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

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

[1] Structural and Computational 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 to accurately predict which secondary metabolites result from these are critically lacking. Here, we present CHAMOIS, a machine learning-based tool for predicting chemical properties of secondary metabolites from protein domains annotated in the input BGCs. CHAMOIS infers 485 chemical properties from the ChemOnt ontology using logistic regression. It accurately predicts 111 such properties (AUPRC > 0.5) in cross-validation against known instances. Although CHAMOIS is not explicitly trained on biosynthetic knowledge, many of the inferred links between protein domains and metabolite properties are consistent with scientific literature, others suggest new biochemical functions of uncharacterized biosynthetic domains. Finally, CHAMOIS can pinpoint which BGC within a given genome produces a pre-specified metabolite (correct BGC in 69% of cases ranked among the top 5), which holds great potential for prioritising experimental BGC characterisation and discovery of novel biosynthetic enzymes.

Introduction

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 3.1 (2022), however, numbers only 2,502 instances[7]. 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 (81 distinct types in v7.0). These types summarise the broad biosynthetic pathway, but reveal little detail about the final molecule itself. Currently available data-driven approaches for BGC categorisation assign BGCs to one or several of seven 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. 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=2,068 BGC corresponding to a total of N=1,156 ChemOnt classes, see Fig. 1 and *Sequence dataset preparation* section of the Online 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 rare classes (N<10) under the assumption that they were unpredictable, or too rare for satisfactory evaluation. The remaining classes (N=485) were used to train LASSO classifiers (see *Compound dataset preparation* section of the Online 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 36.0 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 3,585 dimensions) encoding presence/absence of each domain per BGC. 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). These binary feature vectors 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, 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[33] between MHFP6 fingerprints[25] to measure the distance of produced metabolites in chemical space (see *Compound dataset preparation* section of the Online 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=1,536 examples used for cross-validation) to perform stratified cross-validation preserving these groups, using 5 distinct folds for each of the 485 classifiers (see *Cross-validation* section of the Online Methods).

The median area under the receiver-operating characteristic (ROC) curve (AUROC) across all evaluated classes was 0.82 (Supplementary Table 1). However, because of the imbalance in the predicted classes (Supplementary Fig. 1), 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, 111 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). Among the classes with the highest AUPRC were many high-level classes represented by many training instances, such as Organonitrogens (CHEMONTID:0000278, AUPRC=0.968, N_{pos}=1,512) or Organoheterocyclic compounds (CHEMONTID:0000002, AUPRC=0.913, N_{pos}=1,708). However, the class with the highest AUPRC was Phenazines (CHEMONTID:0000416, AUPRC=1.0, N_{pos}=12), a rare class of tricyclic compounds known to be synthesised by a pathway of 7 conserved genes, *phzA-G*[34]. In addition, CHAMOIS exhibited good performance on other relatively rare classes (Fig. 2c, Supplementary Fig. 2): Cyclohexylamines (CHEMONTID:0002674, AUPRC=0.849, N_{pos}=24), Quinoxalines (CHEMONTID:0000486, AUPRC=0.818, N_{pos}=16), Indolocarbazoles (CHEMONTID:0001866, AUPRC=0.874, N_{pos}=12), or Hydroxamic acids (CHEMONTID:0000376, AUPRC=0.766, N_{pos}=80).

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 97, with a median of 54, 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=83, ~17%) or two domains (N=117, ~24%) (Supplementary Fig. 3, 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 Tryptophan halogenase domains (PF04820) linked to Organohalogen compounds (CHEMONTID:0000267) biosynthesis (for more examples see Table 1). 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 98.0[35]). In our dataset derived from MIBiG 3.1 a large fraction of the BGC-associated domains still lack functional annotations: 1842 of 2068 BGCs (~89%) contain at least one unannotated domain, also as a consequence of recent Pfam releases introducing new domains, such as DUF6194 (PF19694) that are present across several MIBiG BGCs (BGC0000595, BGC0000596, BGC0001597) 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 130 uncharacterized domains for which CHAMOIS inferred biosynthetic functions, for a total of 444 domain/class pairs (Supplementary Table 3, see *Identification of uncharacterized domains* section of the Online 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[36], yet it did not have any 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 Azolines (CHEMONTID:0002490) classifier (Table 1, Supplementary Table 3). 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 Phosphoenolpyruvate phosphomutase domain (PF13714), and the Cucumopine synthase C-terminal helical bundle domain (PF18631) (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 Seven-bladed beta-propeller (SBBP) repeat domain (PF06739, Fig. 3 inset *ix*) received a high weight from the Benzenoids (CHEMONTID:0002448, w=2.151) and Phenols (CHEMONTID:0000134, w=2.592) classifiers. While it was previously reported that a SBBP protein, CylK, was capable of cross-linking resorcinols via Friedel-Crafts alkylation[37], our results suggest more broadly that certain biosynthetic SBBP domains could be selective for benzene or phenol moieties. Among other domains of interest, DUF2378 (PF09536) was associated with the prediction of Pyrrolidines (CHEMONTID:000218, w=3.340), and the SchA/CurD-like domain (PF04486), found in polyketide clusters of *Streptomyces* with no function known, was associated by CHAMOIS with the biosynthesis of Naphthopyranones (CHEMONTID:0001641, w=2.243). 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 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 45 experimentally-annotated bacterial BGCs which had been identified in a complete bacterial genome (Fig. 4a) and most of which showed little 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. This resulted in 1,795 unique BGC predictions across all genomes (on average 37, ranging from 1 to 76 per genome). We then applied CHAMOIS to each BGC to predict a fingerprint consisting of the 485 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[38]. 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 Online Methods).

