eight-category classification results of EPM tested on the CASMI2016_Myxo_plus dataset. (E) Determination of ADM anomaly detection threshold (7.915×10^{-5}). The optimal threshold was determined from the inflection point of the reconstruction error threshold-versus-percentile curve generated by ADM using the IsoHN dataset.

Table 1 Performance comparison of methods for halogen prediction.

Type	Method	TP	TN	FP	FN	Recall	Precision	F1-score	FPR
Simulated isotope pattern dataset (IsoBase_Val_SN_Eva)									
Ours	EPM	65,225	75,356	60	100	0.9985	0.9991	0.9988	0.0008
DNN-based	SIRIUS 4.0	65,270	74,083	1333	55	0.9992	0.9800	0.9895	0.0177
ML-based	ChloroDBPFinder	58,522	71,279	4137	6803	0.8959	0.9340	0.9145	0.0549
Rule-based	MeHaloCoA	65,167	66,284	9128	158	0.9976	0.8771	0.9335	0.1210
	HaloSeeker	55,847	65,501	9907	9478	0.8549	0.8493	0.8521	0.1314
Real isotope pattern dataset (CASMI2016_Myxo_plus)									
Ours	EPM	172	451	0	0	1.0000	1.0000	1.0000	0.0000
DNN-based	SIRIUS 4.0	172	443	8	0	1.0000	0.9556	0.9773	0.0177
ML-based	ChloroDBPFinder	163	447	4	9	0.9477	0.9760	0.9616	0.0089
	DCAnalysis	133	451	0	39	0.7733	1.0000	0.8721	0.0000
Rule-based	HaloSeeker	169	449	2	3	0.9826	0.9883	0.9826	0.0044
	MeHaloCoA	172	448	3	0	1.0000	0.9829	0.9885	0.0067

patterns are corrected and validated by checking the mass differences between the first three peaks (Supporting Information). Only those with Δm1 and Δm2 within the 0.05th to 95.5th percentiles of the IsoHN dataset (Supporting Information Fig. S5), a high-noise augmented dataset containing halogenated (Cl/Br/X-type) and non-halogenated (C-type only) patterns, are retained for further analysis. Subsequently, for validation of isotopic intensities, we established a deep autoencoder-based[49] anomalous isotope pattern detection model (ADM) (Fig. 2C). Using this model, we computed reconstruction errors for intensities of the first five isotope peaks, which served to distinguish normal from anomalous distributions. The optimal discrimination threshold was set at

7.915×10^{-5}, determined from the inflection point on the reconstruction error threshold-*versus*-percentile curve generated by ADM using the IsoHN dataset (Fig. 2E).

Additionally, we introduced a halogen confidence score (H-score) as an indicator of the likelihood that a feature contains halogens. This metric comprehensively evaluates predictions at both the centroid and scan levels. To achieve this, we adapted PyOpenMS to extract isotopic patterns not only for each feature (centroid form) but also for each scan (Fig. 1A). Subsequently, predictions were performed based on isotopic patterns of both levels, and the results were integrated for the calculation of H-score (Supporting Information). This hierarchical assessment approach is a key distinguishing feature of DeepHalo, which not only enhances the reliability of halogen predictions but also allows flexible filtering of target features by adjusting the H-score threshold during analysis. Coupled with visualization tools, the H-score enables rapid differentiation and targeted annotation of halogenated candidates, even in highly complex datasets (Supporting Information Fig. S6).

Automatic dereplication is essential for efficiently annotating known compounds and discovering novel ones from complex samples. In the DeepHalo pipeline, dereplication is achieved through two complementary methods (Fig. 1A). 1) MS^1-based dereplication: If the user provides a chemistry database, the pipeline utilizes MS^1 data for dereplication by verifying the exact *m*/*z* of precursor ions, assessing the presence of Cl/Br, and calculating the cosine similarity between experimental and theoretical intensities. 2) MS^2-based dereplication: If LC−HRMS data are acquired in DDA mode, the pipeline optionally exports the MS^2 spectra of halogenated features based on a specified H-score threshold for further dereplication *via* molecular networking on the GNPS platform[19]. By leveraging the propagation of molecular networks, this method not only utilizes reference MS^2 spectra from GNPS for dereplication but also eliminates false positives caused by the biotransformation of halogenates in blank matrices. Finally, all results are compiled into the network file output by GNPS and visualized using Cytoscape[32], enabling intuitive and effective analysis of all outcomes (Fig. S6). By combining these two methods, known compounds can be efficiently and accurately annotated, while potential new halogenated compounds are identified for further investigation.

Computational efficiency is a critical consideration for large-scale dataset analysis. To address this challenge, we employed a two-tier optimization strategy. First, we integrated PyOpenMS with KD-tree spatial indexing and concurrent. futures-based parallel processing, achieving an order-of-magnitude reduction in per-sample analysis time. Second, we unified all analytical steps (except dereplication) within a single "detect" function, thereby enabling fully automated processing of arbitrarily large LC−HRMS datasets, a capability essential for high-throughput applications.

Following the establishment of DeepHalo, we systematically evaluated the workflow using both simulated and real-world LC−HRMS datasets. First, we established a simulated LC−MS dataset SM1820-Base (Supporting Information) and used it to evaluate the effects of mass spectrometry parameters and isotope pattern validation on DeepHalo's performance. The results (Fig. 3A−F) indicate that higher resolution and accuracy yield higher performance, with a resolution of 20,000 and an accuracy of 3 ppm, typically achieved by high-resolution mass spectrometers, providing acceptable results. Although isotope pattern validation slightly reduces recalls, it substantially improves precisions and, consequently, the F1-scores. Additionally, we employed simulated SM1820-R20K and experimental datasets, including INSD, INSD_SCM, and CASMI 2022, to systematically assay the influence of the H-score threshold on DeepHalo. The results show that the sensitivity of the H-score threshold increases as data quality decreases, and a threshold of 0.4 proves optimal for balancing recall and precision (Fig. 3G−I). Next, we compared the performance of DeepHalo with state-of-the-art comparable tools (Table 2). Across all benchmarks, DeepHalo demonstrated consistent superiority, achieving the highest precision and recall, the lowest FPR, and significantly reduced computational time. Furthermore, its efficiency advantage becomes more pronounced as dataset size increases. For example, with the 145-sample CASMI 2022 dataset, competing methods failed to process the data within 12 h, whereas DeepHalo completed the analysis in just 16.7 min with perfect accuracy and recall (100%).

3.2. DeepHalo-assisted mining reveals a wealth of halogenated metabolites in Streptomyces

To demonstrate the efficiency of DeepHalo, we applied it to explore new halogenated compounds from 1296 HRMS-based metabolomic datasets derived from 53 *Streptomyces* and one *Micromonospora* strain cultures (Supporting Information). Initially, using the fully automatic 'detect' function of DeepHalo with H-score threshold set to 0.4, we identified and extracted potential features with halogen and their corresponding MS^2 spectra from untargeted UPLC−HRMS/MS data of the crude extracts and 60 blank media controls. Following this, the similarity of the obtained MS^2 spectra was analyzed using the molecular networking tool on the GNPS platform[19]. Subsequently, a deeper dereplication process was performed using the 'dereplicate' function of DeepHalo, employing NPAtlas[10] (bacterial-derived data only) and FunGBE[34] as dereplication databases. Ultimately, the results were visualized using Cytoscape for further analysis.

Based on the resulting molecular networks (MNs) (Fig. 4A), blank media-derived variants (Group 1) were readily excluded according to the propagating nature of MNs. Singletons in Group 6 were then also recognized as media-derived variants based on careful examination of their precursors' *m*/*z* and retention times. Concurrently, potential known compounds annotated using the user-provided database, highlighted as rectangles ($[M+H]^+$) or hexagons ($[M+Na]^+$) in Fig. 4A, were efficiently identified. Consequently, five groups of known halogenated metabolites, svetamycins A-C, colibrimycin A1, salinamide B, lysolipin I, and benzastatin C, were detected from five strains, with one group (lysolipin I) confirmed *via* GNPS MS^2 spectral matching (Fig. 4B). The structures of all these halogenates were subsequently confirmed through bioinformatics analysis combined with MS^2 spectra or UV spectroscopy (Supporting Information Figs. S7-S10). Intriguingly, four of these five classes of compounds exhibit antibacterial or antiviral activities, with the chloride atom in benzastatin C proven to be critical for its antiviral activity[50]. As expected, these compounds display significant structural diversity while undergoing halogenation *via* distinct mechanisms. For example, the lipopeptide colibrimycin A1[51] and the aromatic polyketide antibiotic lysolipin I[52] are halogenated through FDH-mediated electrophilic attack on activated sp^2 carbons, while depsipeptide antibiotic svetamycins[53,54] feature KDH-catalyzed radical halogenation of inactivated sp^3 carbons. In contrast, the chloride atom in the depsipeptide antibiotic salinamide B[55] is

