Machine learning inference of natural product chemistry across biosynthetic gene cluster types

Martin Larralde[1,2], Georg Zeller[1,2,3]

[1] Molecular Systems Biology Unit, EMBL, 69117 Heidelberg, Germany
[2] Leiden University Center for Infectious Diseases (LUCID), Leiden University Medical Center, 2333 ZA Leiden, Netherlands
[3] Center for Microbiome Analyses and Therapeutics, Leiden University Medical Center, 2333 ZA Leiden, Netherlands.

Correspondence should be addressed to m.larralde@proton.me or georg.zeller@gmail.com

Abstract

With ever-increasing volumes of sequencing data for biosynthetic gene clusters (BGCs), computational methods for the prediction of resulting secondary metabolites are critically needed. Here, we present CHAMOIS, a machine learning tool inferring metabolite properties from protein domains in BGCs. Out of 539 relevant chemical properties from the ChemOnt ontology, CHAMOIS predicts 120 with an AUPRC > 0.5. Although entirely data-driven, CHAMOIS infers many protein-metabolite links that are consistent with the scientific literature and suggests interesting novel biosynthetic functions of uncharacterized proteins. Finally, to guide experimental BGC characterisation, CHAMOIS can pinpoint which BGC within a given genome produces a pre-specified metabolite.

Keywords

Biosynthetic gene clusters; natural products; biosynthesis; machine learning; chemical ontology

Background

Microorganisms across Earth's habitats are capable of producing an astonishing diversity of natural products. The enzymatic machinery for synthesising many of these compounds is encoded by genes located in genomic proximity and thus referred to as biosynthetic gene clusters (BGCs). BGC-encoded enzymes often interact in biosynthetic pathways in a modular fashion. Evolution acting on these genetic building blocks has given rise to a vast diversity of cluster architectures and a resulting biochemical diversity of produced molecules(1). While BGCs usually contain a core biosynthetic cluster composed of key enzymes(2), these are often surrounded by additional genes involved in regulation or metabolite transport(3).

Historically, BGCs were often studied in microbial isolates known to synthesize a secondary metabolite of interest. BGC identification usually relied on genetic techniques such as cosmid library generation(4) or knock-out studies targeting putative biosynthetic or resistance genes(5). More recently, the deluge of genomic data warranted the development of novel methods for identifying candidate BGCs *in silico*. Several such methods are based on rules derived from expert knowledge, including the popular antiSMASH software(6), which combines the detection of key biosynthetic genes with a system of biochemistry-aware rules. In parallel, standardised repositories of experimentally-validated BGCs, such as MIBiG(7), encouraged the development of a new generation of machine learning (ML) and artificial intelligence (AI) methods to complement rule-based methods. These include ClusterFinder(8), DeepBGC(9), GECCO(10) and SanntiS(11).

With an ever-increasing amount of genomic data, the sprawl of BGC prediction methods has led to an exponential accumulation of BGC predictions: the first large-scale repository of BGC prediction, the antiSMASH-db (2017), contained ~22,000 BGC predictions made by antiSMASH on ~4,000 bacterial genomes(12); six years later, the proGenomes3 database (2023) released more than 3 million BGC predictions made with GECCO(10,13). The most comprehensive resource for experimentally-validated BGCs, MIBiG 4.0 (2024), however, numbers only 2,437 "active" entries (https://mibig.secondarymetabolites.org/stats) . Hence, for the vast majority of *in silico* predicted microbial BGCs, the encoded metabolite is unknown, since natural products remain more challenging to identify than the genomic locations of BGCs. Only BGCs with strong homology to experimentally characterized ones can be confidently hypothesised to produce a similar metabolite. However, even the deletion of small accessory genes can substantially alter the final product, its biochemical properties and biological activity(14). BGCs with remote homology to known instances are very challenging to characterise *in silico*(15). Even today, natural products discovery from genomic data is limited by time-consuming experiments, and the large volumes of available BGC predictions remain relatively unexplored, underlining the need for better computational metabolite inference and prioritisation methods.

Toward this goal, antiSMASH categorises BGCs according to one or more high-level product types based on its ruleset (101 distinct types in v8.0). These types summarise the broad biosynthetic pathway, but often reveal little detail about the final molecule itself. Currently available data-driven approaches for BGC categorisation assign BGCs to one or several of six more coarse-grained MIBiG types, typically using relatively simple ML classifiers, as e.g.

implemented in DeepBGC(9). While the predicted types can be useful for prioritising BGCs for experimentation, their information content is generally too low to derive interesting chemical properties or bioactivities of the putative product.

Among the different types of BGCs, relationship between enzyme sequence and corresponding metabolite properties have been mostly studied for the polyketide (PK) and non-ribosomal peptide (NRP) types, as they often feature multifunctional enzymes composed of multiple biosynthetic domains. These domains form an assembly line and incorporate smaller precursors into the molecule, which is modified sequentially to form a larger metabolite backbone(16). Several studies have shown the specificity of certain PK or NRP synthase domains for particular precursors based on sequence features(17,18). These discoveries have in turn been integrated into computational methods for inferring some aspects of natural product biosynthesis, such as NRPSpredictor2(19), NERPA(20), RiPPMiner(21), transAtor(22), or PRISM(23), the three latter of which can be used to predict the backbone structure of BGC products from sequence features, but only for certain BGC types. While these methods represent a step towards natural product prediction for certain BGCs, consistent and accurate product prediction across all BGC types remains a fundamental challenge, despite the elucidation of more and more biosynthetic pathways.

Here, to approach this problem from a novel data-driven perspective, we developed CHAMOIS, the first method to employ machine-learning for predicting chemical properties of BGCs products from gene sequences across all major types of BGCs. We evaluate its prediction accuracy in carefully stratified cross-validation to assess how CHAMOIS generalizes across distinct BGCs and metabolites. We further demonstrate that CHAMOIS's data-driven model captures relevant known and novel biochemical information about enzymatic domains, and finally show that CHAMOIS is useful to pinpoint which BGC within a given genome gives rise to an a priori known metabolite – a capability that could expedite experimental characterisation of BGCs and discovery of novel biosynthetic enzymes.

Results

CHAMOIS: A novel machine-learning method for prediction of chemical properties from a BGC sequence

Inferring chemical structures directly is nearly impossible because of the sheer size of chemical space. The limited availability of experimentally validated BGCs for training and testing poses an additional challenge which renders the task of directly inferring metabolite structures from BGC sequences extremely difficult. Therefore, instead of generating candidate metabolite structures, we developed a method for predicting chemical properties of these metabolites using machine learning to infer links between gene sequences and chemical properties from experimentally characterized BGCs with known metabolites.

To capture chemical properties of arbitrary natural products for ML inference, we use the classes of a chemical ontology, ChemOnt, which contains 4,825 non-exclusive chemical categories covering all domains of chemistry(24). This representation of molecules into a lower-dimensional space of relevant properties resembles molecular fingerprinting, a common task in cheminformatics(25). Most fingerprinting methods, however, are developed for pairwise molecule comparisons, e.g. for database retrieval or ligand screening, and are not so well suited for ML prediction and interpretability. Compared to molecular fingerprints, chemical ontologies, such as ChemOnt, actually offer interpretable hierarchy classes, and can be more discriminative of closely related molecules, such as natural products of the same family.

To predict the ChemOnt classes for a putative BGC product using only genomic features, we developed a novel ML method, called CHAMOIS (Chemical Hierarchy Approximation for secondary Metabolism clusters Obtained In Silico), which is trained on ChemOnt annotations of known BGC-derived compounds from MIBiG 3.1 (N=1,598 bacterial BGC corresponding to a total of N=1,034 ChemOnt classes, see Fig. 1 and *Sequence dataset preparation* section of the Methods). Instead of manual annotation, we used ClassyFire(24), a tool to automatically assign any molecule with a known chemical structure to the ChemOnt ontology. The hierarchical ChemOnt class annotation was subsequently treated as a binary vector (ones indicating class memberships) and for simplicity and scalability approached as a series of (independent) binary classification tasks. For reasons of interpretability, we opted for LASSO logistic regression as binary classification models(26). We excluded classes occurring in less than 5 compound groups, to enable evaluation with 5-fold stratified grouped cross-validation (*vide infra*). The remaining classes (N=539) were used to train LASSO classifiers (see *Compound dataset preparation* section of the Methods).

To prepare genomic sequence-derived features for training these classifiers, we first identified open reading frames (ORFs) for each BGC region(27,28), translated these and annotated them with Pfam 38.0 domains(29,30). Pfam is a curated database of protein domains which offers profile Hidden Markov Models (pHMMs) for automated domain annotation. This approach has been repeatedly shown to yield informative features for subsequent ML modeling(9,10,31). Here, we restricted Pfam feature space to domains that appeared in at least one BGC in MIBiG 3.1 resulting in binary feature vectors (of 3,334 dimensions) encoding presence/absence of each domain per BGC. On average, 26.5 unique

Pfam domains were identified in the MIBiG 3.1 BGCs (Supplementary Fig. 1). Using Pfam domains facilitated interpretability, as LASSO models learnt to associate specific Pfam domains with compound classes (see section *CHAMOIS infers the biosynthetic capabilities of protein domains* of the Results). To train the classifiers, we excluded features appearing in less than 5 compound groups. The binary feature vectors with the remaining domains (N=927) were passed to every independent logistic regression classifier.

Evaluating prediction of chemical properties

Since we decomposed the hierarchical topology of the ChemOnt class labels into binary classification tasks, we could separately evaluate each classifier using cross-validation. As MIBiG contains homologous clusters for the same compound across different species (e.g. coformycin(32), BGC0002039-40), as well as groups of closely related BGCs producing compounds of the same family (e.g kanamycin and tobramycin(33)), a naively sampled cross-validation would lead to biased accuracy estimates that are inflated by highly similar BGC-metabolite instances. To assess if CHAMOIS would be able to generalize beyond very similar instances, we first grouped BGCs based on the similarity of their produced compound, using Hamming distance(34) between MHFP6 fingerprints(25) to measure the distance of produced metabolites in chemical space (see *Compound dataset preparation* section of the Methods). We then used these groups (where all BGCs/compounds with pairwise chemical similarity of 0.5 and greater would be assigned to the same fold, total N=1,247 examples used for cross-validation) to perform stratified cross-validation preserving these groups, using 5 distinct folds for each of the 539 classifiers (see *Cross-validation* section of the Methods).

The median area under the receiver-operating characteristic (ROC) curve (AUROC) across all evaluated classes was 0.79 (Supplementary Table 1). However, because of the imbalance in the predicted classes (Supplementary Fig. 2a), AUROC values can be misleadingly high despite the underlying classifier having very low precision (i.e. having excess false-positives among the instances it predicts to belong to a given class). As an evaluation metric better reflecting classifier precision for these unbalanced tasks, we focused on the area under the precision-recall curve (AUPRC).

In our cross-validation, 120 ChemOnt classes were found to be predicted by CHAMOIS with an AUPRC of 0.5 or higher (Fig. 2a, see also Supplementary Table 1 for the metrics for each class). This constitutes an assessment of generalization across chemically dissimilar metabolites due to the grouping of similar metabolites in cross-validation folds (Fig. 2b). We also evaluated the performance of Random Forest classifiers in place of LASSO logistic regressions and found no significant differences in overall performance (Supplementary Fig. 3).

Among the classes with the highest AUPRC were some high-level classes represented by many training instances, such as Organonitrogens (CHEMONTID:0000278, AUPRC=0.979, N_{pos}=1,284) or Organoheterocyclic compounds (CHEMONTID:0000002, AUPRC=0.937 N_{pos}=1,325). However, the class with the highest AUPRC was Phenazines (CHEMONTID:0000416, AUPRC=1.0, N_{pos}=13), a rare class of tricyclic compounds known to be synthesised by a pathway of 7 conserved genes, *phzA-G*(35). In addition, CHAMOIS exhibited good performance on other relatively rare classes (Fig. 2c-d, Supplementary Fig.

2b): Cyclohexylamines (CHEMONTID:0002674, AUPRC=0.861, N_{pos}=24), Thiazoles (CHEMONTID:0000095, AUPRC=0.769, N_{pos}=81), Organohalogens (CHEMONTID:0000267, AUPRC=0.800, N_{pos}=150), or Hydroxamic acids (CHEMONTID:0000376, AUPRC=0.832, N_{pos}=71). Furthermore, the classes best predicted by CHAMOIS correspond to classes found across various types of BGCs, demonstrating that CHAMOIS predictions do not simply correlate with broad BGC types (Fig. 2d).

CHAMOIS infers the biosynthetic capabilities of protein domains

CHAMOIS's ML model is entirely data driven and unaware of any knowledge of natural product biosynthesis. Because we used LASSO classifiers, the learnt coefficients are sparse, which reduces the risk of overfitting and facilitates interpretation. To summarise the importance of few decisive domains in ChemOnt class prediction as learned by CHAMOIS, we extracted the 2 domains with the highest weight for every class, and used these relations to build a bipartite network capturing the strongest associations between ChemOnt classes and Pfam domains (Fig. 3). While each ChemOnt class could be predicted by a combination of many domains (on average 63, with a median of 34, domains received a non-zero LASSO coefficient), CHAMOIS classifiers for many ChemOnt classes indeed assigned a weight greater than 2.0 to a single domain (N=201, ~37%) or two domains (N=106, ~20%) (Supplementary Fig. 4, see Supplementary Table 2 for the complete matrix of weights learned by CHAMOIS).