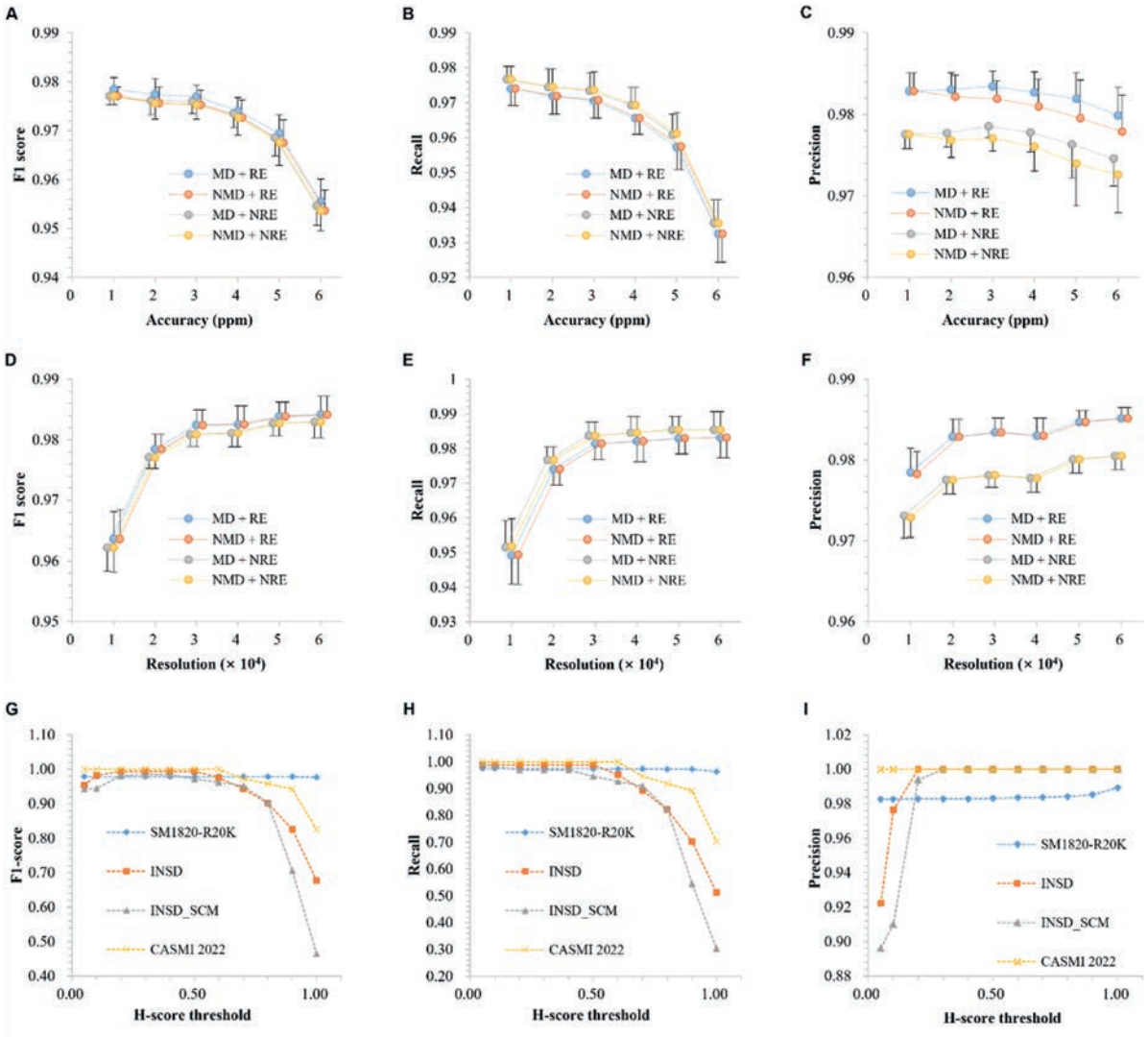

Figure 3 DeepHalo's performance sensitivity to mass spectrometry parameters and H-score thresholds. The effects of MS accuracy (A−C) and resolution (D−F), as well as H-score threshold (G−I) on Recalls, Precisions, and F1-scores. Key settings for DeepHalo analysis: MD/NMD: mass difference validation applied *vs*. not applied. RE/NRE: Reconstruction error (ADM-based) used *vs*. not used. Panels (A−F): $n = 10$, using simulated dataset SM1820-R20K. Panels (G−I): all evaluations performed under the MD + RE setting using simulated SM1820-R20K and experimental datasets (INSD, INSD_SCM, CASMI 2022). The results demonstrate that isotope pattern validation (MD and RE) slightly reduces recall but significantly improves precision and F1-scores. Furthermore, threshold sensitivity increases with lower data quality, and a threshold of 0.4 provides an optimal balance between recall and precision.

introduced *via* non-enzymatic epoxide opening, and most intriguingly, the halogenation of benzastatin derivative benzastatin C[56] involves an elusive cytochrome P450-catalyzed mechanism. Collectively, these results demonstrate DeepHalo's efficacy in identifying structurally diverse halogenated metabolites, as well as its unique capability to uncover atypical halogenation pathways.

Further analysis revealed 11 additional strains producing halogenated metabolites (Fig. 4A, Supporting Information Table S5), bringing the total to 16 strains, 30% of the 54 strains examined, which is considerably higher than previously believed[7,8]. Earlier studies, relying on PCR-based screening of flavin-dependent halogenases in actinomycetes strains, reported a detection rate of ∼20%[7,8]. Since most of our tested strains are *Streptomyces* spp., this discrepancy suggests that the potential for halogenated metabolite production in this genus may have been significantly underestimated. To further explore this hidden biosynthetic capacity, we sequenced the genomes of all 53 *Streptomyces* strains analyzed in this study. Using these sequences, we systematically investigated the distribution of halogenases by probing for two major classes: FDHs and KDHs using HMMer[57]. Strikingly, more than 79% (42 out of 53) of the strains encoded either FDHs or KDHs, with 36 strains harboring 82 FDHs and 17 strains containing 17 KDHs (Supporting Information Table SD2). Sequence similarity network (SSN) analysis revealed significant diversity among these halogenases, exhibiting broad substrate specificities (Supporting Information Fig. S11). To ascertain whether the high recovery rate of halogenases was universal or unique to our Tibetan strains, we analyzed potential halogenases in 2110 *Streptomyces* genomic sequences downloaded from NCBI (as of January 14, 2021). The

Table 2 Performance comparison of our DeepHalo workflow with other counterpart methods.

Methods	TP	TN	FP	FN	Recall	Precision	F1-score	FPR	Time (min)[a]
	Simulated LCMS dataset (SM1820-R20K)[b]								
DeepHalo	4471	12,991	78	119	0.9741	0.9829	0.9784	0.0060	7.3
HaloSeeker	4421	13,586	435	169	0.9632	0.9104	0.9361	0.0310	47
MeHaloCoA	3712	12	1206	878	0.8087	0.7548	0.7808	0.9901	90
	Real LCMS dataset (INSD)[c]								
DeepHalo	83	48	0	1	0.9881	1.0000	0.9940	0.0000	1.6
HaloSeeker	27	23	3	40	0.3241	0.9000	0.4737	0.1154	70
MeHaloCoA	76	0	9	2	0.9047	0.8941	0.8994	1.0000	170
ChloroDBPFinder	49	48	0	31	0.5833	1.0000	0.7368	0.0000	243
	Real LCMS dataset (INSD_SCM)[d]								
DeepHalo	160	91	0	8	0.9524	1.0000	0.9756	0.0000	9.1
HaloSeeker	39	70	4	101	0.2786	0.9070	0.4262	0.0541	240
MeHaloCoA	146	0	17	22	0.8690	0.8957	0.8822	1.0000	480
ChloroDBPFinder	102	96	0	66	0.6071	1.0000	0.7556	0.0000	380
	Real LC-HRMS dataset (CASMI 2022)[e]								
DeepHalo	37	446	0	0	1.0000	1.0000	1.0000	0.0000	16.7

[a]Pure computational runtime.
[b]Simulated LC−MS dataset derived from 1820 molecules with ten replicates.
[c]In-house analytical standard dataset.
[d]In-house standards spiked in culture media.
[e]Competing methods failed to process the data within 12 h. The DeepHalo pipeline demonstrated superior precision and recall with dramatically reduced computational time, significantly outperforming all comparable tools (Supporting Information Tables SD4-SD7).

results showed that approximately 70% (1476 out of 2110) of the strains harbored halogenases, with 1324 strains containing 2297 FDHs and 388 containing 430 KDHs (Supporting Information Table SD3), consistent with the discovery rate observed in our in-house strains. These findings underscore the widespread presence of halogenases in the genus *Streptomyces* and point to a vast, untapped reservoir of halogenated natural products.

3.3. Identification of a new class of halogenated depsipeptides

We prioritized the remaining potential halogenated natural products based on their yields and H-scores. Consequently, three strains (11-23, 19-293, and 19-259) were selected for repeated fermentation. While all three strains produced the target compounds (Fig. 4C), only strain 11-23 yielded enough for further study. After optimizing the fermentation conditions (Supporting Information Fig. S12), a scale-up fermentation combined with MS-guided targeted isolation yielded aglomycins A (**1**) and B (**2**).

Aglomycin A (**1**) was evidenced to have a molecular formula of $C_{35}H_{45}ClN_8O_8S$ from HRESI(+)MS and ^{13}C NMR data, requiring 17 double bond equivalents (DBE). Acquisition of 1D and 2D NMR data of **1** (Supporting Information Table S6 and Figs. S13-S20) and detailed analysis suggested the presence of one threonine (Thr^1), one proline (Pro^2), one piperazic acid (Piz^3), and one *N*-methylvaline (*N*-$MeVal^5$). Additionally, a 3,4-epoxyvaline (epoxy-Val^4) residue was deduced by the $^1H-^1H$ COSY correlation of its NH and H-2 (δ_H 5.22), combined with the key HMBC correlations from H-2 of epoxy-Val^4 to its carbonyl carbon (C-1), H_2-4 to C-2, C-3 (quaternary), and C-5 (isolated methyl), as well as characteristic chemical shifts of epoxy group ($\delta_{C\text{-}3}$ 57.4 and $\delta_{C\text{-}4}$ 49.6). Moreover, based on the $^1H-^1H$ COSY correlation of aromatic H-4′/H-5′/H-6′, and the key HMBC correlations from highly downfield-shifted H-5 (δ_H 8.37, s) to aromatic C-2 and C-4 as well as the carbonyl carbon (C-1″), from NH_2 to C-1′ and C-3′ from H-6′ to C-2, C-1′, C-2′, and from H-4′ to C-3′, combined with characteristic isotope distributions of the chlorine atom in MS (Supporting Information Fig. S14), one 2-(2-amino-3-chlorophenyl)-4-thiazolecarboxylic acid (ACTA) residue was characterized. The connectivity among these six moieties was established by the key HMBC correlations (Fig. 4D) and the MS^2 fragmentations of the sodium adduct ion (Supporting Information Fig. S15), ultimately confirming the planar structure of **1** (Fig. 4D). The amide bond between Pro^2 and Thr^1 was determined to have a *trans* geometry based on the small $\Delta\delta_{C\beta-C\gamma}$ value of 4.2 ppm (Table S6)[58,59], whereas the amide bond between Pro^2 and Piz^3 was inferred to have a *cis* geometry based on the ROESY correlation of H-2-Pro^2 and H-2-Piz^3 (Supporting Information Fig. S21)[60,61]. The absolute configuration of **1** was determined by combining the Marfey's analysis[62] (Supporting Information Supporting Information Figs. S22-S25) with GIAO NMR calculations[63] and DP4+ analysis[64] (Supporting Information Tables S7-S17 and Fig. S26), ultimately confirming its amino acid residues as *S*-Thr^1, *S*-Pro^2, *S*-Piz^3, *S*,*S*-epoxy-Val^4 and *S*-*N*-Me-Val^5.

Compound **2** with a much lower molecular weight ($[M+H]^+$ *m/z* 254.9995) was determined by NMR analyses (Supporting Information Table S18 and Figs. S27-S33) as a dipeptide corresponding to the truncated side chain of **1**. Additionally, two other minor components, aglomycins C (**3**) and D (**4**), were tentatively identified by MS^2 due to their extremely low yields (Supporting Information Fig. S34). Both resemble **1**, except that epoxy-Val^4 in **1** is replaced by dehydro-Val^4 in **3** and dihydroxy-Val^4 in **4** (Fig. 4D).

3.4. Deciphering the biosynthesis of aglomycin A

Structurally, aglomycin A (**1**) is characterized by a cyclic pentapeptide core framework incorporating an exceptionally rare

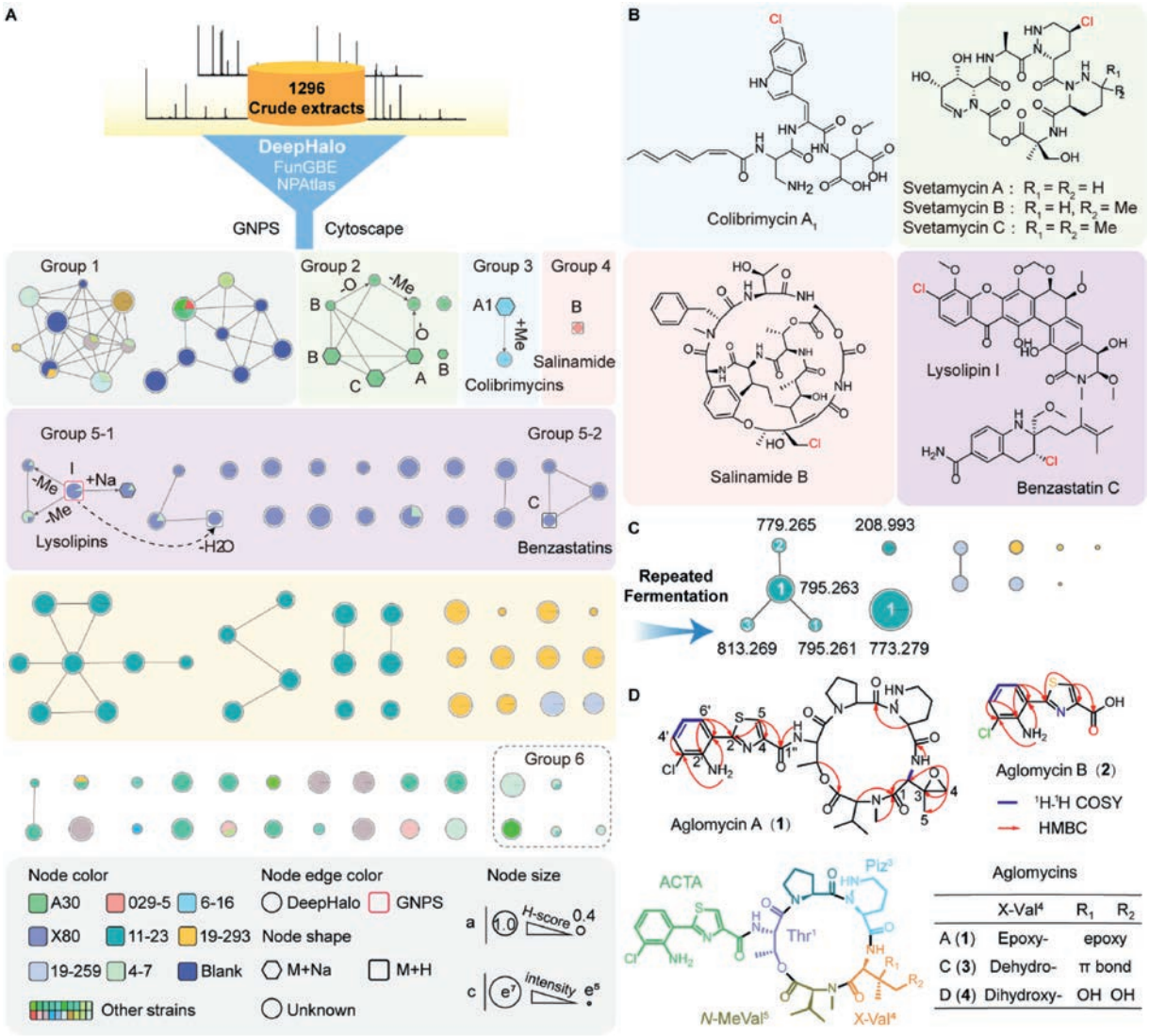

Figure 4 High-throughput mining of halogenated metabolites using DeepHalo. (A) Molecular networks of potential halogenated compounds generated by mining 1296 crude extracts using DeepHalo, revealing that sixteen strains harbor halogenated features (Table S5). (B) The structures of five groups of known halogenated compounds identified by DeepHalo. (C) Molecular networks constructed from the potential halogenated compounds of three strains (11-23, 19-293, and 19-259) in flask-scale cultures reveal a dominant compound, aglomycin A, with an $[M+H]^+$ ion at *m/z* 773.279 and an $[M+Na]^+$ ion at *m/z* 795.263. (D) The structures and key 2D NMR correlations of aglomycins A (**1**) and B (**2**) isolated from *Streptomyces* sp.11-23, together with aglomycins C (**3**) and D (**4**) identified through MS^2 fragmentation.