In CHAMOIS's classification model, a domain receiving a high weight suggests that it contributes to the formation of metabolite features that are characteristic of a given class. Without prior knowledge, CHAMOIS was able to recover known biosynthetic capabilities of certain Pfam domains. Among these examples of known biosynthetic domains were the Tryptophan halogenase domain (PF04820) linked to Aryl halides (CHEMONTID:0002866) biosynthesis, or the P-aminobenzoate N-oxygenase domain (PF11583) linked to Nitrobenzenes (CHEMONTID:0000036) biosynthesis (for 20 examples see Table 1 and Supplementary Table 3). This proof of principle that CHAMOIS extracted meaningful information about biosynthetic domains, prompted us to explore domain/metabolite class associations more broadly.

CHAMOIS gives insight into the function of unknown domains

To identify the potential involvement of uncharacterized Pfam domains in the biosynthesis of certain compound classes, we collected a subset of Pfam domains devoid of functional annotation (as per InterPro 107.0(36)). In our dataset derived from MIBiG 3.1 an overwhelming majority of the BGC-associated domains still lack functional annotations: 1,428 of 1,576 BGCs (~91%) contain at least one unannotated domain, also as a consequence of recent Pfam releases introducing new domains, such as DUF6069 (PF19545) present in 14 PKS BGCs, but still lack detailed annotations of their biosynthetic functions. From the set of domains with unknown function, we retained those to which CHAMOIS assigned a high weight (>2.0) in any ChemOnt classifier. This screening resulted in a list of 106 uncharacterized domains for which CHAMOIS inferred biosynthetic functions, for a total of 292 domain/class pairs (Supplementary Table 4, see *Identification of uncharacterized domains* section of the Methods) representing a rich resource for the discovery of biosynthetic functions.

While by construction, biosynthetic functional annotations could not be automatically retrieved for any of these domains, targeted literature searches recovered several of them to be well-described as such. For instance, the YcaO cyclodehydratase domain (PF02624) has been extensively studied for its involvement in post-translational modification of ribosomal peptides(37), yet it did not have any functional annotation in the corresponding InterPro entry (IPR003776). The associated ChemOnt classes in CHAMOIS were consistent with the literature, with a high weight assigned by the Thiazoles (CHEMONTID:0000095, w=4.005) classifier (Table 1, Supplementary Table 4). Other domains lacking annotations in the databases, but for which our approach recovered a function consistent with the scientific literature, included the Thiazolinyl imine reductase Irp3-like domain (PF21390), the Nitroreductase domain (PF00881), and the Lantibiotic biosynthesis dehydratase C-terminal domain (PF14028) (Table 1).

Beyond these confirmatory results, CHAMOIS also made novel predictions for which no direct literature support could be found, e.g. by connecting domains with truly unknown functions to a whole branch of the ChemOnt hierarchy. For instance, the Conserved hypothetical protein 95 (PF03602) received a high weight from the Dialkylarylamines (CHEMONTID:0003901, w=2.858, P=1.8E-3) and Tertiary alkylarylamines (CHEMONTID:0002454, w=2.784, P=1.3E-3) classifiers, hinting at a putative *N,N*-dimethyltransferase function. Among domains of unknown function *stricto sensu*, DUF742 (PF05331) received a high weight from the 1,2-diols (CHEMONTID:0002467, w=2.132, P=0.04) classifier, DUF4135 (PF13575) from Hydroxy acids and derivatives (CHEMONTID:0000472, w=2.392, P=6.9E-3) classifiers, DUF5837 (PF19155) from the Thiazolecarboxylic acids and derivatives (CHEMONTID:0002007, w=3.515, P=1.91E-10) classifier, and DUF6531 (PF20148) from the Hydroxypyridines (CHEMONTID:0004151, w=2.203, P=1.1E-3) classifier. Altogether these results suggest that CHAMOIS is capable of learning relevant associations between chemical classes and protein domains highlighting the promise of exploring these associations for the functional elucidation of uncharacterized biosynthetic domains.

CHAMOIS outperforms PRISM 4.0 for structure prediction of Polyketides and NRPs

PRISM(23) is a state-of-the-art method for prediction of BGCs and their putative metabolites using a mixture of rules and *in silico* combinatorial generation of chemical structures. To compare CHAMOIS and PRISM 4 performance, we used the set of 1,279 BGCs known as the “gold standard BGCs”(38). We applied CHAMOIS to predict ChemOnt classes for these BGCs, and then searched the Natural Product Atlas(39) for the entry most similar to each CHAMOIS prediction. As done in Skinnider *et al.*, we computed the Morgan fingerprint of each prediction, and then took the median of the Tanimoto similarity to the true compound for each cluster (Supplementary Table 5, see the *Comparison to PRISM 4* section of the Methods). In this evaluation, CHAMOIS significantly outperformed PRISM 4.0 on structure prediction for BGCs encoding Polyketide or NRPs, while PRISM 4.0 slightly outperformed CHAMOIS on RiPP BGCs (Supplementary Fig. 5).

CHAMOIS can prioritize putative BGCs for experimental characterization

A common task encountered by natural product biochemists is the identification of the producing cluster in the genome of a producer strain from which a metabolite of interest has already been identified. To pinpoint the biosynthetic machinery producing this metabolite, researchers typically first analyze the genome sequence with BGC prediction tools such as antiSMASH or GECCO, and subsequently manually inspect all predicted BGCs to identify the producing cluster. This search can be supported by CHAMOIS by predicting a ChemOnt fingerprint for each BGC and then quantifying its similarity to the fingerprint of the compound of interest, identifying the BGC with the highest likelihood of producing a certain metabolite.

To benchmark CHAMOIS's ability to screen BGC predictions for a known compound, we assembled a test dataset of 70 experimentally-annotated bacterial BGCs which had been identified in a complete bacterial genome (Fig. 4a, Supplementary Table 6) and most of which showed moderate to no homology to any BGC from MIBiG 3.1, on which CHAMOIS was trained. The known metabolites were analysed with ClassyFire, as similarly done in the training dataset preparation to obtain a ChemOnt class vector. The source genomes were analyzed with antiSMASH and GECCO, to generate sets of candidate BGCs. In case of overlapping BGC predictions (made with different BGC finding tools), consensus clusters were obtained by taking the intersection gene set across overlapping predictions; when neither tool predicted a region overlapping the experimentally-validated BGC, it was manually added to the predictions. This resulted in 2,331 unique BGC predictions across all genomes (on average 37, ranging from 5 to 76 per genome). We then applied CHAMOIS to each BGC to predict a fingerprint consisting of the 530 ChemOnt classes covered by the model. For each genome, we calculated the similarity between all corresponding BGC fingerprints and that of the true compound using a probabilistic Jaccard index(40). Although most compounds in this benchmarking dataset belong to the Polyketide and NRP classes, the dataset represents an otherwise diverse set in the space of ChemOnt labels (Fig. 4b, see *Contextualization of benchmark data* section of the Methods).

In this benchmark, CHAMOIS identified the correct cluster in 31 of the 70 cases (~44%). In an additional 19 instances (~27%), it ranked the true BGC among the top 5 clusters. That CHAMOIS in nearly three quarters of the benchmarking examples ranked the correct BGC very highly appears remarkable given that many of the genomes in this benchmark contained a large number of predicted BGCs of the same biosynthetic type and that fingerprint predictions by CHAMOIS often show only moderate similarity to that of the produced compound (<0.6, except for one BGC, Fig. 4c). Moreover, most of the examples used for benchmarking here share low chemical or sequence similarity with any secondary metabolites or BGC sequences contained in MIBiG 3.1 (used to train CHAMOIS), which underlines CHAMOIS's capacity to generalize to diverse cluster architectures and metabolite classes. For instance, CHAMOIS successfully recovered the BGC for 8,8a-deoxyoleandolide across 64 predicted loci despite the cognate BGC only sharing ~2% identity with the closest MIBiG BGC (BGC0000898, desosamine). Overall these results suggest that CHAMOIS could be a useful tool to expedite the identification of the BGC corresponding to interesting metabolites in isolated producer strains or low-complexity metagenomes.

Discussion

Predicting secondary metabolites from their cognate BGCs *in silico* is an important problem, the urgency and potential of which is highlighted by the growing divide between available sequence and metabolite data. While previous research has resulted in prediction tools for certain BGC types, CHAMOIS is the first-of-its-kind universal open-source machine-learning method for predicting the ontological classification of a BGC product from its genomic sequence. In our study we carefully assessed its predictive capacity across dissimilar compound subsets in cross-validation to avoid overoptimistic performance reports (Fig. 2). In addition we performed a benchmark on an additional external dataset, which mostly contained BGCs and metabolites without similarity to the training data (Fig. 4). A limitation of our method, which implicitly relies on protein domains involved in biosynthesis, is the limited prediction performance for RiPP BGC products due to several RiPP-specific challenges. They often produce large molecules, the complexity of which can sometimes not be fully captured by ClassyFire. As an additional challenge, the structure of RiPP molecules largely depends on the precursor peptide sequence, not only on tailoring enzymes; while the latter can be captured by Pfam annotations, the precursor peptide typically remains unannotated. Even so, as CHAMOIS allows the prediction of tailoring enzyme modifications, it may potentially be useful in conjunction with type-specific methods that predict a BGC compound backbone, such as RiPPMiner(21).

CHAMOIS is an entirely data-driven method using multilabel LASSO logistic regression to infer ChemOnt classes. While LASSO regularisation avoids overfitting and enforces sparse weights useful for feature selection, it however comes with a drawback in the case of co-occurring features, where it will often arbitrarily select one and disregard the others(41). From a biological point of view, co-occurring features may often be found in BGCs encoding enzyme complexes, which require the presence of more than one protein to catalyse a given reaction. In these cases, LASSO-derived associations between domains and ChemOnt classes will on the one hand tend to capture only one complex member, so that the others remain undetected. On the other hand false-positive predictions may occur when an incomplete pathway is encountered. On the upside, CHAMOIS's LASSO classifiers are fast to retrain and evaluate, and will easily cope with – and benefit from – more data becoming available in future BGC database updates. Likewise, the inclusion of new domain features will be straightforward as CHAMOIS does not rely on their functional annotation. In particular, when more natural products belonging to rare ChemOnt classes become available in the future, CHAMOIS will be capable of predicting a more diverse set of classes and will provide an automated approach to obtain clues as to the protein domains characteristic of these classes and their key biosynthetic reactions (Fig. 3). Eventually, an even larger amount of training data could support modeling more complex relationships between genes, e.g. using modern AI architectures such as Autoencoders in the future. However, as currently available data sets are limited in size, such methods are unlikely to outperform CHAMOIS's more classical ML approach.

As CHAMOIS enables (inferential) translation of BGCs into a chemical space, it holds potential for exploring natural product space beyond what is feasible based on genomic representations alone. This translation can for instance be used to scan sets of BGC predictions for members of a chemical class of interest, or to cluster BGCs based on their putative product properties, or to assess biosynthetic diversity in terms of not only (meta-

)genomic sequence similarity, but also biochemical class representation. Additionally, the genetic basis of this chemical space (in terms of biosynthetic protein domains predicted by CHAMOIS) can be explored as an additional layer of information to guide the discovery of new enzymatic domains in well-characterized as well as newly discovered BGCs.

Conclusion

Chamois represents a scalable, entirely data-driven computational method for inferring secondary metabolite properties from their cognate BGCs that is – unlike most previous methods – not restricted to certain types of clusters. It facilitates exploring the biochemical space of natural products on the basis of rapidly growing genomic BGC resources as an important step towards genotype-phenotype prediction for this highly relevant area of microbial metabolism.

Methods

Sequence dataset preparation. BGC sequences for MIBiG 3.1(42) entries were downloaded from the MIBiG website (https://mibig.secondarymetabolites.org) as GenBank files. 567 BGCs were filtered based on a manually curated list, to exclude eukaryotic records, and records with known annotation issues (Supplementary Table 7). The cluster sequences were trimmed to the coordinates of the outermost genes of the clusters. The coordinates of 82 clusters were programmatically corrected to retain only the biosynthetic core based on the reference literature (see `data/scripts/mibig/download_records.py` in the project repository). The BGC records were then annotated with the CHAMOIS CLI (`chamois annotate`). Cluster genes were called using Pyrodigal(28) v3.6.3 wrapping Prodigal(27) v2.6.3 in *meta* mode with closed ends (`-c -p meta`). The genes were then mapped against Pfam(43) v38.0 using `hmmsearch` as implemented in PyHMMER(30) v0.11.1 wrapping HMMER (http://hmmer.org) v3.4. The Pfam domains were filtered using the "trusted" bitscore cutoffs of each profile HMM (`--cut_tc`). From each BGC, a boolean feature vector of annotated Pfam domains (d=3,146) was constructed, which was stored in an HDF5 file using the `anndata` library(44) v0.12.3.