epoxyvaline residue, as well as a dipeptide side chain formed by the condensation of 3-chloroanthranilic acid (3-Cl-AA) and cysteine. Given the distinctive structure of **1**, the producer strain, *Streptomyces* sp. 11-23, was sequenced to elucidate its biosynthetic pathway. Biogenetically, the 3-Cl-AA moiety is predicted to be derived from tryptophan, which subsequently undergoes halogenation catalyzed by a tryptophan halogenase. Therefore, we utilized the tryptophan halogenase, PrnA[65], as a probe, together with comprehensive bioinformatics analysis, revealing a putative *agl* cluster (Fig. 5A and Supporting Information Table S19). This cluster encodes two discrete giant NRPSs, AlgN and AglJ, holding two and five modules, respectively, expected to be responsible for the construction of the dipeptide side chain and the cyclic pentapeptide core scaffold (Fig. 5B and Supporting Information Table S20). The CAL domain of the first module of AlgN was expected to activate the precursor of 3-Cl-AA and transfer it to the adjacent T domain[66]. However, detailed bioinformatic analysis revealed that CAL harbors a mutation in the A10 motif, where the conserved lysine, crucial for adenylation activity[67], is replaced by proline (Supporting Information Fig. S35). Concurrently, *aglM* encoding an NRPS with a stand-alone A domain was found to be adjacent to *algN*, possibly compensating the adenylation function of the CAL and activate 3-Cl-AA. Intriguingly, a $C_{starter}$ domain, typically found in lipid peptides to condense a fatty acid and the first amino acid residue for initiating the elongation of the peptide chain[68], is located in the first module of AglJ (Fig. 5B and Supporting Information Fig. S36) and proposed to condense dipeptide acyl and threonine acyl to initiate the biosynthesis of cyclic pentapeptide. Notably, the fourth module of NRPS AglJ lacks an A domain and comprises only a C-T didomain, whereas a separate A-T didomain NRPS AglB and a free-standing type II thioesterase (TEII) AglI, were present in the *agl* cluster (Fig. 5B and

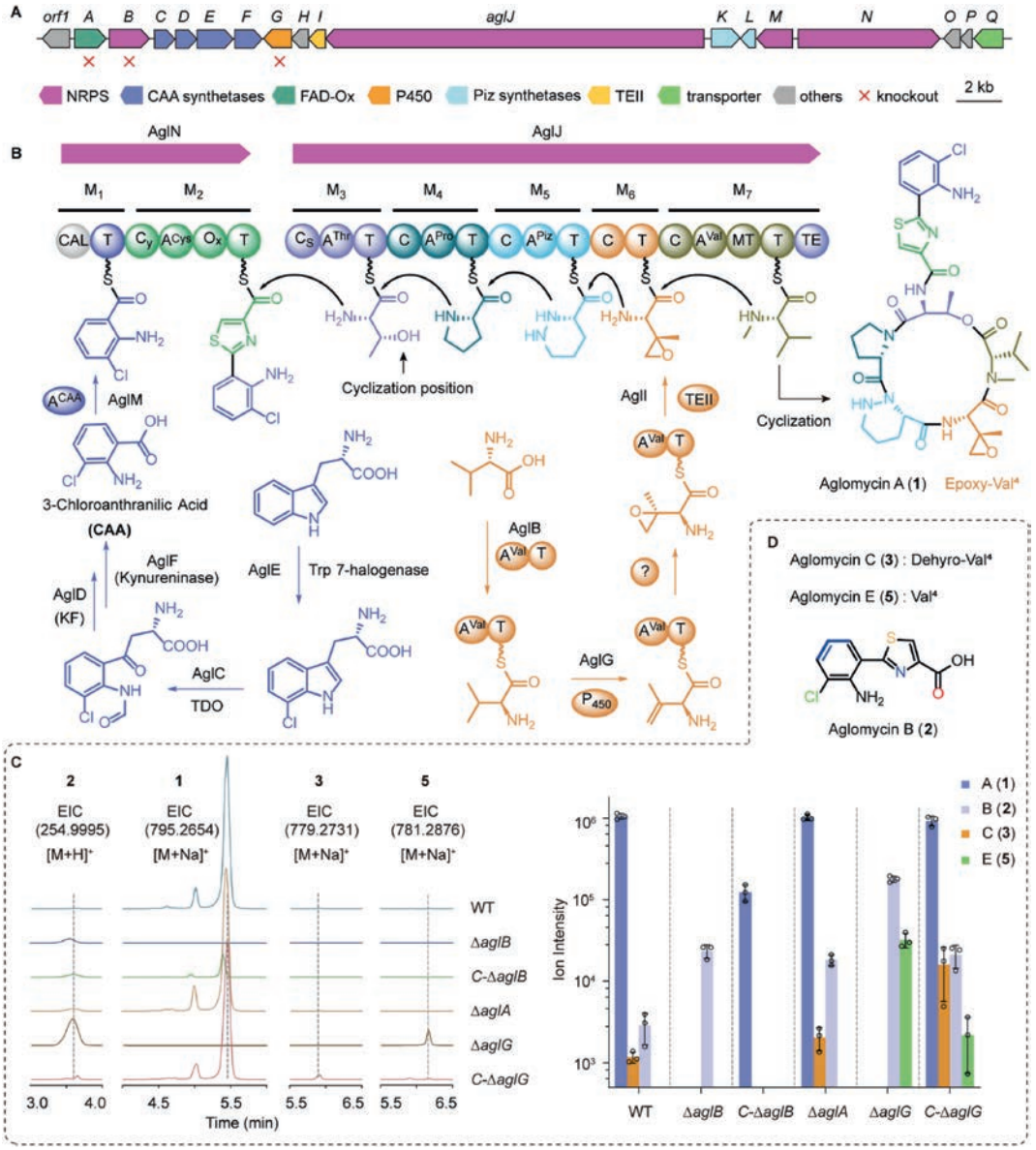

Figure 5 Proposed and identification of aglomycin A biosynthesis. (A) Proposed aglomycin A biosynthetic gene cluster, confirmed by *aglA/B/G* knockout experiments. (B) A plausible aglomycin A biosynthesis pathway. TDO: tryptophan 2,3-dioxygenase, KF: kynurenine formamidase. (C) Extracted ion chromatograms (EICs) of compounds **1**, **2**, **3**, and **5** from wild type (WT), knockout (Δ*aglA*, Δ*aglB*, and Δ*aglG*), and complemented strains (C-Δ*aglB* and C-Δ*aglG*) and their relative yields ($n = 3$). (D) Chemical structures of aglomycin derivatives.

Supporting Information Fig. S37). This modular configuration, reminiscent of the assembly lines of WS9326A[69], legonmycin[70], and rotihibin[71], is predicted to cooperatively incorporate the unusual amino acid moiety, epoxy-Val[4]. This hypothesis was further verified through an in-frame deletion experiment targeting *aglB*. The *ΔaglB* knockout strain completely abolished production of the cyclic peptide **1** and its cyclic analogues, while leading to accumulation of its truncated dipeptide derivative **2** (Fig. 5C). Moreover, metabolite analysis of the complemented strain (C-Δ*aglB*) revealed the restoration of **1** production concurrent with the disappearance of **2**, indicating that *aglB* is involved in the biosynthesis of the core pentapeptide scaffold of **1**.

Apart from the NRPSs responsible for core-scaffold biosynthesis, the enzymes for the unusual building blocks are also identified in *agl.* AglK and AglL, exhibiting 62% and 51% amino acid sequence identities, respectively, to their homologues KtzI and KtzT[72,73], are proposed to catalyze Piz biosynthesis. The precursor 3-Cl-AA is presumed to be derived from tryptophan through sequential catalysis by AglE (tryptophan halogenase), AglC (tryptophan 2,3-dioxygenase, TDO), AglD (kynurenine formamidase, KF), and AglF (kynureninase), following a route partly analogous to kynurenine pathway of tryptophan metabolism and that for antidepressant prodrug L-4-chlorokynurenine (L-4-Cl-Kyn) in taromycins[74] (Fig. 5B). Intriguingly, bioinformatics analysis of Streptomyces sp. 11-23 genome sequence disclosed a second triad of homologues of AglC/D/F, which are likely to be involved in theL-Trp metabolism[75] (Supporting Information Fig. S38). AglC/D/F appear to have diverged from their canonical counterparts after gene duplication and coevolved with the tryptophan 7-halogenase, AglE, to specifically process 7-Cl-Trp.

To our best knowledge, AglC represents the first reported TDO preferring 7-Cl-Trp, which complements the recently identified TDO acting on 6-Cl-Trp[74] and expands the biosynthetic toolbox for halogenated tryptophan-derived molecules[76].

Given the extreme rarity of epoxy valine in natural products, reported only recently in a patent for lipopeptide streptocinnamide B (Scm B)[77], we conducted a comprehensive analysis alongside targeted gene knockout experiments to interrogate its biosynthesis pathway. Epoxidation of alkyl carbons, a process rarely observed in nature[78-80], is hypothesized to occur either through a stepwise dehydrogenation−epoxidation cascade[78] or *via* a hydroxylation−epoxidation route[80]. In the wild-type strain, we detected **3**, featuring dehydro-Val[4], while no analogues containing hydroxy-Val[4] were observed, suggesting that the epoxy group might be formed *via* a dehydrogenation−epoxidation cascade. In the aglomycin pathway, only two oxygenases remain without assigned functions, AlgA (a FAD-dependent oxidase) and AlgG (a cytochrome P450 monooxygenase), which are proposed to jointly mediate the two required oxygenation steps. To test this hypothesis, we individually inactivated *aglA* and *aglG*, through in-frame deletion in *Streptomyces* sp. 11-23 wild-type strain. Unexpectedly, the production of **1** was almost unaffected in Δ*aglA* mutant. In contrast, the Δ*aglG* mutant completely abolished the production of **1** and **3**, while leading to prominent production of **2** and a new halogenated compound **5** (named aglomycin E) with *m/z* 781.2876 $[M+Na]^+$. Subsequent MS^2 fragmentation analysis identified **5** as an analogue of **1**, featuring Val[4] instead of epoxy-Val[4] (Fig. 5D and Supporting Information Fig. S34). Reintroduction of *aglG* into its mutant strain (C-Δ*aglG*) restored the production of **1** and **3** while dramatically reducing the levels of **2** and **5** (Fig. 5C). These results indicate that AglG is involved in forming the epoxide moiety of epoxy-valine, likely through a dehydrogenation step. However, the detailed biosynthetic mechanism remains elusive and warrants further investigation.

*3.5. The synergistic antibacterial effects of aglomycin A (**1**) and linezolid*

We evaluated the antibacterial activity of purified aglomycins A (**1**) and B (**2**) and found them displaying potent antibacterial effects against Gram-positive bacteria. The minimum inhibitory concentrations (MICs) against *E. faecium*, *E. faecalis*, *S. aureus* and methicillin-resistance *Staphylococcus aureus* (MRSA) ranged from 4 to 8 μg/mL (Fig. 6A). Subsequently, we further investigated their antibacterial activity against nine multidrug-resistant *Enterococcus* strains isolated from clinical samples. Both aglomycins A and B exhibited potent activity against clinical isolates of *E. faecium* and *E. faecalis* (Fig. 6B). Notably, aglomycins A and B were effective against VRE regardless of the *van*A or *van*B genotype, suggesting their potential as a novel therapeutic option for combating drug-resistant enterococcal infections.

Next, we investigated the potential synergistic or additive effects of **1** in combination with clinically used antibiotics against *Enterococci*, including linezolid, quinupristin, dalfopristin, and daptomycin (Fig. 6C and Supporting Information Fig. S39), with **2** excluded for significant cytotoxicity (Supporting Information Fig. S40A). Results disclosed that only the combination of **1** and linezolid demonstrated a fractional inhibitory concentration index below 0.5, specifically 0.375, indicating a synergistic effect against VRE. Furthermore, time-kill kinetics assay demonstrated that **1** exerted time-dependent bacteriostatic effects comparable to those of linezolid (Fig. 6D). When combined with linezolid, **1** significantly enhanced bactericidal activity at 24 h, reducing bacterial counts by approximately 1−2 $\log_{10}$ CFU/mL (Fig. 6E and Supporting Information Fig. S40B). These findings confirm that compound **1** and linezolid act synergistically against *E. faecium*. Moreover, their combination converted two individually bacteriostatic agents into a bactericidal outcome, suggesting a strong mechanistic interaction that alters the overall pharmacological effect and warrants further investigation.

Then, we employed a *Galleria mellonella* infection model[81] to evaluate **1** *in vivo* synergistic antibacterial activity in combination with linezolid. As shown in Fig. 6F, co-administration of **1** (2.5 mg/kg) and linezolid (0.5 mg/kg) in *E. faecium*-infected larvae significantly improved the survival rate compared to the vehicle control (DMSO in saline) ($P = 0.0006$) and to monotherapy with either aglomycin A ($P = 0.0044$) or linezolid ($P = 0.0005$), both of which provided only limited protection. This synergistic effect was further confirmed by co-administration of both drugs at 1 mg/kg each (Supporting Information Fig. S40C). Of note, while only a moderate synergistic effect (FIC = 0.375) was observed in the checkerboard assay, aglomycin A exhibited a pronounced synergy with linezolid *in vivo*. We propose that this discrepancy may partially stem from a transformation of the pharmacological mode from bacteriostatic to bactericidal upon combination. Collectively, these results reinforce the therapeutic potential of aglomycin A in combination with linezolid for combating infections with drug-resistant *Enterococci*.

4. Discussion

Halogenated compounds constitute an important class of drug-like molecules[2], with microorganisms representing a major natural source[10,82]. In biological systems, halogenation is generally catalyzed by diverse halogenases, which are widely distributed in nature[3,4]. However, a significant discrepancy exists between the prevalence of halogenases and the number of identified natural halogenated compounds, suggesting that many such molecules remain undiscovered[6,7]. Traditional approaches to their exploration are hampered by high rediscovery rates and the frequent inactivity of biosynthetic gene clusters (BGCs)[83], highlighting the need for efficient, high-throughput methods for systematic analysis.

To address these challenges, we developed DeepHalo, a deep learning-powered workflow optimized for high-throughput mining of halogenated molecules in large-scale LC−HRMS data. The workflow features five key advancements that collectively enhance both the confidence and efficiency of halogenated molecule identification. First, we developed a deep learning-based Element Prediction Model (EPM) that achieves state-of-the-art performance in detecting halogenated compounds leveraging only seven isotope peak features. Surprisingly, ablation experiments revealed that the monoisotopic mass, previously considered critical[12], had no impact on model performance (Fig. S3b). This phenomenon can be explained by the linear correlation between the P1/P0 intensity ratio and monoisotopic mass (Supporting Information Fig. S41), suggesting the model inherently learns mass information through isotopic patterns. Second, we implemented an optional isotope pattern validation step that significantly improves prediction accuracy by mitigating errors in isotope peak extraction. Although this type of validation has been used for detecting

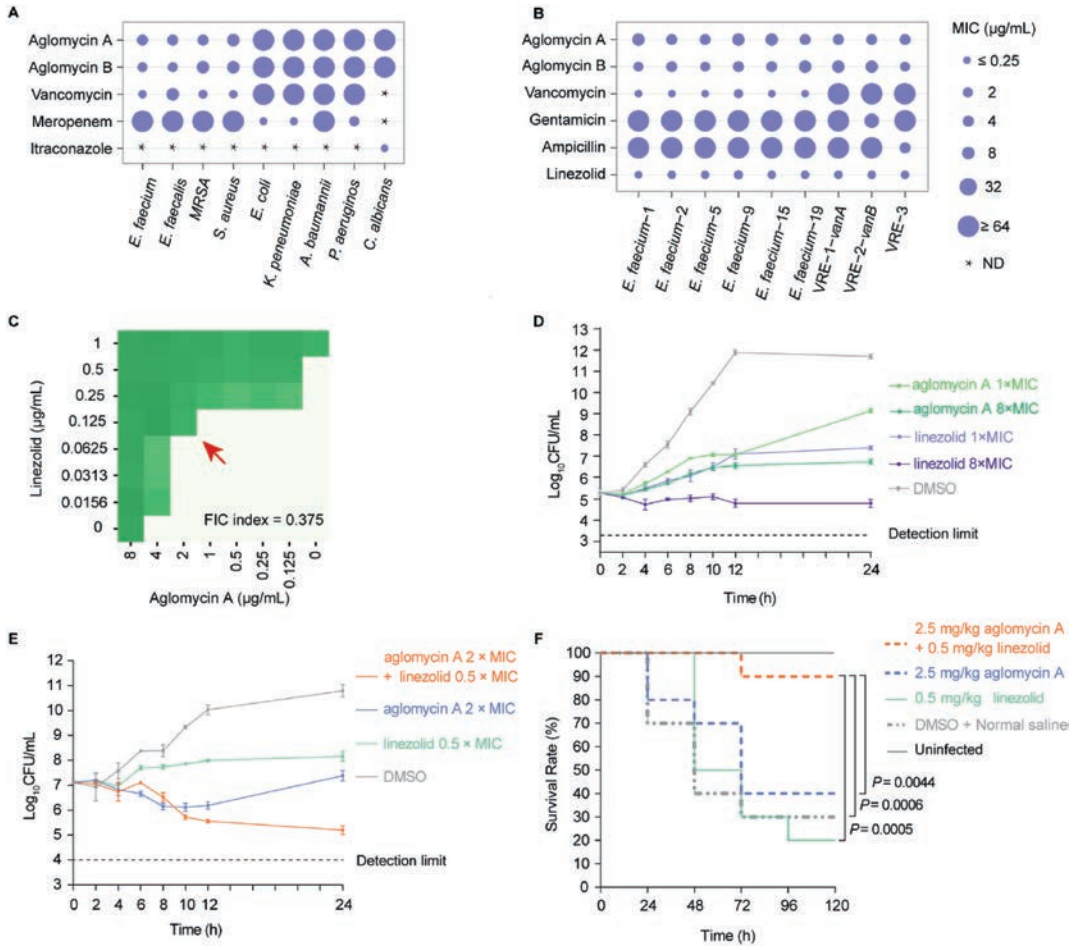

Figure 6 Aglomycin A (**1**) exhibits synergistic antibacterial effects in combination with linezolid. (A) MIC values against various microbes. ND: not detected. (B) MIC values against various drug-resistant *Enterococci* clinical isolates. (C) Checkerboard assay of aglomycin A with linezolid against *E. faecium*; dark green indicates higher inhibition. (D) Time-kill curve of aglomycin A and linezolid alone on *E. faecium*. Data represent the mean of three technical replicates with similar results. (E) Time-kill curve of aglomycin A in combination with linezolid on *E. faecium*. Data represent mean of three technical replicates. (F) The synergistic antibacterial effects of aglomycin A and linezolid in a *G. mellonella* infection model. Details see Supporting Information Tables SD8-SD13. Cox regression analysis was performed using R software (version 4.2.2), along with MSTATA software (https://www.mstata.com/).