Compound dataset preparation. The BGC metadata for MIBiG 3.1(42) and 4.0(7) entries were downloaded from the MIBiG website (https://mibig.secondarymetabolites.org) as JSON files, containing the structures of the compounds as SMILES. 110 cluster compounds were programmatically corrected for incorrect name or SMILES annotations (see the download script `data/scripts/mibig/download_compounds.py` in the project repository). For compounds missing a structure in the metadata, the corresponding NPAtlas structure was fetched if a cross-reference existed. When compounds contained no cross-reference, they were mapped by name against the NPAtlas(45), or, failing that, against PubChem(46)

(https://pubchem.ncbi.nlm.nih.gov). All obtained SMILES strings were then converted into InChi using RDKit(47) v2023.9.6 (http://www.rdkit.org). The InChi strings of each compound were submitted to annotation on the ClassyFire(24) web server (http://classyfire.wishartlab.com). The classification of each compound was then encoded into a binary vector of ChemOnt classes for each compound and assigned to its cognate BGCs; for BGCs producing more than one compound, the compound with the highest number of ChemOnt classes was used. For MIBiG 3.1, 1,598 BGCs could be successfully labelled this way. The resulting labels were stored in an HDF5 file using the `anndata` library(44) v0.12.3. For further stratification, we grouped BGCs using a pairwise Hamming distance(34) cutoff of 0.5 between their MHFP6 fingerprints(25) computed using RDKit(47) v2023.9.6 (http://www.rdkit.org) resulting in 1,180 groups of BGCs that were distinct to each other with respect to the encoded metabolites.

Model training. The training dataset was first filtered to remove clusters shorter than 1,000 bp. Classes with less than 5 positive or negative occurrences across groups were excluded from the label matrix. A LASSO logistic regression model was trained on each target label independently (N=539) with the `LogisticRegression` classifier implemented in `scikit-learn`(48) v1.7.2, using the `LIBLINEAR`(49) solver with L1 regularisation and 100 maximum iterations. Features that received zero weights (coefficients) from all classifiers were discarded, and only the remaining features were retained in the model (d=896). The weights were then combined into a single weight matrix and stored alongside the model metadata in JSON format. The whole training procedure is implemented in the `train` command of the `chamois` command-line tool.

Cross-validation. The model was validated using a cross-validation strategy. First, classes with less than 5 positive or negative occurrences across groups were excluded from the label matrix. Then, for each class of the label matrix, a 5-fold cross-validation was run, using a `StratifiedGroupKFold` from `scikit-learn`(48) v1.7.2 to split the data according to the molecular-similarity groups built from MHFP6 fingerprints(25), found in the "group" column of the observation table of the dataset. The probabilities generated for each label were combined into a probability matrix, for which the area under the receiver-operator characteristic curve (AUROC) and area under the precision-recall curve (AUPRC) were computed with micro- and macro-averaging using the `roc_auc_score` and `average_precision_score` functions of `scikit-learn` respectively. The whole evaluation procedure is implemented in the `cvi` command of the `chamois` command-line tool.

Visualization of cross-validation performance across the ChemOnt hierarchy. The `pronto` Python library (https://github.com/althonos/pronto) v2.6.0 was used to load the ChemOnt ontology from OBO format, and converting the class hierarchy into an adjacency matrix, treating each class as a node and each subclassing relationship as an edge in the tree. This tree was then displayed with the Vega visualization framework (https://vega.github.io/) in a radial tree layout, using AUPRC values computed with the `average_precision_score` as described above to color the nodes. Individual precision-recall curves were generated with `matplotlib`(50) v3.10.6 (Fig. 2).

Network visualization. The weights of the trained model were extracted from the CHAMOIS LASSO classifiers (Supplementary Table 2). For each classifier corresponding to a ChemOnt class, the two domains with the highest weights were extracted to form a bipartite graph with chemical classes and protein domains as nodes. To categorize the Pfam domains, the predictions from the ECDomainMiner tool(51) were used to label domains by top-level Enzyme Commision (EC) number. To categorize ChemOnt classes, their top-level ancestor in the ChemOnt hierarchy was extracted. The graph was then displayed with the Vega visualization framework (https://vega.github.io/) using force-directed layout (Fig. 3).

Identification of uncharacterized domains. Domain annotations were downloaded from InterPro(36) 107.0 in XML format (https://www.ebi.ac.uk/interpro/download/). We collected all Pfam(43) 38.0 domains, and retained those as "uncharacterized" which did not have a corresponding InterPro entry with an Enzyme Commission (EC) number or a Gene Ontology(52) term. We further analyzed domains that were assigned a weight of at least 2.0 by any classifier within CHAMOIS (Supplementary Tables 3 and 4).

Significance of domain to chemical class associations. For each uncharacterized domain/chemical class pair extracted , a contingency table was built, counting occurrences inside the MIBiG 3.1 dataset. For each table, Fisher's exact p-value was calculated using the scipy.stats.fisher_exact function of SciPy(53) v1.16.3. The p-values are reported both uncorrected and corrected with Bonferroni correction (Supplementary Table 4).

Comparison to PRISM 4. The BGCs records of the 1,281 "Gold Standard" BGCs from PRISM were downloaded from Zenodo(38). The PRISM 4-predicted SMILES for each BGC were obtained from the Supplementary Material of the PRISM 4 paper(23). The ChemOnt classes for each BGCs were predicted with the `chamois predict` command using default parameters. The predicted ChemOnt classes were used to search the Natural Product Atlas(39) with the `chamois search` command using default parameters. For each cluster, the highest ranking compounds were selected (allowing ties) and their SMILES extracted as the predicted CHAMOIS structures. Using the evaluation code of Skinnider *et al.* (https://github.com/Adapsyn/prism-4-paper), we computed the Tanimoto coefficient of the ECFP6 fingerprints of the predicted molecules to the true compound for CHAMOIS as well as the methods present in the PRISM4 "Gold Standard" dataset: PRISM 1(54), PRISM 4, NP.searcher(55) and antiSMASH 4.0(56) (Supplementary Table 5). For BGCs where all methods successfully predicted a structure (n=373), the median Tanimoto coefficient was extracted, and summary boxplots were generated similarly to Fig. 2c and 2e of Skinnider *et al.* (Supplementary Fig. 5a-b). For all BGCs, we extracted the median Tanimoto of CHAMOIS and PRISM 4 predictions, and grouped them by MIBiG BGC types across PK, NRP, RiPP and Other BGCs, respectively, to allow direct comparison (Supplementary Fig. 5c). Significance was assessed using unpaired *t*-test computed with the `scipy.stats.ttest_ind` function of SciPy(53) v1.16.3.

BGC screening in producer genomes. To evaluate how well CHAMOIS could pinpoint the cognate BGC for a given compound, we assembled a dataset of 70 experimentally-validated BGCs not included in MIBiG 3.1(42) that were contained in 65 (near-)complete bacterial genomes (Supplementary Table 6). Each of these genomes was newly annotated with the antiSMASH(6) web server v8.0 and GECCO(10). For regions where both tools predicted a

BGC, their intersection was kept. When neither tool predicted a region overlapping the experimentally-validated BGC, it was manually added to the predictions. All predicted BGCs (N=2,527) were then annotated with CHAMOIS to predict a ChemOnt classification. For each validated BGC/compound pair, the distance between the true compound and all predictions was computed using Probabilistic Jaccard similarity(40) between their ChemOnt classes. The BGC predictions were then sorted and assigned a relative rank based on their similarity to the ChemOnt classes of the true compound.

Contextualization of benchmark data. The NPAtlas v2024_03 entries(57) were downloaded from the NPAtlas(45) server. Each compound was binarized into an indicator label vector as described earlier (see *Compound dataset preparation*), using the pre-computed ChemOnt annotations provided by the NPAtlas. The complete indicator matrix was used to train a Principal Component Analysis (PCA) using the `PCA` class from `scikit-learn`(48) v1.7.2 with default parameters. The compounds from the benchmark dataset were labeled with the ClassyFire(24) web server (http://classyfire.wishartlab.com) as described earlier (see *Compound dataset preparation*). The compounds were then projected into the aforementioned principal component space. The resulting principal components were plotted with `matplotlib`(50) v3.10.0 (Fig. 4b).

BGC ranking significance. To evaluate the significance of the ranking produced by CHAMOIS on the whole dataset, the relative rank of the true compound in the CHAMOIS predictions sorted by Probabilistic Jaccard similarity was computed using the `rankdata` function of SciPy(53) v1.15.2 (with parameters `method="dense"`). An empirical background distribution was computed using randomization (N=100,000) by selecting a "true" BGC at random in each genome, and averaging their relative rank to the true compound. A lower relative rank than the one obtained by CHAMOIS was not found in any of the random permutations.

Establishing dissimilarity between benchmark and training sets. We queried the BGC predictions intersecting the true cognate BGCs against MIBiG 3.1 BGCs using `skani`(58) v0.2.2 (with parameters `-c10 -m50`) wrapped in Pyskani(59) v0.1.3. For each hit, the query ANI was computed, where □□□□□ □□□ = □□□□□□□□□ * □□□□□ □□□□□□□□□ with □□□□□□□□□ and □□□□□ □□□□□□□□□ computed by `skani`. In addition, we measured the chemical similarity between the true compound and every MIBiG 3.1 BGC using Hamming distance between MHFP6 fingerprints. The highest query ANI and the highest chemical similarity for each BGC were plotted with `matplotlib`(50) v3.10.0 (Fig. 4c).

Declaratations

Ethics approval and consent to participate.

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

Code availability. CHAMOIS code is open source under the GNU General Public License 3.0 or later (GPL-3.0-or-later), publicly hosted in a git repository on GitHub at https://github.com/zellerlab/CHAMOIS. An archive of the version used for this study (v0.2.0) is available in the Zenodo repository (https://zenodo.org/records/17849623). CHAMOIS can be installed for Python 3.7 and later on UNIX operating systems (Linux, MacOS, Windows WSL, etc.) from the Python Package Index (PyPI; https://pypi.org/project/chamois-tool) and from the Bioconda(60) channel of the `conda` package manager (https://anaconda.org/bioconda/chamois). A self-contained Docker image with all dependencies can be obtained from the GitHub Container Registry (https://ghcr.io/zellerlab/chamois).

Data availability. The datasets generated during the current study are available in the Zenodo repository (https://zenodo.org/records/17849853). CHAMOIS weights are included in this published article in Supplementary Table 2.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the European Molecular Biology Laboratory (EMBL); the SFB 1371 of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) [395357507 to G.Z.], and a LUMC Fellowship [to G.Z.].

Author's contributions

Software development and computational analyses were performed by ML, with suggestions from GZ. ML and GZ conceived the study, designed the figures and co-wrote the manuscript.

Acknowledgements

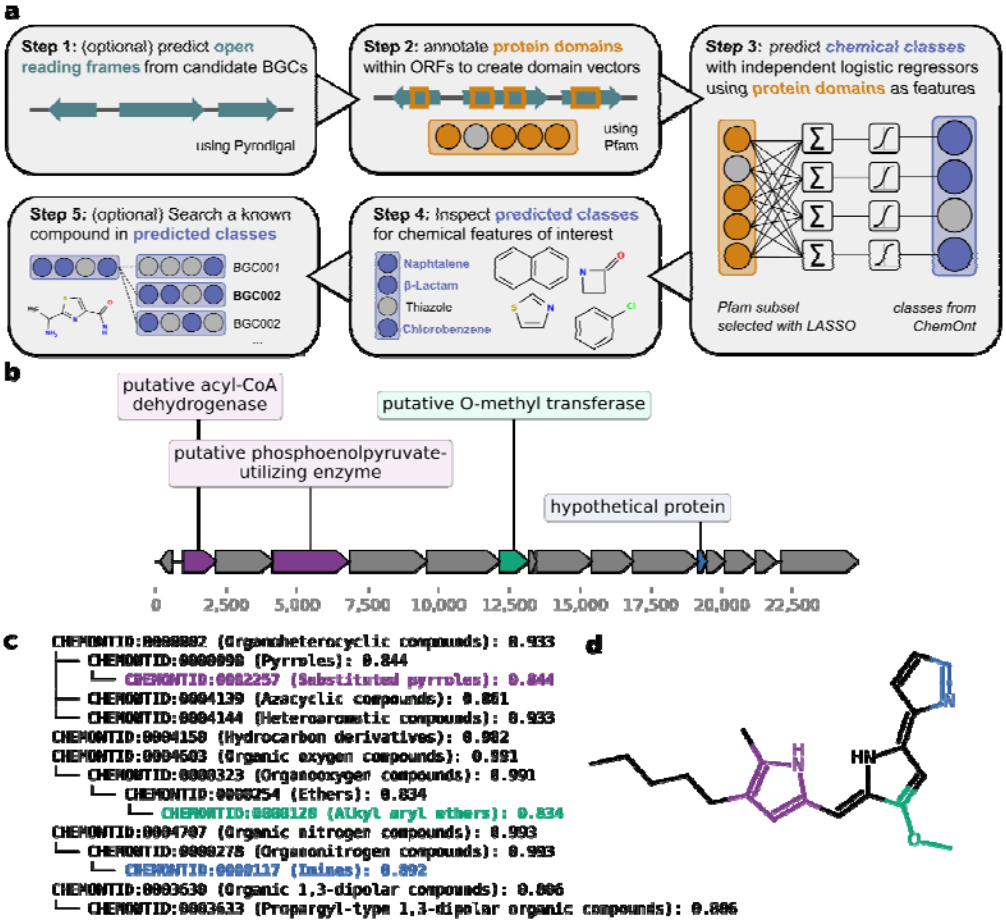
We are grateful to Joachim Hug and Michael Zimmerman (both EMBL Heidelberg), to Laura Carroll (Umeå University), to Justin van der Hooft and Marnix Medema (both Wageningen University) for their suggestions during development. We are indebted to the European Molecular Biology Laboratory (EMBL) and its IT Services Team for providing and administrating HPC resources.

Tables

Pfam domain	Accession	Class	Ref.	Fig. 3 inset
Pyridoxine 5'-phosphate oxidase, C-terminal	PF10590	Phenazines, Quinoxalines, Benzodiazines	(35)	
Tryptophan halogenase	PF04820	Organohalogens, Aryl halides	(61,62)	*i*
Nitric oxide synthase, oxygenase domain	PF02898	Nitroaromatic compounds, Organic nitro compounds	(63)	*xiv*
Flavoprotein	PF02441	Thioenol ethers	(64)	
Condensation domain	PF00668	Macrolide lactams, alpha amino-acid amides, peptides	(65)	
Cyclodipeptide synthase	PF16715	Pyrazines, lactams	(66)	
Formyl transferase	PF00551	N-formyl-alpha amino acids	(67)	
GTP cyclohydrolase II	PF00925	Pyrimidones	(68)	
Phosphoribosyl transferase	PF00156	Pentoses, nucleosides and analogues	(69)	*ix*
Methyltransferase Ppm1/Ppm2/Tcmp	PF04072	Methyl esters, carboxylic acid esters	(70)	
Lysine/ornithine monooxygenase	PF13434	Hydroxamic acids	(71)	*viii*
Squalene-hopene cyclase, N-terminal	PF13249	Steroids, hydroxysteroids, triterpenoids	(72)	
Ferritin-like VioB	PF12902	Indolocarbazoles, Pyrrolocarbazoles, Isoindolones	(73)	*ii*
Isochorismatase domain	PF00857	Catechols, Benzenediols	(74)	*iii*
Carbamoyltransferase, C-terminal	PF16861	Carbamate esters, carbamic acids	(75)	
Erythromycin biosynthesis protein CIII-like, C-terminal domain	PF06722	O-glycosyl compounds, carbohydrates	(76)	*vii*
Queuosine biosynthesis protein QueC	PF06508	Nitriles	(77)	
Class I-like SAM-dependent O-methyltransferase	PF01596	Anisoles, Methoxyphenols	(78)	*iv*
TfuA-like protein	PF07812	Thioamides, Imidothioic acids	(79)	*xi*
P-aminobenzoate N-oxygenase AurF	PF11583	Nitrobenzenes, Organic nitro compounds	(80)	*xiv*
Beta-lactamase class A, catalytic domain	PF13354	Beta lactams, Azetidines	(81)	*xii*
APS kinase domain	PF01583	Organic sulfuric acids, Alkyl sulfates	(82)	
DegT/DnrJ/EryC1/StrS aminotransferase	PF0104	1,2-aminoalcohols, Aminoglycosides	(83)	
SnoaL-like domain	PF13577	Tetracyclines, Anthracyclines, Ketones	(84)	
Thiazolinyl imine reductase Irp3-like, C-terminal	PF21390	Azolines, Azolidines, Thiazolines, Thiazolidines	(85)	
Lantibiotic biosynthesis dehydratase, C-terminal	PF14028	Thiazoles, Pyridines	(86)	
Cucumopine synthase, C-terminal	PF18631	Beta carbolines, Indoles, Pyridines	(87)	
YcaO cyclodehydratase, ATP-ad Mg2+-binding	PF02624	Oxazoles, Thiazoles, Azoles	(88)	
MvdD-like, pre-ATP grasp domain	PF21068	Lactones, Macrolides, Macrolactams	(89)	*vi*
Phosphoenolpyruvate phosphomutase	PF13714	Organic phosphonic acids, Organophosphorus compounds	(90)	*xiii*
Cupin 2, conserved barrel	PF07883	Epoxides	(91,92)	
SchA/CurD like domain	PF04486	Naphthopyranones, Dibenzopyrans	(93)	
PaaD, zinc beta ribbon domain	PF23451	Anilides, Acylaminobenzoic acids, 1-hydroxy-4-unsubstituted benzenoids	(94)	
Aminotransferase class IV	PF01063	Anilides, Acylaminobenzoic acids	(94)	
VCPO second helical-bundle domain	PF2278	Alkyl chlorides, alpha-chloroketones, sesquiterpenoids	(95)	

Table 1: Select subset of highly weighted protein domain / chemical class associations learnt by CHAMOIS. Literature references consistent with CHAMOIS-inferred domain-class

associations are included. Corresponding insets in Fig. 3 are indicated in the last column. An extended version of this table can be found in Supplementary Table 3.



Figures

Figure 1: CHAMOIS, a novel method for predicting chemical properties of secondary metabolites from biosynthetic gene clusters (BGCs). **(a)** Workflow depiction of the chemical ontology prediction as implemented in CHAMOIS. First, CHAMOIS identifies the open reading frames in the input BGCs. Then, protein domains are annotated in the resulting ORFs using a subset of profile Hidden Markov Models from Pfam (N=896) selected as informative for CHAMOIS modeling. Each BGC is then encoded as a boolean feature vector indicating presence/absence of these Pfam domains. From these feature vectors, a multilabel logistic regression classifier predicts classes in the ChemOnt hierarchy (N=539). Resulting predictions can be inspected for classes of interest, or compared for similarity against a query compound. **(b)** Example of a genomic locus and **(c)** CHAMOIS prediction for the prodigiosin BGC of *Serratia marcescens* (BGC0000259)(96). The predicted classes are depicted as hierarchy alongside their posterior probabilities. CHAMOIS accurately predicted the Substituted pyrroles (CHEMONTID:0002257), Alkyl aryl ethers (CHEMONTID:0000128) and Imines (CHEMONTID:0000117) classes, highlighted on **(d)** the prodigiosin molecule in

purple, green and blue respectively, with the PiKAChU(97) library. Genes containing domains that contributed to these class predictions are highlighted in the corresponding colour. CHAMOIS can produce an equivalent summary in tabular format with the `chamois explain cluster` command.

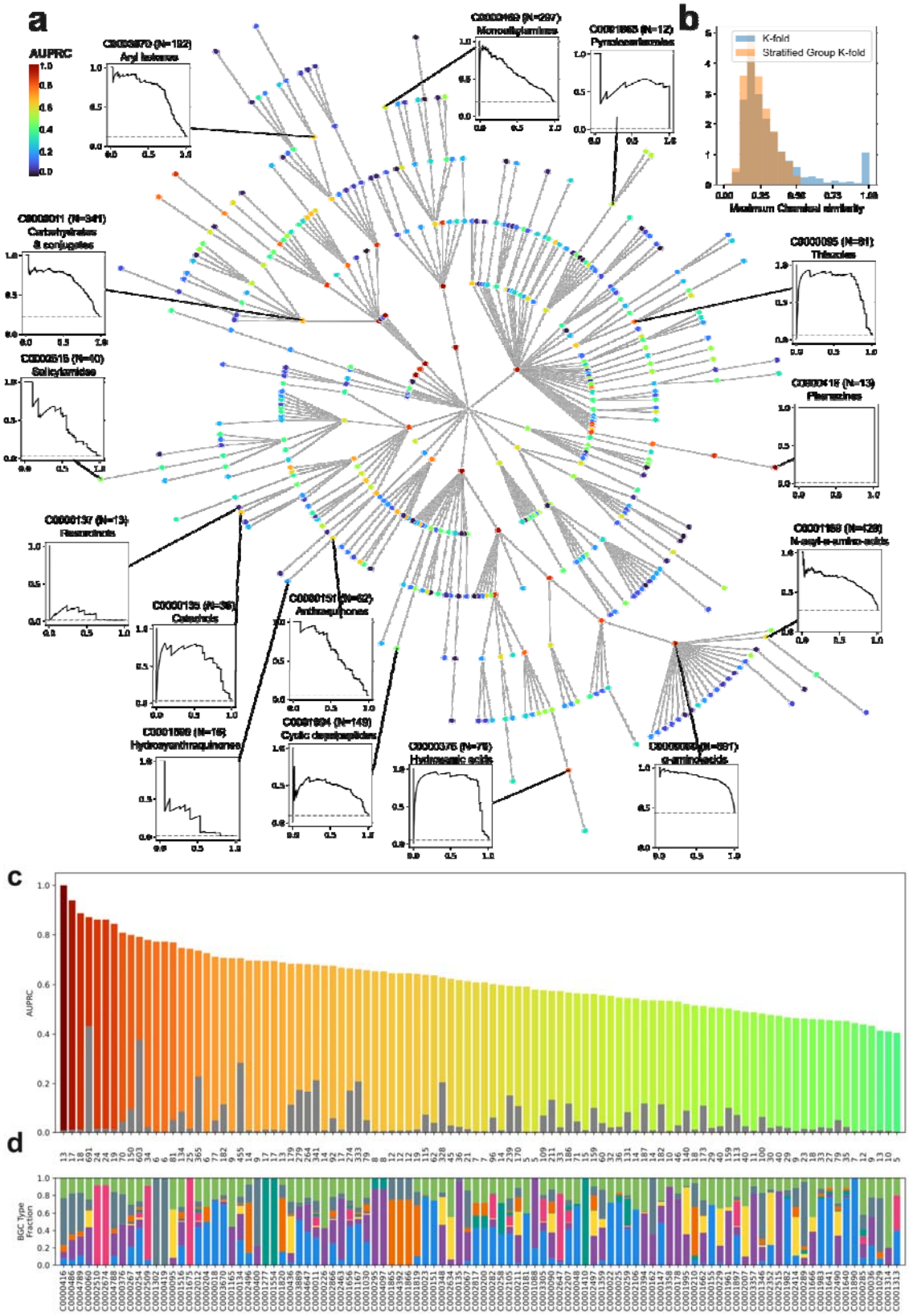

a
AUPRC
b
K-fold
Stratified Group K-fold
Maximum Chemical similarity
c
AUPRC
d
BGC Type Fraction

Figure 2: Evaluation of CHAMOIS against N=1,598 BGCs with known metabolites from MIBiG 3.1. **(a)** Model performance on every predicted ChemOnt class (N=539) using stratified grouped 5-fold cross-validation. Classes are displayed as nodes within the ChemOnt hierarchy and coloured according to CHAMOIS's performance for the respective class, as assessed by the area under the precision-recall curve (AUPRC, see colour key). Precision-recall curves for selected classes are shown as insets against the baseline obtained by random guessing (dashed horizontal lines corresponding to class proportions). An interactive version of this panel can be found as part of CHAMOIS' documentation at https://chamois.readthedocs.io/en/v0.2.0/figures/cv.html. **(b)** The barplot on the top-right corner indicates the distribution of chemical similarities (Hamming distance of MHFP6 fingerprints) between cross-validation folds, using either a naive 5-fold cross validation (blue) or a stratified cross-validation in which groups of similar compounds are kept in the same fold to assess generalization across distinct compounds (orange, used here with a distance cutoff of 0.5). **(c)** AUPRC values and for the 100 ChemOnt classes with the highest difference between CHAMOIS classifier AUPRC and random-guessing baseline AUPRC, shown as the coloured and grey section of each bar, respectively, sorted and coloured according to CHAMOIS AUPRC. Overall, predictors with >0.5 AUPRC were obtained for a total of 120 ChemOnt classes (Supplementary Table 1). **(d)** Distribution of BGCs in each ChemOnt class grouped by MIBiG types: *blue* Polyketide, *purple* NRP, *yellow* RiPP, *orange* Alkaloid, *pink* Saccharide, *green* Mixed (mostly Polyketide/NRP), *dark gray* Other.

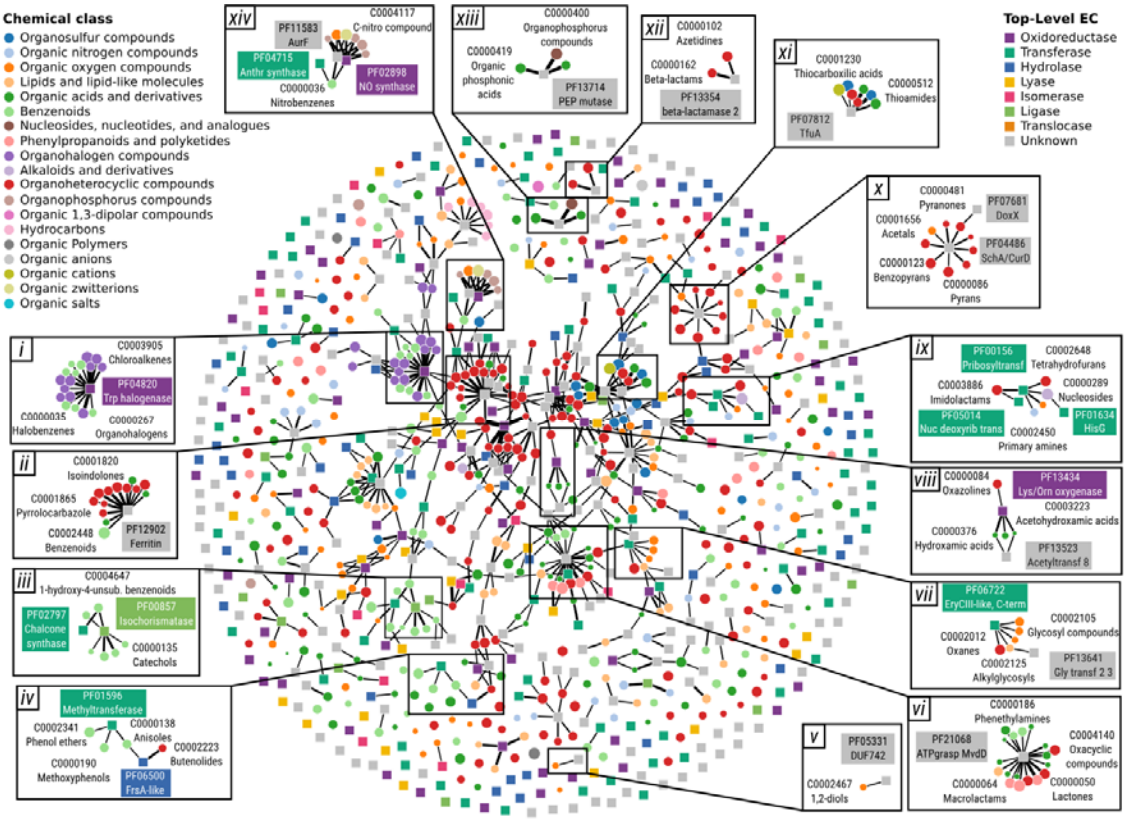


Figure 3: Network visualisation of associations between protein domains and chemical classes extracted from CHAMOIS's logistic regression model coefficients (Supplementary Table 2). Each class (circle) is linked to Pfam domains (square) with weights >=2.0. Classes are coloured according to their highest-level ancestor in the ChemOnt taxonomy. Pfam domains are coloured by their top-level EC number, obtained with the ECDomainMiner tool(51). Several subgraphs feature associations supported by the literature (insets *i* to *iv, vi* to *ix*, and *xi* to *xiv)*, but also associations suggesting novel functions of uncharacterized domains (inset *v* and *x*, Table 1*)*. An interactive version of this panel can be found as part of CHAMOIS' documentation at https://chamois.readthedocs.io/en/v0.2.0/figures/network.html.