compounds with common elements, DeepHalo is the first to extend this approach to the detection of halogenated compounds. Third, DeepHalo introduces a hierarchical halogen confidence score (H-score) that integrates centroid- and scan-level isotopic pattern analyses to reliably assess halogenation likelihood. This innovative metric enables adjustable threshold filtering and visualization-guided prioritization of halogenated features in complex metabolomic datasets. Fourth, by integrating MS^1 dereplication (optimized for halogenated compounds) with GNPS-compatible MS^2 molecular networking, our workflow achieves both high accuracy and efficiency in annotating halogenated molecules. Fifth, through computational optimizations, DeepHalo achieves unmatched efficiency, processing large-scale LC−HRMS datasets orders of magnitude faster than conventional tools. Collectively, these advancements establish DeepHalo as a transformative workflow for halogenate metabolomics, enabling high-throughput discovery and identification of halogenated compounds with unprecedented speed and accuracy.

Following the development of DeepHalo, we employed this workflow to systematically investigate halogenated metabolites in 1296 LC−HRMS/MS datasets obtained from 54 actinomycete strains. The analysis revealed an unexpectedly high prevalence of halometabolite production, with 16 strains (30%) yielding detectable halogenated compounds. DeepHalo successfully identified five structurally distinct classes of known halogenated metabolites in these strains, all verified through HRMS, MS^2, UV spectroscopy, and bioinformatics analyses. Moreover, genome mining targeting halogenases corroborated the high detection rate of halogenated metabolites and revealed a vast, untapped reservoir within the genus *Streptomyces*. These findings further demonstrate DeepHalo as an effective tool for halogenated compound discovery.

Additionally, from prioritized candidates, we characterized a novel cyclic depsipeptide, aglomycin A, along with its three derivatives from *Streptomyces* sp. 11-23. Aglomycin A features two rare structural blocks: 3-chloroanthranilic acid, first identified in natural products despite its common use in chemical synthesis, and an epoxide valine moiety, an exceptionally rare residue documented only recently in a patent[77]. Given the characteristic structure of aglomycin A, we subsequently investigated its

biosynthesis using bioinformatics analyses and targeted gene knockout experiments. Bioinformatics analysis clearly disclosed that 3-chloroanthranilic acid is derived from tryptophan *via* the catalysis of the enzyme quartet AglE/C/D/F, with AglC representing the first reported TDO preferring 7-Cl-Trp. In addition, gene knockout experiments indicated that the epoxide moiety is formed through a dehydrogenation−epoxidation cascade mediated by AglG, a cytochrome P450 enzyme rarely reported to catalyze such a reaction[84]. Notably, using AglG as a probe, we identified a series of diverse biosynthetic gene clusters characterized by a conserved arrangement of genes encoding an A-T didomain NRPS (most predicted to selectively activate valine), a cytochrome P450, and a TEII (Supporting Information Fig. S42). These findings suggest that the occurrence of epoxidized valine in natural products may be underestimated, while the AglG offers a practical genetic probe for their systematic discovery.

Finally, we evaluated the antibacterial activity of aglomycins and found that aglomycin A exhibits potent efficacy against clinically relevant multidrug-resistant *Enterococcus* strains. Notably, aglomycin A acts synergistically with linezolid in both *in vitro* and *in vivo* models. We acknowledge that the *Galleria mellonella* model used here has limitations for profiling time-dependent antibiotics like aglomycin A and linezolid. Nevertheless, given the significant clinical challenges posed by linezolid's adverse effects[85] and narrow therapeutic window[86], our findings provide a strong rationale for future studies in mammalian models to thoroughly evaluate the therapeutic potential of this combination.

Despite DeepHalo's successful large-scale screening performance, several limitations warrant discussion. First, polyhalogenated compound identification occasionally fails due to inaccuracies in extracting their complex isotope patterns (Supporting Information Table SD5). Second, although implementing synthetic false isotope peaks and rigorous verification procedures, residual errors persist. These errors predominantly localize in the 100% methanol elution fractions, where severe co-elution of analytes coupled with detector saturation effects substantially distort the isotopic peak profiles and compromise their accurate extraction[87] (Supporting Information Fig. S43). Third, while our EPM model demonstrates high accuracy in classifying Se-, B-, and Fe-containing compounds, their detection in raw metabolomic data remains challenging: Se/Fe detection suffers from unreliable isotopic peak extraction due to the exceptionally low abundance of their first two isotopic peaks (Fig. S2), while B analysis shows elevated false positives from hydrogen loss interference[88]. Collectively, these challenges all fundamentally stem from limitations in isotopic pattern extraction, suggesting that future development should prioritize more robust algorithms for isotope peak detection.

5. Conclusions

Overall, we present DeepHalo, an advanced HRMS-based metabolomics workflow that integrates deep learning with hierarchical optimization to deliver state-of-the-art performance in detecting halogenated compounds from complex matrices. As a proof of concept, DeepHalo enabled high-throughput screening of microbial halogenated natural products and led to the discovery of a novel antibiotic depsipeptide with potential against VRE infections. Beyond natural product research, DeepHalo's robust detection capability opens a wide range of applications, including, but not limited to, precise tracking of halogenated drug metabolism in pharmacological studies, comprehensive monitoring of persistent halogenated pollutants in environmental science, and sensitive detection of chlorinated disinfection byproducts in water quality analysis. Although the current implementation is optimized for halogenated molecules, DeepHalo's underlying framework is designed for broader elemental-feature analysis. Its dual-branch Isotope Patterns Neural Network (IsoNN) offers an efficient architecture, enabling the rapid development of customized models for diverse elemental signatures. This versatile design establishes DeepHalo as a powerful platform for targeted discovery and precise annotation of uncommon element-bearing molecules, with the potential to significantly advance element-driven metabolomics research.

Acknowledgments

This research was co-funded by CAMS Innovation Fund for Medical Sciences (CIFMS, 2021-I2M-1-028, China), the National Natural Science Foundation of China (81973219 and 82104047, China) and Beijing-Tianjin-Hebei Natural Science Cooperation Project (25JJJC0023, China), and supported by Biomedical High Performance Computing Platform, Chinese Academy of Medical Sciences. We sincerely thank Professor Zhengyan Guo, Professor Haiyan He, and Professor Xinyi Yang for their thorough review of our manuscript and constructive suggestions.

Author contributions

Yunying Xie, Bin Hong, and Xingxing Li conceptualized and managed this study. Xin Qi established DeepHalo workflow. Shanshan Chang, Mengyuan Wang, Xinyue Huang, Ning He, Mingxu Chen, Qing Lv, Jiahan Wang, Yu Du, Xinran Chen, and Quanxiu Gao carried out the wet experiments. Xin Qi, Shanshan Chang, Mengyuan Wang, Xinyue Huang, Yu Du, Shuchen Wang, and Yihong Li analyzed the data. Shanshan Chang, Xin Qi, Mengyuan Wang, Xinyue Huang, and Yunying Xie drafted the manuscript. Shuchen Wang, Xingxing Li, Bin Hong and Yunying Xie edited the manuscript.

Conflicts of interest

The authors declare no competing financial interests.

Appendix A. Supporting information

Supporting information to this article can be found online at https://doi.org/10.1016/j.apsb.2026.02.010.

References

1. Faleye OS, Boya BR, Lee JH, Choi I, Lee J. Halogenated antimicrobial agents to combat drug-resistant pathogens. *Pharmacol Rev* 2024;**76**:90−141.
2. Xu Z, Yang Z, Liu Y, Lu Y, Chen K, Zhu W. Halogen bond: its role beyond drug−target binding affinity for drug discovery and development. *J Chem Inf Model* 2014;**54**:69−78.
3. Crowe C, Molyneux S, Sharma SV, Zhang Y, Gkotsi DS, Connaris H, et al. Halogenases: a palette of emerging opportunities for synthetic biology−synthetic chemistry and C−H functionalisation. *Chem Soc Rev* 2021;**50**:9443−81.