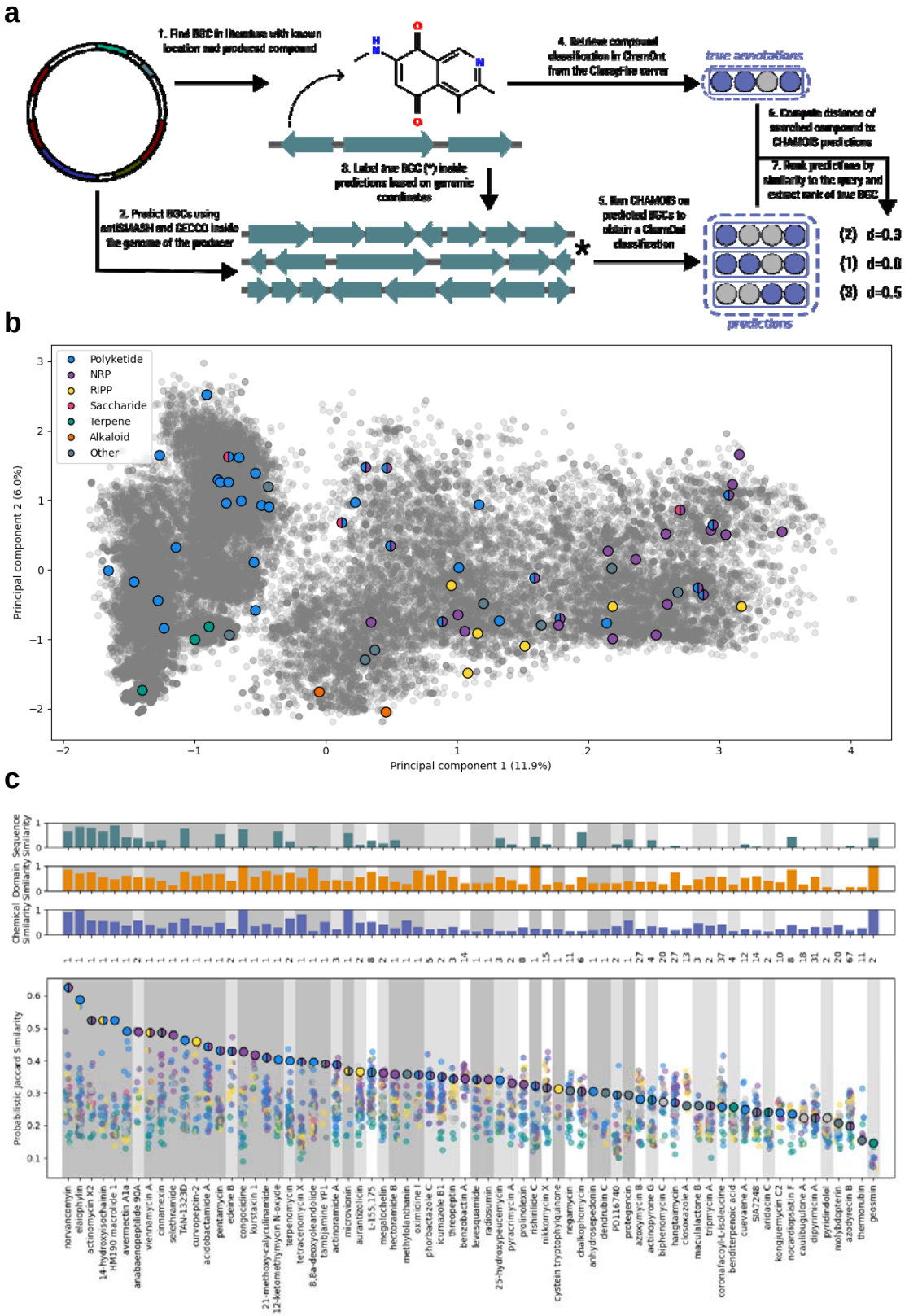

a
1. Find BGC in literature with known location and produced compound
2. Predict BGCs using antiSMASH and GECCO inside the genome of the producer
3. Label true BGC (*) inside predictions based on genomic coordinates
4. Retrieve compound classification in ChemOnt from the ClassyFire server
5. Run CHAMOIS on predicted BGCs to obtain a ClassOut classification
6. Compute distance of searched compound to CHAMOIS predictions
7. Rank predictions by similarity to the query and extract rank of true BGC
true annotations
predictions
(2) d=0.3
(1) d=0.0
(3) d=0.5
b
Polyketide
NRP
RiPP
Saccharide
Terpene
Alkaloid
Other
Principal component 2 (6.0%)
Principal component 1 (11.9%)
c
Sequence Similarity
Domain Similarity
Chemical Similarity
Probabilistic Jaccard Similarity
norvancomycin
elaiophylin
actinomycin X2
14-hydroxyisochainin
HM190 macrolide 1
avermectin A1a
anabaenopeptilide 90A
viennamycin A
cinnamexin
selethramide
TAN-1323D
curvopeptin-2
acidobactamide A
pentamycin
edeine B
congocidine
kurstakin 1
21-methoxy-calyculinamide
12-ketomethymycin N-oxyde
terpenomycin X
tetracenomycin X
8,8a-deoxyoleandolide
tambjamine YP1
actinoramide A
microvionin
aurantizolicin
L135,175
megalochelin
hectoramide B
methylolanthanin
oximidine I
phorbactazole C
icumazole B1
benzobactin A
threopeptin
levesquamide
radiosumin
25-hydroxypeucemycin
pyracrimycin A
prolinolexin
rishirilide C
nikkomycin X
cysteine tryptophylquinone
negamycin
chalkophomycin
anhydrosepedonin
dendrubin C
P0116740
protegencin
azoxymycin B
actinopyrone G
biphenomycin C
hangtaimycin
closoxazole A
maculalactone B
tripmycin A
coronafacoyl-L-isoleucine
benditerpenoic acid
cuevaene A
SIA7248
aridacin C
konguemycin C2
nocardiopsistin F
caulibugulone A
dipyrimicin A
pyridindolol
molybdopterin
azodyrecin B
thermorubin
geosmin

Figure 4: CHAMOIS can screen BGCs from within a given producer genome to pinpoint the cluster producing a given metabolite. **(a)** Graphical depiction of the prediction and evaluation framework. Shortly, as ground truth, experimentally validated BGCs with a known compound and found in a complete genome were extracted from literature (N=67). The known compounds were annotated with ClassyFire to obtain their reference ChemOnt classification. The source genomes were annotated for putative BGCs using antiSMASH (v7.0) and GECCO (v0.9.10) (see Methods). BGC predictions (N=2,331) were then passed to CHAMOIS to predict their ChemOnt classification. For each true compound, all predictions for a given genome were ranked using Probabilistic Jaccard similarity(40) between the predicted ChemOnt classes and the reference ChemOnt classification. **(b)** Principal component analysis (PCA) of ChemOnt class fingerprints highlighting compound diversity of the examples included in this benchmark (see *Contextualization of benchmark data* section of the Methods). NPAtlas compounds(39) are shown as gray background, while the compounds used for benchmarking are coloured by MIBiG classes. **(c)** Evaluation of CHAMOIS predictions using Probabilistic Jaccard similarity (y-axis) between predicted and true compounds for each producer genome (labeled on the x-axis). Each dot represents a BGC prediction, coloured by GECCO or antiSMASH predicted type (see colour code in (b)). The BGC producing the true compound is highlighted with an enlarged circle, and its rank among the predictions indicated above. A dark grey background indicates molecules where CHAMOIS correctly ranked the producing BGC with the highest similarity to the query (N=35, ~52%), light-grey backgrounds highlight instances where the true compound ranked among the top five predictions (together N=50, ~75%). The ranking produced by CHAMOIS is highly significant (P=1e-5, empirical permutation test with N=1e5 randomized rankings, see *BGC ranking significance* section of the Methods). The two barplots on top indicate sequence similarity (teal) and chemical similarity (lavender) to the closest examples in MIBiG 3.1 (used to train CHAMOIS) assessed by average nucleotide identity and Hamming distance between MHFP6 fingerprints, respectively (see *Establishing dissimilarity between benchmark and training sets* section of the Methods).

References

1. Fischbach MA, Walsh CT, Clardy J. The evolution of gene collectives: How natural selection drives chemical innovation. Proc Natl Acad Sci. 2008 Mar 25;105(12):4601–8.

2. Walsh CT, Tang Y. Natural product biosynthesis: chemical logic and enzymatic machinery. 2nd edition. London: Royal Society of Chemistry; 2023. 794 p.

3. Crits-Christoph A, Bhattacharya N, Olm MR, Song YS, Banfield JF. Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity. Genome Res. 2021 Feb;31(2):239–50.

4. Brady SF. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. Nat Protoc. 2007;2(5):1297–305.

5. Stohl EA, Milner JL, Handelsman J. Zwittermicin A biosynthetic cluster. Gene. 1999 Sept 17;237(2):403–11.

6. Blin K, Shaw S, Vader L, Szenei J, Reitz ZL, Augustijn HE, et al. antiSMASH 8.0: extended gene cluster detection capabilities and analyses of chemistry, enzymology, and regulation. Nucleic Acids Res. 2025 July 7;53(W1):W32–8.

7. Zdouc MM, Blin K, Louwen NLL, Navarro J, Loureiro C, Bader CD, et al. MIBiG 4.0: advancing biosynthetic gene cluster curation through global collaboration. Nucleic Acids Res. 2025 Jan 6;53(D1):D678–90.

8. Cimermancic P, Medema MH, Claesen J, Kurita K, Wieland Brown LC, Mavrommatis K, et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell. 2014 July 17;158(2):412–21.

9. Hannigan GD, Prihoda D, Palicka A, Soukup J, Klempir O, Rampula L, et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res. 2019 Oct 10;47(18):e110–e110.

10. Carroll LM, Larralde M, Fleck JS, Ponnudurai R, Milanese A, Cappio E, et al. Accurate de novo identification of biosynthetic gene clusters with GECCO. BioRxiv Prepr Serv Biol [Internet]. 2021; Available from: https://www.biorxiv.org/content/early/2021/05/04/2021.05.03.442509

11. Sanchez S, Rogers JD, Rogers AB, Nassar M, McEntyre J, Welch M, et al. Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS [Internet]. bioRxiv; 2023 [cited 2024 Sept 16]. p. 2023.05.23.540769. Available from: https://www.biorxiv.org/content/10.1101/2023.05.23.540769v3

12. Blin K, Medema MH, Kottmann R, Lee SY, Weber T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. Nucleic Acids Res. 2017 Jan 4;45(D1):D555–9.

13. Fullam A, Letunic I, Schmidt TSB, Ducarmon QR, Karcher N, Khedkar S, et al. proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. Nucleic Acids Res. 2023 Jan 6;51(D1):D760–6.

14. Liu T, Kharel MK, Zhu L, Bright SA, Mattingly C, Adams VR, et al. Inactivation of the ketoreductase gilU gene of the gilvocarcin biosynthetic gene cluster yields new analogues with partly improved biological activity. Chembiochem Eur J Chem Biol. 2009 Jan 26;10(2):278–86.

15. Huo L, Hug JJ, Fu C, Bian X, Zhang Y, Müller R. Heterologous expression of bacterial natural product biosynthetic pathways. Nat Prod Rep. 2019 Oct 16;36(10):1412–36.

16. Nivina A, Yuet KP, Hsu J, Khosla C. Evolution and Diversity of Assembly-Line Polyketide Synthases: Focus Review. Chem Rev. 2019 Dec 26;119(24):12524–47.

17. Marahiel MA, Stachelhaus T, Mootz HD. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. Chem Rev. 1997 Nov 1;97(7):2651–74.

18. Stachelhaus T, Mootz HD, Marahiel MA. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chem Biol. 1999 Aug;6(8):493–505.