4. Menon BRK, Richmond D, Menon N. Halogenases for biosynthetic pathway engineering: toward new routes to naturals and non-naturals. *Catal Rev* 2022;**64**:533−91.
5. Chiang CY, Ohashi M, Le J, Chen PP, Zhou Q, Qu S, et al. Copper-dependent halogenase catalyses unactivated C−H bond functionalization. *Nature* 2025;**638**:126−32.
6. Ferrinho S, Connaris H, Mouncey NJ, Goss RJM. Compendium of metabolomic and genomic datasets for cyanobacteria: mined the gap. *Water Res* 2024;**256**:121492.
7. Hornung A, Bertazzo M, Dziarnowski A, Schneider K, Welzel K, Wohlert SE, et al. A genomic screening approach to the structure-guided identification of drug candidates from natural sources. *Chembiochem* 2007;**8**:757−66.
8. Gao P, Huang Y. Detection, distribution, and organohalogen compound discovery implications of the reduced flavin adenine dinucleotide-dependent halogenase gene in major filamentous actinomycete taxonomic groups. *Appl Environ Microbiol* 2009;**75**: 4813−20.
9. Zhang L, Liu B, Zhu T, Tian X, Chen N, Wang Y. Advances in the study of halogenated natural products. *Curr Top Med Chem* 2025;**25**: 1217−50.
10. van Santen JA, Poynton EF, Iskakova D, McMann E, Alsup TA, Clark TN, et al. The natural products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res* 2022;**50**: D1317−23.
11. Roullier C, Guitton Y, Valery M, Amand S, Prado S, Robiou du Pont T, et al. Automated detection of natural halogenated compounds from LC−MS profiles-application to the isolation of bioactive chlorinated compounds from marine-derived Fungi. *Anal Chem* 2016;**88**: 9143−50.
12. Andersen AJ, Hansen PJ, Jorgensen K, Nielsen KF. Dynamic cluster analysis: an unbiased method for identifying A + 2 element containing compounds in liquid chromatographic high-resolution time-of-flight mass spectrometric data. *Anal Chem* 2016;**88**:12461−9.
13. Léon A, Cariou R, Hutinet S, Hurel J, Guitton Y, Tixier C, et al. HaloSeeker 1.0: a user-friendly software to highlight halogenated chemicals in nontargeted high-resolution mass spectrometry data sets. *Anal Chem* 2019;**91**:3500−7.
14. Zhao T, Wawryk NJP, Xing S, Low B, Li G, Yu H, et al. ChloroDBPFinder: machine learning-guided recognition of chlorinated disinfection byproducts from nontargeted LC−HRMS analysis. *Anal Chem* 2024;**96**:2590−8.
15. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS online: a web-based platform to process untargeted metabolomic data. *Anal Chem* 2012;**84**:5035−9.
16. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat Methods* 2019;**16**: 299−302.
17. Chang S, Qi X, Wang M, Huang X, He N, Chen M, et al. DeepHalo: deep learning-powered exploration of halogenated metabolites uncovering antibacterial depsipeptides. bioRxiv 2025.2025.08.11.669588 Available from: https://doi.org/10.1101/2025.08.11.669588.
18. Pfeuffer J, Bielow C, Wein S, Jeong K, Netz E, Walter A, et al. OpenMS 3 enables reproducible analysis of large-scale mass spectrometry data. *Nat Methods* 2024;**21**:365−7.
19. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat Biotechnol* 2016;**34**:828−37.
20. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. COCONUT online: collection of open natural products database. *J Cheminf* 2021;**13**:2.
21. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, et al. ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res* 2016;**44**:D1214−9.
22. Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, et al. HMDB 5.0: the human metabolome database for 2022. *Nucleic Acids Res* 2022;**50**:622−31.
23. Toxicology EsNCfC. *Chemical metadata from the CompTox chemicals dashboard linked to predicted spectra*. 2019. Available from: https://epa.figshare.com/articles/dataset/Chemical_Metadata_from_the_CompTox_Chemicals_Dashboard_Linked_to_Predicted_Spectra/7660976.
24. Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, et al. ZINC20-A free Ultralarge-scale chemical database for ligand discovery. *J Chem Inf Model* 2020;**60**:6065−73.
25. Meusel M, Hufsky F, Panter F, Krug D, Müller R, Böcker S. Predicting the presence of uncommon elements in unknown biomolecules from isotope patterns. *Anal Chem* 2016;**88**:7556−66.
26. Schymanski EL, Ruttkies C, Krauss M, Brouard C, Kind T, Dührkop K, et al. Critical assessment of small molecule identification 2016: automated methods. *J Cheminf* 2017;**9**:22.
27. Casmi Nikolić D. 2016: a manual approach for dereplication of natural products using tandem mass spectrometry. *Phytochem Lett* 2017;**21**: 292−6.
28. Kösters M, Leufken J, Leidel SA. SMITER-A python library for the simulation of LC−MS/MS experiments. *Genes* 2021;**12**:396.
29. Aalizadeh R, Alygizakis NA, Schymanski EL, Krauss M, Schulze T, Ibáñez M, et al. Development and application of liquid chromatographic retention time indices in HRMS-based suspect and nontarget screening. *Anal Chem* 2021;**93**:11601−11.
30. *Revisiting CASMI*. Fiehn Laboratory; 2022. Available from: https://fiehnlab.ucdavis.edu/casmi.
31. Zhang W, Zhao PX. Quality evaluation of extracted ion chromatograms and chromatographic peaks in liquid chromatography/mass spectrometry-based metabolomics data. *BMC Bioinf* 2014;**15**:S5.
32. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003;**13**:2498−504.
33. Chang S, Luo Y, Wang M, He N, Chen M, Huang X, et al. Pairing comparative genomics with tandem mass-based molecular networking allows to highly efficient discovery of nonribosomal peptides from Nocardia spp. *J Chromatogr, A* 2023;**1708**:464343.
34. Xie Y. FunGBE: functional group biosynthesis explore. Available from: http://www.npba-xielab.com:8502/.
35. Hammer PE, Burd W, Hill DS, Ligon JM, van Pée K. Conservation of the pyrrolnitrin biosynthetic gene cluster among six pyrrolnitrin-producing strains. *FEMS Microbiol Lett* 1999;**180**:39−44.
36. Podzelinska K, Latimer R, Bhattacharya A, Vining LC, Zechel DL, Jia Z. Chloramphenicol biosynthesis: the structure of CmlS, a flavin-dependent halogenase showing a covalent flavin-aspartate bond. *J Mol Biol* 2010;**397**:316−31.
37. Mitchell AJ, Zhu Q, Maggiolo AO, Ananth NR, Hillwig ML, Liu X, et al. Structural basis for halogenation by iron- and 2-oxo-glutarate-dependent enzyme WelO5. *Nat Chem Biol* 2016;**12**:636−40.
38. Kissman EN, Neugebauer ME, Sumida KH, Swenson CV, Sambold NA, Marchand JA, et al. Biocatalytic control of site-selectivity and chain length-selectivity in radical amino acid halogenases. *Proc Natl Acad Sci U S A* 2023;**120**:e2214512120.
39. Oberg N, Zallot R, Gerlt JA. EFI-EST, EFI-GNT, and EFI-CGFP: enzyme function initiative (EFI) web rresource for genomic enzymology tools. *J Mol Biol* 2023;**435**:168018.
40. Doerschuk AP, Johnson JR, Goodman JJ, Szumski SA, Growich JA, Miller PA, et al. Biosynthesis of tetracyclines. I. The halide metabolism of *Streptomyces aureofaciens* mutants. The preparation and characterization of tetracycline, 7-chloro36-tetracycline and 7-bromotetracycline. *J Am Chem Soc* 1959;**81**:3069−75.
41. Smith CG. Effect of halogens on the chloramphenicol fermentation. *J Bacteriol* 1958;**75**:577−83.
42. Glasser NR, Cui D, Risser DD, Okafor CD, Balskus EP. Accelerating the discovery of alkyl halide-derived natural products using halide depletion. *Nat Chem* 2024;**16**:173−82.

43. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, et al. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* 2021;**49**:29−35.
44. Cockerill F, Wikler M, Alder J, Dudley M, Eliopoulos G, Ferraro M, et al. *Methods for dilution antimicrobial susceptibility tests for bacteria that grow aerobically: approved standard*. Wayne, PA: Clinical and Laboratory Standards Institute; 2012. Available from: https://clsi.org/shop/standards/m07/.
45. Hall MJ, Middleton RF, Westmacott D. The fractional inhibitory concentration (FIC) index as a measure of synergy. *J Antimicrob Chemother* 1983;**11**:427−33.
46. Odds FC. Synergy, antagonism, and what the chequerboard puts between them. *J Antimicrob Chemother* 2003;**52**:1.
47. Mackay ML, Milne K, Gould IM. Comparison of methods for assessing synergic antibiotic interactions. *Int J Antimicrob Agents* 2000;**15**:125−9.
48. Li S, Siddiqa A, Thapa M, Chi Y, Zheng S. Trackable and scalable LC−MS metabolomics data processing using asari. *Nat Commun* 2023;**14**:4113.
49. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006;**313**:504−7.
50. Lee JG, Yoo ID, Kim WG. Differential antiviral activity of benzastatin C and its dechlorinated derivative from Streptomyces nitrosporeus. *Biol Pharm Bull* 2007;**30**:795−7.
51. Prado-Alonso L, Pérez-Victoria I, Malmierca MG, Montero I, Rioja-Blanco E, Martín J, et al. Colibrimycins, novel halogenated hybrid polyketide synthase-nonribosomal peptide synthetase (PKS-NRPS) compounds produced by *Streptomyces* sp. sstrain CS147. *Appl Environ Microbiol* 2022;**88**:e0183921.
52. Lopez P, Hornung A, Welzel K, Unsin C, Wohlleben W, Weber T, et al. Isolation of the lysolipin gene cluster of *Streptomyces tendae* Tu 4042. *Gene* 2010;**461**:5−14.
53. Morshed MT, Lacey E, Vuong D, Lacey AE, Lean SS, Moggach SA, et al. Chlorinated metabolites from *Streptomyces* sp. highlight the role of biosynthetic mosaics and superclusters in the evolution of chemical diversity. *Org Biomol Chem* 2021;**19**:6147−59.
54. Dardic D, Lauro G, Bifulco G, Laboudie P, Sakhaii P, Bauer A, et al. Svetamycins A-G, unusual piperazic acid-containing peptides from *Streptomyces* sp. *J Org Chem* 2017;**82**:6032−43.
55. Ray L, Yamanaka K, Moore BS. A peptidyl-transesterifying Type I thioesterase in salinamide biosynthesis. *Angew Chem Int Ed Engl* 2016;**55**:364−7.
56. Tsutsumi H, Katsuyama Y, Izumikawa M, Takagi M, Fujie M, Satoh N, et al. Unprecedented cyclization catalyzed by a cytochrome P450 in benzastatin biosynthesis. *J Am Chem Soc* 2018;**140**:6631−9.
57. Eddy SR. A new generation of homology search tools based on probabilistic inference. *Genome inform* 2009;**23**:205−11.
58. Siemion Tw IZ, Pook KH. Influence of the distance of the proline carbonyl from the beta and gamma carbon on the ^{13}C chemical shifts. *Angew Chem Int Ed Engl* 1975;**14**:702−3.
59. Tang WZ, Liu JT, Hu Q, He RJ, Guan XQ, Ge GB, et al. Pancreatic lipase inhibitory cyclohexapeptides from the marine sponge-derived fungus *Aspergillus* sp. 151304. *J Nat Prod* 2020;**83**:2287−93.
60. Ikai KSK, Takesako K, Kato I, Naganawa H. NMR studies of aureobasidins A and E. *J Antibiot (Tokyo)* 1991;**44**:1199−207.
61. Li S, Li Y, Yu T, Wei X, Zhang Y, Chen B, et al. Discovery of potent antiosteoporotic cyclic depsipeptides with an unusual nitrile hydroxy acid from *Microascus croci*. *Bioorg Chem* 2025;**155**:108133.
62. Vijayasarathy S, Prasad P, Fremlin LJ, Ratnayake R, Salim AA, Khalil Z, et al. C3 and 2D C3 Marfey's methods for amino acid analysis in natural products. *J Nat Prod* 2016;**79**:421−7.
63. Grimblat N, Sarotti AM. Computational chemistry to the rescue: modern toolboxes for the assignment of complex molecules by GIAO NMR calculations. *Chemistry* 2016;**22**:12246−61.
64. Grimblat N, Zanardi MM, Sarotti AM. Beyond DP4: an improved probability for the stereochemical assignment of isomeric compounds using quantum chemical calculations of NMR shifts. *J Org Chem* 2015;**80**:12526−34.
65. Selin C, Habibian R, Poritsanos N, Athukorala SN, Fernando D, de Kievit TR. Phenazines are not essential for *Pseudomonas chlororaphis* PA23 biocontrol of *Sclerotinia sclerotiorum*, but do play a role in biofilm formation. *FEMS Microbiol Ecol* 2010;**71**:73−83.
66. Gulick AM. Conformational dynamics in the Acyl-CoA synthetases, adenylation domains of non-ribosomal peptide synthetases, and firefly luciferase. *ACS Chem Biol* 2009;**4**:811−27.
67. Horswill AR, Escalante-Semerena JC. Characterization of the propionyl-CoA synthetase (PrpE) enzyme of *Salmonella enterica*: residue Lys592 is required for propionyl-AMP synthesis. *Biochemistry* 2002;**41**:2379−87.
68. Rausch C, Hoof I, Weber T, Wohlleben W, Huson DH. Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol Biol* 2007;**7**:78.
69. Kim MS, Bae M, Jung YE, Kim JM, Hwang S, Song MC, et al. Unprecedented noncanonical features of the nonlinear nonribosomal peptide synthetase assembly line for WS9326A biosynthesis. *Angew Chem Int Ed Engl* 2021;**60**:19766−73.
70. Wang S, Brittain WDG, Zhang Q, Lu Z, Tong MH, Wu K, et al. Aminoacyl chain translocation catalysed by a type II thioesterase domain in an unusual non-ribosomal peptide synthetase. *Nat Commun* 2022;**13**:62.
71. Planckaert S, Deflandre B, de Vries AM, Ameye M, Martins JC, Audenaert K, et al. Identification of novel rotihibin analogues in streptomyces scabies, including discovery of its biosynthetic gene cluster. *Microbiol Spectr* 2021;**9**:e0057121.
72. Fujimori DG, Hrvatin S, Neumann CS, Strieker M, Marahiel MA, Walsh CT. Cloning and characterization of the biosynthetic gene cluster for kutznerides. *Proc Natl Acad Sci U S A* 2007;**104**:16498−503.
73. Du YL, He HY, Higgins MA, Ryan KS. A heme-dependent enzyme forms the nitrogen-nitrogen bond in piperazate. *Nat Chem Biol* 2017;**13**:836−8.
74. Luhavaya H, Sigrist R, Chekan JR, McKinnie SMK, Moore BS. Biosynthesis of l-4-chlorokynurenine, an antidepressant prodrug and a non-proteinogenic amino acid found in lipopeptide antibiotics. *Angew Chem Int Ed Engl* 2019;**58**:8394−9.
75. Zummo FP, Marineo S, Pace A, Civiletti F, Giardina A, Puglia AM. Tryptophan catabolism *via* kynurenine production in *Streptomyces coelicolor*: identification of three genes coding for the enzymes of tryptophan to anthranilate pathway. *Appl Microbiol Biotechnol* 2012;**94**:719−28.
76. Reed KB, Brooks SM, Wells J, Blake KJ, Zhao M, Placido K, et al. A modular and synthetic biosynthesis platform for *de novo* production of diverse halogenated tryptophan-derived molecules. *Nat Commun* 2024;**15**:3188.
77. Makareva TN, Rebachuk NM, Lubennaya LA, Mineev KS, Shubina LK, Guglya EB, et al. *Elyakov Pacific Institute of Bioorganic Chemistry, Far Eastern Branch, Russian Academy of Sciences (PIBOC FEB RAS), assignee. New depsipeptide compounds with antibacterial activity.* Russian Federation patent RU2795449C1; 2022 Oct 27.
78. Gaisser S, Lill R, Staunton J, Méndez C, Salas J, Leadlay PF. Parallel pathways for oxidation of 14-membered polyketide macrolactones in Saccharopolyspora erythraea. *Mol Microbiol* 2002;**44**:771−81.
79. Shah S, Xue Q, Tang L, Carney JR, Betlach M, McDaniel R. Cloning, characterization and heterologous expression of a polyketide synthase and P-450 oxidase involved in the biosynthesis of the antibiotic oleandomycin. *J Antibiot (Tokyo)* 2000;**53**:502−8.
80. Hashimoto T, Matsuda J, Yamada Y. Two-step epoxidation of hyoscyamine to scopolamine is catalyzed by bifunctional hyoscyamine 6β-hydroxylase. *FEBS (Fed Eur Biochem Soc) Lett* 1993;**329**:35−9.
81. Menard G, Rouillon A, Cattoir V, Donnio PY. *Galleria mellonella* as a suitable model of bacterial infection: past, present and future. *Front Cell Infect Microbiol* 2021;**11**:782733.
82. Gribble GW. A survey of recently discovered naturally occurring organohalogen compounds. *J Nat Prod* 2024;**87**:1285−305.

83. Walsh CT, Tang Y. *Natural product biosynthesis*. The royal society of chemistry; 2022.
84. Parisi G, Freda I, Exertier C, Cecchetti C, Gugole E, Cerutti G, et al. Dissecting the cytochrome P450 OleP substrate specificity: evidence for a preferential substrate. *Biomolecules* 2020;**10**:1411.
85. Conradie F, Bagdasaryan TR, Borisov S, Howell P, Mikiashvili L, Ngubane N, et al. Bedaquiline-pretomanid-linezolid regimens for drug-resistant tuberculosis. *N Engl J Med* 2022;**387**:810−23.
86. Cairns KA, Udy AA, Peel TN, Abbott IJ, Dooley MJ, Peleg AY. Therapeutics for vancomycin-resistant enterococcal bloodstream infections. *Clin Microbiol Rev* 2023;**36**:e0005922.
87. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. *Anal Bioanal Chem* 2007;**389**:1017−31.
88. Treutler H, Neumann S. Prediction, Detection, and Validation of Isotope Clusters in Mass Spectrometry Data. *Metabolites* 2016;**6**:37.