19. Röttig M, Medema MH, Blin K, Weber T, Rausch C, Kohlbacher O. NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Res. 2011 July 1;39(Web Server issue):W362–7.

20. Kunyavskaya O, Tagirdzhanov AM, Caraballo-Rodríguez AM, Nothias LF, Dorrestein PC, Korobeynikov A, et al. Nerpa: A Tool for Discovering Biosynthetic Gene Clusters of Bacterial Nonribosomal Peptides. Metabolites. 2021 Oct 11;11(10):693.

21. Agrawal P, Khater S, Gupta M, Sain N, Mohanty D. RiPPMiner: a bioinformatics resource for deciphering chemical structures of RiPPs based on prediction of cleavage and cross-links. Nucleic Acids Res. 2017 July 3;45(W1):W80–8.

22. Helfrich EJN, Ueoka R, Dolev A, Rust M, Meoded RA, Bhushan A, et al. Automated structure prediction of trans-acyltransferase polyketide synthase products. Nat Chem Biol. 2019 June 17;15(8):813–21.

23. Skinnider MA, Johnston CW, Gunabalasingam M, Merwin NJ, Kieliszek AM, MacLellan RJ, et al. Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. Nat Commun. 2020 Nov 27;11(1):6058.

24. Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminformatics. 2016;8:61.

25. Probst D, Reymond JL. A probabilistic molecular fingerprint for big data settings. J Cheminformatics. 2018 Dec 18;10(1):66.

26. Tibshirani R. Regression Shrinkage and Selection via the Lasso. J R Stat Soc Ser B Methodol. 1996;58(1):267–88.

27. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010 Mar 8;11:119.

28. Larralde M. Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. J Open Source Softw. 2022 Apr 25;7(72):4296.

29. Paysan-Lafosse T, Andreeva A, Blum M, Chuguransky SR, Grego T, Pinto BL, et al. The Pfam protein families database: embracing AI/ML. Nucleic Acids Res. 2025 Jan 6;53(D1):D523–34.

30. Larralde M, Zeller G. PyHMMER: A Python library binding to HMMER for efficient sequence analysis. Bioinformatics. 2023 Apr 19;btad214.

31. Walker AS, Clardy J. A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters. J Chem Inf Model. 2021 June 28;61(6):2560–71.

32. Zhang M, Zhang P, Xu G, Zhou W, Gao Y, Gong R, et al. Comparative Investigation into Formycin A and Pyrazofurin A Biosynthesis Reveals Branch Pathways for the Construction of C-

Nucleoside Scaffolds. Appl Environ Microbiol. 2020 Jan 7;86(2):e01971-19.

33. Kharel MK, Subba B, Basnet DB, Woo JS, Lee HC, Liou K, et al. A gene cluster for biosynthesis of kanamycin from *Streptomyces kanamyceticus*: comparison with gentamicin biosynthetic gene cluster. Arch Biochem Biophys. 2004 Sept 15;429(2):204–14.

34. Hamming RW. Error detecting and error correcting codes. Bell Syst Tech J. 1950 Apr;29(2):147–60.

35. Hendry S, Steinke S, Wittstein K, Stadler M, Harmrolfs K, Adewunmi Y, et al. Functional Analysis of Phenazine Biosynthesis Genes in Burkholderia spp. Appl Environ Microbiol. 2021 May 11;87(11):e02348-20.

36. Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, et al. The InterPro protein families and domains database: 20 years on. Nucleic Acids Res. 2020 Nov 6;49(D1):D344–54.

37. Burkhart BJ, Schwalen CJ, Mann G, Naismith JH, Mitchell DA. YcaO-Dependent Posttranslational Amide Activation: Biosynthesis, Structure, and Function. Chem Rev. 2017 Apr 26;117(8):5389–456.

38. Skinnider M. Gold standard BGCs [Internet]. Zenodo; 2020 [cited 2024 Sept 16]. Available from: https://zenodo.org/record/3985982

39. 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 Jan 7;50(D1):D1317–23.

40. Martire I, da Silva PN, Plastino A, Fabris F, Freitas AA. A novel probabilistic Jaccard distance measure for classification of sparse and uncertain data. In: Rebeiro de Faria Paiva E, Merschmann L, Cerri R, editors. Uberlandia, MG, Brazil; 2017 [cited 2025 Feb 11]. p. 81–8. Available from: http://www.facom.ufu.br/~kdmile/proceedings/anais-kdmile-2017.pdf

41. Segal MR, Dahlquist KD, Conklin BR. Regression approaches for microarray data analysis. J Comput Biol J Comput Mol Cell Biol. 2003;10(6):961–80.

42. Terlouw BR, Blin K, Navarro-Muñoz JC, Avalon NE, Chevrette MG, Egbert S, et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. Nucleic Acids Res. 2023 Jan 6;51(D1):D603–10.

43. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021 Jan 8;49(D1):D412–9.

44. Virshup I, Rybakov S, Theis FJ, Angerer P, Wolf FA. anndata: Access and store annotated data matrices. J Open Source Softw. 2024 Sept 16;9(101):4371.

45. van Santen JA, Jacob G, Singh AL, Aniebok V, Balunas MJ, Bunsko D, et al. The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. ACS Cent Sci. 2019 Nov 27;5(11):1824–33.

46. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2023 update. Nucleic Acids Res. 2023 Jan 6;51(D1):D1373–80.

47. Landrum G, Tosco P, Kelley B, Rodriguez R, Cosgrove D, Vianello R, et al. rdkit/rdkit: 2024_09_6 (Q3 2024) Release [Internet]. Zenodo; 2025 [cited 2025 Mar 12]. Available from: https://zenodo.org/records/14943932

48. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12(85):2825–30.

49. Fan RE, Chang KW, Hsieh CJ, Wang XR, Lin CJ. LIBLINEAR: A Library for Large Linear Classification. J Mach Learn Res. 2008 June 1;9:1871–4.

50. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007 May;9(3):90–5.

51. Alborzi SZ, Devignes MD, Ritchie DW. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. BMC Bioinformatics. 2017 Feb 13;18(1):107.

52. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan 8;47(D1):D330–8.

53. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar;17(3):261–72.

54. Skinnider MA, Dejong CA, Rees PN, Johnston CW, Li H, Webster ALH, et al. Genomes to natural products PRediction Informatics for Secondary Metabolomes (PRISM). Nucleic Acids Res. 2015 Nov 16;43(20):9645–62.

55. Li MH, Ung PM, Zajkowski J, Garneau-Tsodikova S, Sherman DH. Automated genome mining for natural products. BMC Bioinformatics. 2009 June 16;10:185.

56. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0-improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017 July 3;45(W1):W36–41.

57. van Santen JA, Linington RG, Contributors NA. The Natural Products Atlas - data download [Internet]. Zenodo; 2024 [cited 2025 Feb 21]. Available from: https://zenodo.org/records/13756408

58. Shaw J, Yu YW. Fast and robust metagenomic sequence comparison through sparse chaining with skani. Nat Methods. 2023 Nov;20(11):1661–5.

59. Larralde M, Zeller G, Carroll LM. PyOrthoANI, PyFastANI, and Pyskani: a suite of Python libraries for computation of average nucleotide identity [Internet]. bioRxiv; 2025 [cited 2025 Mar 12]. p. 2025.02.13.638148. Available from: https://www.biorxiv.org/content/10.1101/2025.02.13.638148v1

60. Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods. 2018 July;15(7):475–6.

61. Adak S, Lukowski AL, Schäfer RJB, Moore BS. From Tryptophan to Toxin: Nature's Convergent Biosynthetic Strategy to Aetokthonotoxin. J Am Chem Soc. 2022 Feb 23;144(7):2861–6.

62. Lingkon K, Bellizzi JJ. Structure and Activity of the Thermophilic Tryptophan-6 Halogenase BorH. Chembiochem Eur J Chem Biol. 2020 Apr 17;21(8):1121–8.

63. Barry SM, Kers JA, Johnson EG, Song L, Aston PR, Patel B, et al. Cytochrome P450–catalyzed L-tryptophan nitration in thaxtomin phytotoxin biosynthesis. Nat Chem Biol. 2012 Oct;8(10):814–6.

64. Kupke T, Stevanović S, Sahl HG, Götz F. Purification and characterization of EpiD, a flavoprotein involved in the biosynthesis of the lantibiotic epidermin. J Bacteriol. 1992 Aug 1;174(16):5354–61.

65. Bloudoff K, Schmeing TM. Structural and functional aspects of the nonribosomal peptide synthetase condensation domain superfamily: discovery, dissection and diversity. Biochim Biophys Acta BBA - Proteins Proteomics. 2017 Nov 1;1865(11, Part B):1587–604.

66. Sauguet L, Moutiez M, Li Y, Belin P, Seguin J, Le Du MH, et al. Cyclodipeptide synthases, a family of class-I aminoacyl-tRNA synthetase-like enzymes involved in non-ribosomal peptide synthesis. Nucleic Acids Res. 2011 May 1;39(10):4475–89.

67. Wang X, Zhou H, Ren X, Chen H, Zhong L, Bai X, et al. Recombineering enables genome mining of novel siderophores in a non-model Burkholderiales strain. Eng Microbiol. 2023 Aug 2;3(3):100106.

68. Bacher A, Eberhardt S, Fischer M, Kis K, Richter G. Biosynthesis of Vitamin B2 (Riboflavin). Annu Rev Nutr. 2000 July 1;20(Volume 20, 2000):153–67.

69. Ngivprom U, Kluaiphanngam S, Ji W, Siriwibool S, Kamkaew A, Ketudat Cairns JR, et al. Characterization of NucPNP and NucV involved in the early steps of nucleocidin biosynthesis in Streptomyces calvus. RSC Adv. 11(6):3510–5.

70. Weinig S, Hecht HJ, Mahmud T, Müller R. Melithiazol biosynthesis: further insights into myxobacterial PKS/NRPS systems and evidence for a new subclass of methyl transferases. Chem Biol. 2003 Oct;10(10):939–52.

71. Mihara K, Tanabe T, Yamakawa Y, Funahashi T, Nakao H, Narimatsu S, et al. Identification and transcriptional organization of a gene cluster involved in biosynthesis and transport of acinetobactin, a siderophore produced by Acinetobacter baumannii ATCC 19606T. Microbiol Read Engl. 2004 Aug;150(Pt 8):2587–97.

72. Nair IM, Kochupurackal J. Squalene hopene cyclases and oxido squalene cyclases: potential targets for regulating cyclisation reactions. Biotechnol Lett. 2023 June 1;45(5):573–88.

73. Antônio RV, Creczynski-Pasa TB. Genetic analysis of violacein biosynthesis by Chromobacterium violaceum. Genet Mol Res GMR. 2004 Mar 31;3(1):85–91.

74. Rowland BM, Taber HW. Duplicate isochorismate synthase genes of Bacillus subtilis: regulation and involvement in the biosyntheses of menaquinone and 2,3-dihydroxybenzoate. J Bacteriol. 1996 Feb;178(3):854–61.

75. Petras D, Kerwat D, Pesic A, Hempel BF, von Eckardstein L, Semsary S, et al. The O-Carbamoyl-Transferase Alb15 Is Responsible for the Modification of Albicidin. ACS Chem Biol. 2016 May 20;11(5):1198–204.

76. Moncrieffe MC, Fernandez MJ, Spiteller D, Matsumura H, Gay NJ, Luisi BF, et al. Structure of the Glycosyltransferase EryCIII in Complex with its Activating P450 Homologue EryCII. J Mol Biol. 2012 Jan 6;415(1):92–101.

77. Gaur R, Varshney U. Genetic Analysis Identifies a Function for the queC (ybaX) Gene Product at an Initial Step in the Queuosine Biosynthetic Pathway in Escherichia coli. J Bacteriol. 2005 Oct 15;187(20):6893–901.

78. Palma PN, Rodrigues ML, Archer M, Bonifácio MJ, Loureiro AI, Learmonth DA, et al. Comparative Study of ortho- and meta-Nitrated Inhibitors of Catechol-O-methyltransferase: Interactions with the Active Site and Regioselectivity of O-Methylation. Mol Pharmacol. 2006 July 1;70(1):143–53.

79. Liu A, Si Y, Dong SH, Mahanta N, Penkala HN, Nair SK, et al. Functional elucidation of TfuA in peptide backbone thioamidation. Nat Chem Biol. 2021 May;17(5):585–92.

80. Choi YS, Zhang H, Brunzelle JS, Nair SK, Zhao H. In vitro reconstitution and crystal structure of p-aminobenzoate N-oxygenase (AurF) involved in aureothin biosynthesis. Proc Natl Acad Sci. 2008 May 13;105(19):6858–63.

81. Urbach C, Evrard C, Pudzaitis V, Fastrez J, Soumillion P, Declercq JP. Structure of PBP-A from Thermosynechococcus elongatus, a Penicillin-Binding Protein Closely Related to Class A β-

Lactamases. J Mol Biol. 2009 Feb;386(1):109–20.

82. Soto-Liebe K, Murillo AA, Krock B, Stucken K, Fuentes-Valdés JJ, Trefault N, et al. Reassessment of the toxin profile of Cylindrospermopsis raciborskii T3 and function of putative sulfotransferases in synthesis of sulfated and sulfonated PSP toxins. Toxicon Off J Int Soc Toxinology. 2010 Dec;56(8):1350–61.

83. Ahlert J, Distler J, Mansouri K, Piepersberg W. Identification of stsC, the gene encoding the L-glutamine:scyllo-inosose aminotransferase from streptomycin-producing Streptomycetes. Arch Microbiol. 1997 Aug 1;168(2):102–13.

84. Sultana A, Kallio P, Jansson A, Wang J, Niemi J, Mäntsälä P, et al. Structure of the polyketide cyclase SnoaL reveals a novel mechanism for enzymatic aldol condensation. EMBO J. 2004 May 5;23(9):1911–21.

85. Meneely KM, Ronnebaum TA, Riley AP, Prisinzano TE, Lamb AL. Holo Structure and Steady State Kinetics of the Thiazolinyl Imine Reductases for Siderophore Biosynthesis. Biochemistry. 2016 Sept 27;55(38):5423–33.

86. Cogan DP, Hudson GA, Zhang Z, Pogorelov TV, van der Donk WA, Mitchell DA, et al. Structural insights into enzymatic [4+2] aza-cycloaddition in thiopeptide antibiotic biosynthesis. Proc Natl Acad Sci U S A. 2017 Dec 5;114(49):12928–33.

87. Mori T, Hoshino S, Sahashi S, Wakimoto T, Matsui T, Morita H, et al. Structural Basis for β-Carboline Alkaloid Production by the Microbial Homodimeric Enzyme McbB. Chem Biol. 2015 July;22(7):898–906.

88. Dunbar KL, Chekan JR, Cox CL, Burkhart BJ, Nair SK, Mitchell DA. Discovery of a new ATP-binding motif involved in peptidic azoline biosynthesis. Nat Chem Biol. 2014 Oct 1;10(10):823–9.

89. Zhao G, Kosek D, Liu HB, Ohlemacher SI, Blackburne B, Nikolskaya A, et al. Structural Basis for a Dual Function ATP Grasp Ligase That Installs Single and Bicyclic ω-Ester Macrocycles in a New Multicore RiPP Natural Product. J Am Chem Soc. 2021 June 1;143(21):8056–68.

90. Schwartz D, Recktenwald J, Pelzer S, Wohlleben W. Isolation and characterization of the PEP-phosphomutase and the phosphonopyruvate decarboxylase genes from the phosphinothricin tripeptide producer *Streptomyces viridochromogenes* TÃ¼494. FEMS Microbiol Lett. 1998 June;163(2):149–57.

91. Zheng Z, Clardy J, Liu H wen. Biosynthesis of the Unusual Epoxy Isonitrile-Containing Antibiotics Aerocyanidin and Amycomicin. J Am Chem Soc. 2024 July 31;146(30):21061–8.

92. Wang SY, Chiu KW, Lin KL, Wei HY, Chen YR, Tu Z, et al. Epoxide Stereochemistry Controls Regioselective Ketoreduction in Epoxyquinoid Biosynthesis. J Am Chem Soc. 2025 Aug 13;147(32):29582–91.

93. Blanco G, Brian P, Pereda A, Méndez C, Salas JA, Chater KF. Hybridization and DNA sequence analyses suggest an early evolutionary divergence of related biosynthetic gene sets encoding polyketide antibiotics and spore pigments in Streptomyces spp. Gene. 1993 Aug 16;130(1):107–16.

94. Liu J, Zhu X, Kim SJ, Zhang W. Antimycin-type depsipeptides: discovery, biosynthesis, chemical synthesis, and bioactivities. Nat Prod Rep. 2016;33(10):1146–65.

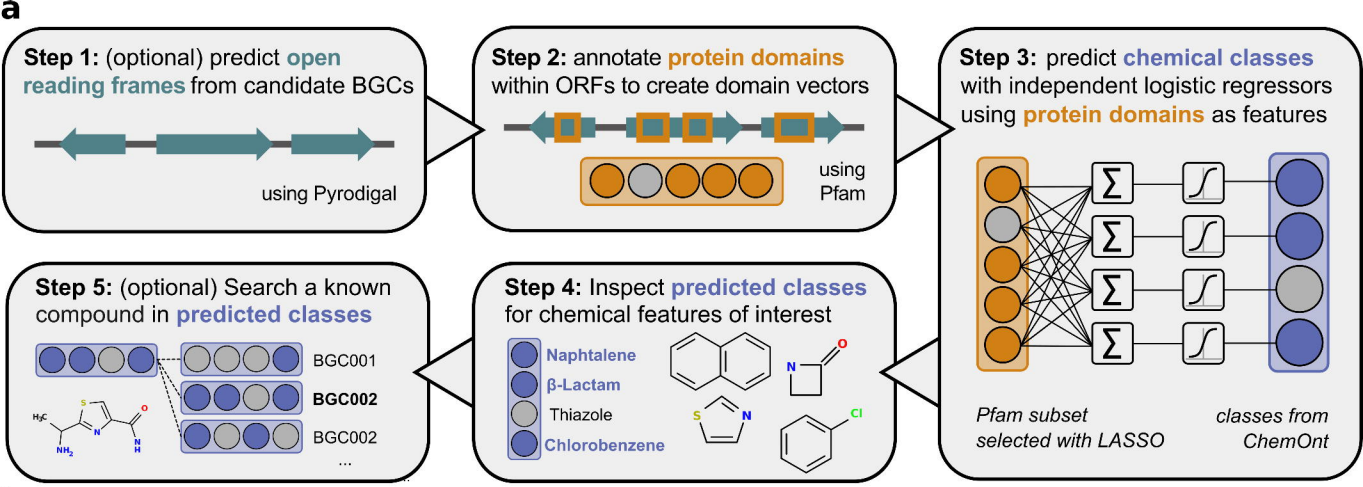
95. Chen PYT, Adak S, Chekan JR, Liscombe DK, Miyanaga A, Bernhardt P, et al. Structural Basis of Stereospecific Vanadium-Dependent Haloperoxidase Family Enzymes in Napyradiomycin Biosynthesis. Biochemistry. 2022 Sept 6;61(17):1844–52.

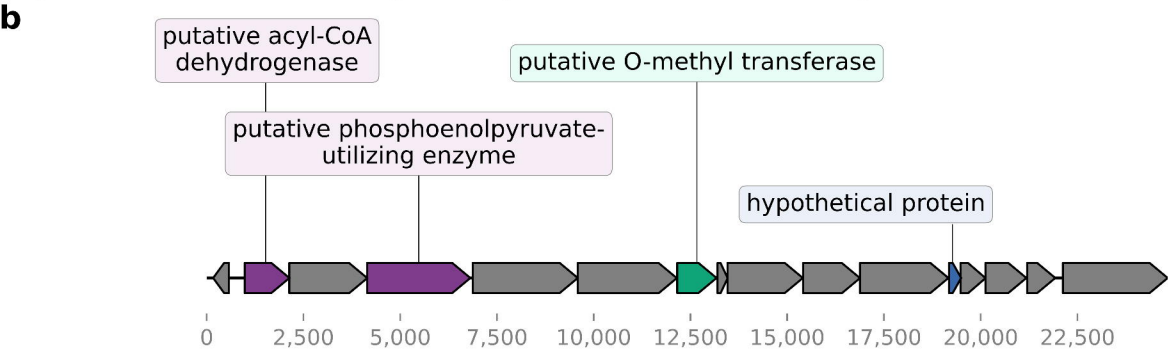
96. Harris AKP, Williamson NR, Slater H, Cox A, Abbasi S, Foulds I, et al. The Serratia gene cluster encoding biosynthesis of the red antibiotic, prodigiosin, shows species- and strain-dependent

genome context variation. Microbiol Read Engl. 2004 Nov;150(Pt 11):3547–60.

97. Terlouw BR, Vromans SPJM, Medema MH. PIKAChU: a Python-based informatics kit for analysing chemical units. J Cheminformatics. 2022 June 7;14(1):34.

98. Ayikpoe RS, Zhu L, Chen JY, Ting CP, van der Donk WA. Macrocyclization and Backbone Rearrangement During RiPP Biosynthesis by a SAM-Dependent Domain-of-Unknown-Function 692. ACS Cent Sci. 2023 May 24;9(5):1008–18.

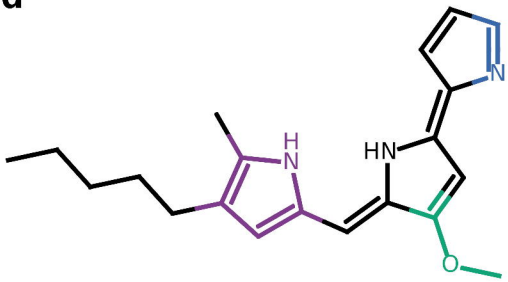


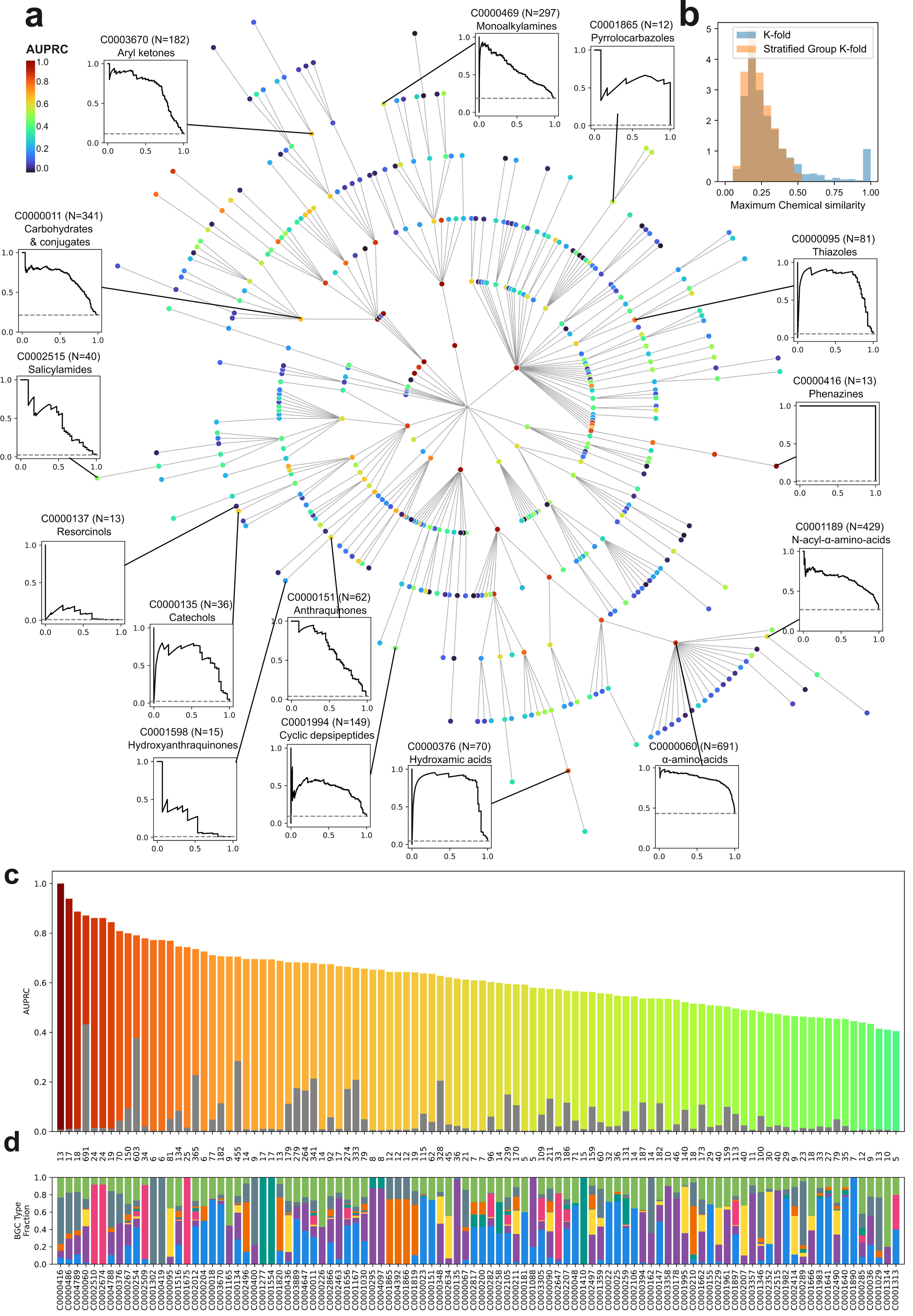


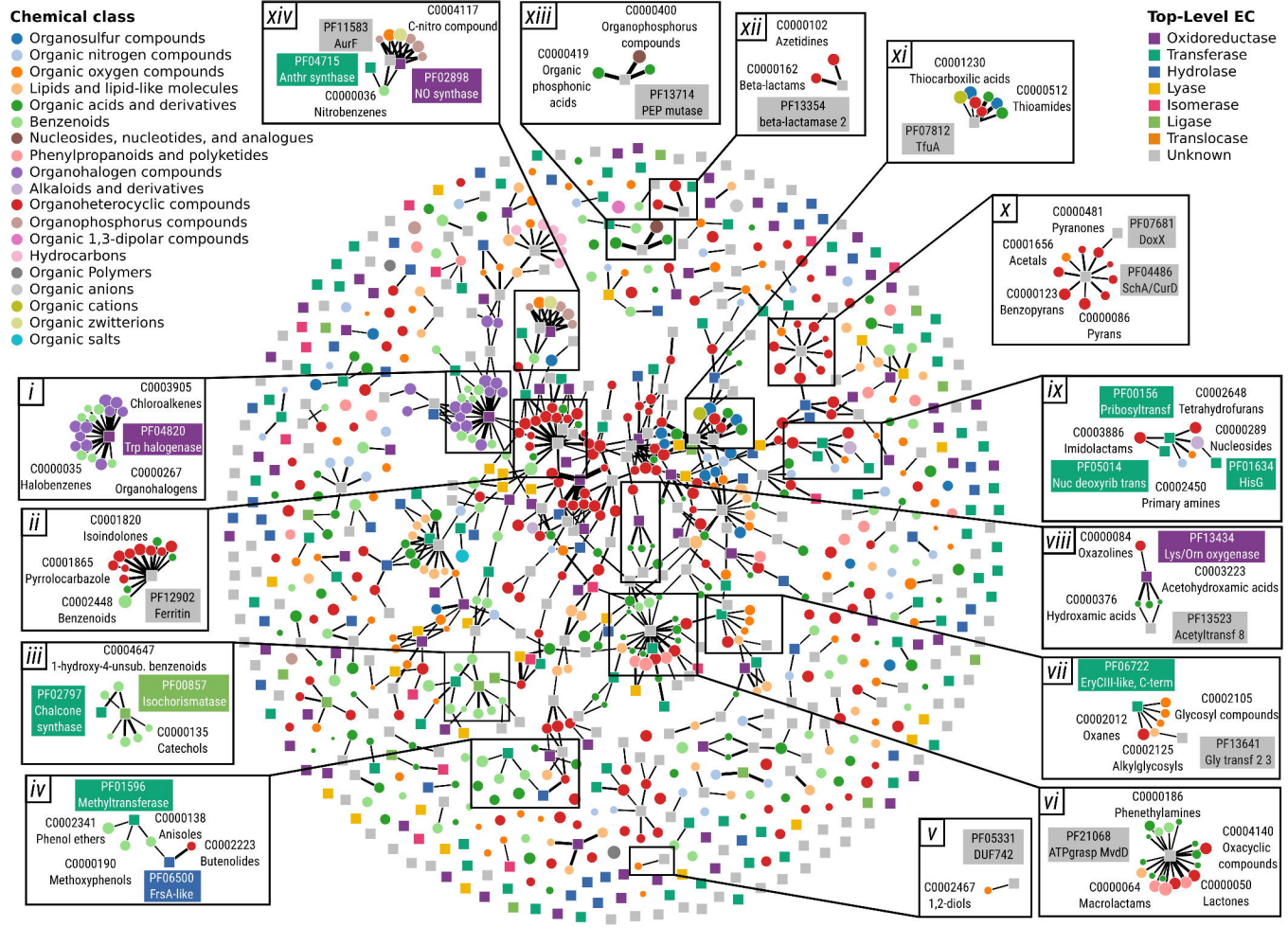
c

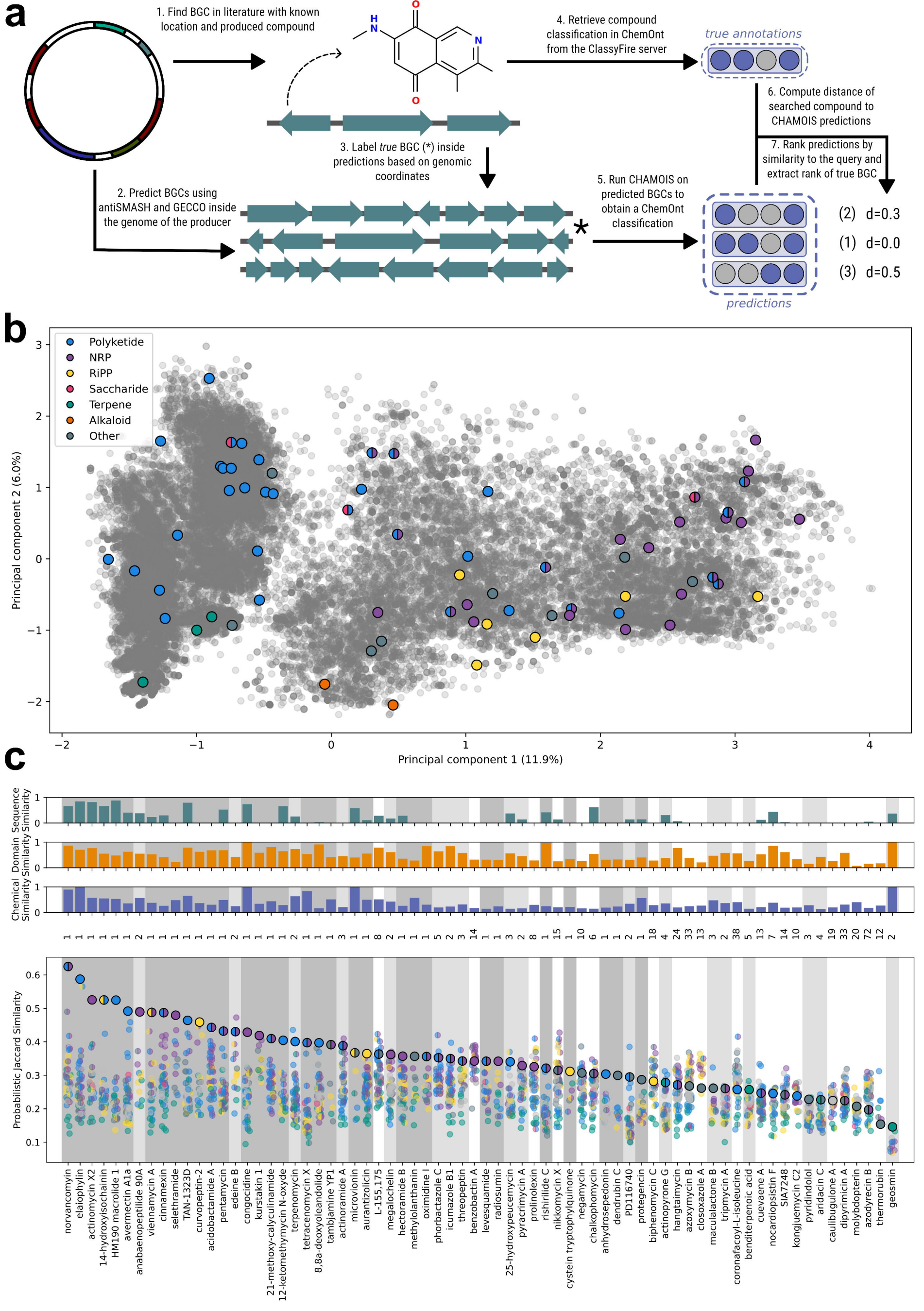
```
CHEMONTID:0000002 (Organoheterocyclic compounds): 0.933
├── CHEMONTID:0000090 (Pyrroles): 0.844
│   └── CHEMONTID:0002257 (Substituted pyrroles): 0.844
├── CHEMONTID:0004139 (Azacyclic compounds): 0.861
└── CHEMONTID:0004144 (Heteroaromatic compounds): 0.933
CHEMONTID:0004150 (Hydrocarbon derivatives): 0.982
CHEMONTID:0004603 (Organic oxygen compounds): 0.991
└── CHEMONTID:0000323 (Organooxygen compounds): 0.991
    └── CHEMONTID:0000254 (Ethers): 0.834
        └── CHEMONTID:0000128 (Alkyl aryl ethers): 0.834
CHEMONTID:0004707 (Organic nitrogen compounds): 0.993
└── CHEMONTID:0000278 (Organonitrogen compounds): 0.993
    └── CHEMONTID:0000117 (Imines): 0.892
CHEMONTID:0003630 (Organic 1,3-dipolar compounds): 0.806
└── CHEMONTID:0003633 (Propargyl-type 1,3-dipolar organic compounds): 0.806
```

d




a
AUPRC
C0003670 (N=182)
Aryl ketones
C0000469 (N=297)
Monoalkylamines
C0001865 (N=12)
Pyrrolocarbazoles
b
K-fold
Stratified Group K-fold
Maximum Chemical similarity
C0000011 (N=341)
Carbohydrates
& conjugates
C0000095 (N=81)
Thiazoles
C0002515 (N=40)
Salicylamides
C0000416 (N=13)
Phenazines
C0000137 (N=13)
Resorcinols
C0001189 (N=429)
N-acyl-α-amino-acids
C0000135 (N=36)
Catechols
C0000151 (N=62)
Anthraquinones
C0001598 (N=15)
Hydroxyanthraquinones
C0001994 (N=149)
Cyclic depsipeptides
C0000376 (N=70)
Hydroxamic acids
C0000060 (N=691)
α-amino-acids
c
AUPRC
d
BGC Type
Fraction


Chemical class
Organosulfur compounds
Organic nitrogen compounds
Organic oxygen compounds
Lipids and lipid-like molecules
Organic acids and derivatives
Benzenoids
Nucleosides, nucleotides, and analogues
Phenylpropanoids and polyketides
Organohalogen compounds
Alkaloids and derivatives
Organoheterocyclic compounds
Organophosphorus compounds
Organic 1,3-dipolar compounds
Hydrocarbons
Organic Polymers
Organic anions
Organic cations
Organic zwitterions
Organic salts
Top-Level EC
Oxidoreductase
Transferase
Hydrolase
Lyase
Isomerase
Ligase
Translocase
Unknown
xiv
C0004117
C-nitro compound
PF11583
AurF
PF04715
Anthr synthase
PF02898
NO synthase
C0000036
Nitrobenzenes
xiii
C0000400
Organophosphorus compounds
C0000419
Organic phosphonic acids
PF13714
PEP mutase
xii
C0000102
Azetidines
C0000162
Beta-lactams
PF13354
beta-lactamase 2
xi
C0001230
Thiocarboxilic acids
C0000512
Thioamides
PF07812
TfuA
x
C0000481
Pyranones
PF07681
DoxX
C0001656
Acetals
PF04486
SchA/CurD
C0000123
Benzopyrans
C0000086
Pyrans
i
C0003905
Chloroalkenes
PF04820
Trp halogenase
C0000035
Halobenzenes
C0000267
Organohalogens
ii
C0001820
Isoindolones
C0001865
Pyrrolocarbazole
C0002448
Benzenoids
PF12902
Ferritin
iii
C0004647
1-hydroxy-4-unsub. benzenoids
PF02797
Chalcone synthase
PF00857
Isochorismatase
C0000135
Catechols
iv
PF01596
Methyltransferase
C0002341
Phenol ethers
C0000138
Anisoles
C0002223
Butenolides
C0000190
Methoxyphenols
PF06500
FrsA-like
v
PF05331
DUF742
C0002467
1,2-diols
vi
C0000186
Phenethylamines
PF21068
ATPgrasp MvdD
C0004140
Oxacyclic compounds
C0000064
Macrolactams
C0000050
Lactones
vii
PF06722
EryCIII-like, C-term
C0002105
Glycosyl compounds
C0002012
Oxanes
C0002125
Alkylglycosyls
PF13641
Gly transf 2 3
viii
C0000084
Oxazolines
PF13434
Lys/Orn oxygenase
C0003223
Acetohydroxamic acids
C0000376
Hydroxamic acids
PF13523
Acetyltransf 8
ix
PF00156
Pribosyltransf
C0002648
Tetrahydrofurans
C0003886
Imidolactams
C0000289
Nucleosides
PF05014
Nuc deoxyrib trans
PF01634
HisG
C0002450
Primary amines


a
1. Find BGC in literature with known location and produced compound
2. Predict BGCs using antiSMASH and GECCO inside the genome of the producer
3. Label true BGC (*) inside predictions based on genomic coordinates
4. Retrieve compound classification in ChemOnt from the ClassyFire server
5. Run CHAMOIS on predicted BGCs to obtain a ChemOnt classification
6. Compute distance of searched compound to CHAMOIS predictions
7. Rank predictions by similarity to the query and extract rank of true BGC
true annotations
predictions
(2) d=0.3
(1) d=0.0
(3) d=0.5
b
Polyketide
NRP
RiPP
Saccharide
Terpene
Alkaloid
Other
Principal component 2 (6.0%)
Principal component 1 (11.9%)
c
Sequence Similarity
Domain Similarity
Chemical Similarity
Probabilistic Jaccard Similarity
norvancomyin
elaiophylin
actinomycin X2
14-hydroxyisochainin
HM190 macrolide 1
avermectin A1a
anabaenopeptilide 90A
viennamycin A
cinnamexin
selethramide
TAN-1323D
curvopeptin-2
acidobactamide A
pentamycin
edeine B
congocidine
kurstakin 1
21-methoxy-calyculinamide
12-ketomethymycin N-oxyde
terpenomycin
tetracenomycin X
8,8a-deoxyoleandolide
tambjamine YP1
actinoramide A
microvionin
aurantizolicin
L-155,175
megalochelin
hectoramide B
methylolanthanin
oximidine I
phorbactazole C
icumazole B1
threopeptin
benzobactin A
levesquamide
radiosumin
25-hydroxypeucemycin
pyracrimycin A
prolinolexin
rishirilide C
nikkomycin X
cystein tryptophylquinone
negamycin
chalkophomycin
anhydrosepedonin
dendrubin C
PD116740
protegencin
biphenomycin C
actinopyrone G
hangtaimycin
azoxymycin B
closoxazole A
maculalactone B
tripmycin A
coronafacoyl-L-isoleucine
benditerpenoic acid
cuevaene A
nocardiopsistin F
SIA7248
kongjuemycin C2
pyridindolol
aridacin C
caulibugulone A
dipyrimicin A
molybdopterin
azodyrecin B
thermorubin
geosmin