In this benchmark, CHAMOIS identified the correct cluster in 19 of the 45 genomes. In an additional 12 instances, it ranked the true BGC among the top 5 clusters. That CHAMOIS in two thirds 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 (<0.4, 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. 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 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 dataset, which was not part of the MIBiG-derived training data, and which mostly contained BGCs and metabolites without similarity to that (Fig. 4). A limitation of our method, which implicitly relies on protein domains involved in biosynthesis, is the prediction of RiPP BGC products due to several RiPP-specific challenges. They often produce large molecules, the complexity of which can sometimes not be fully captured with 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 discard the other[39]. 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.

Tables

Domain name	Domain accession	InterPro	EC*	GO*	Class name	Class accession	CHAMOIS weight	Reference	Fig. 3 inset
Pyridoxine 5'-phosphate oxidase C-terminal dimerisation region	PF10590	x	x		Phenazines and derivatives	CHEMONTID:0000416	7.743	[34]	*v*
					Quinoxalines	CHEMONTID:0000486	6.285		
					Benzodiazines	CHEMONTID:0004789	6.285		
Oligosaccharide biosynthesis protein Alg14 like	PF08660	x		x	Pyranoid amino-acids and derivatives	CHEMONTID:0001697	6.383	[40]	*iv*
					N-acyl-alpha-hexosamines	CHEMONTID:0002276	5.258	[41]	
Tryptophan halogenase	PF04820	x	x	x	Organohalogen compounds	CHEMONTID:0000267	5.539	[42]	*iii*
					Aryl halides	CHEMONTID:0002866	5.537	[43]	
Chorismatase FkbO/Hyg5-like, N-terminal	PF21168	x	x		Cyclohexanols	CHEMONTID:0002647	5.262	[44]	
Nitric oxide synthase, oxygenase domain	PF02898	x		x	Nitroaromatic compounds	CHEMONTID:0004612	5.200	[45]	
					Organic nitro compounds	CHEMONTID:0001152	4.978		
Ferritin-like	PF12902	x	x		Indolocarbazoles	CHEMONTID:0001866	4.589	[46]	*ii*
					Pyrrolocarbazoles	CHEMONTID:0001865	4.589		
					Isoindolones	CHEMONTID:0001820	4.137		
TfuA-like protein	PF07812	x	x		Organosulfur compounds	CHEMONTID:0000004	2.248	[47]	
					Imidothioic acids and derivatives	CHEMONTID:0002462	3.734		
Thiazolinyl imine reductase Irp3-like, C-terminal	PF21390	x			Azolines	CHEMONTID:0002490	3.894	[48]	
					Thiazolines	CHEMONTID:0000229	4.804		
Cucumopine synthase C-terminal helical bundle domain	PF18631	x			Indoles and derivatives	CHEMONTID:0000211	2.118	[49]	
					Pyridines and derivatives	CHEMONTID:0000089	4.935		

Protein of unknown function (DUF692)	PF05114	x			Imidothioic acids and derivatives	CHEMONTID:0002462	5.634	[50]	*vi*
YcaO cyclodehydratase, ATP-ad Mg2+-binding	PF02624	x			Oxazoles	CHEMONTID:0000083	3.643	[51]	
					Thiazoles	CHEMONTID:0000095	3.531		
					Azoles	CHEMONTID:0000436	3.000		
					Carboximidamides	CHEMONTID:0003152	2.271	[36]	
Phosphoenolpyruvate phosphomutase	PF13714				Organophosphorus compounds	CHEMONTID:0000400	6.264	[52]	
Beta-propeller repeat	PF06739	x			Phenols	CHEMONTID:0000134	2.592		*ix*
					1-hydroxy-4-unsubstituted benzenoids	CHEMONTID:0004647	3.356		
SchA/CurD like domain	PF04486	x			Naphthopyranones	CHEMONTID:0001641	2.243		
					Acetals	CHEMONTID:0001656	2.259		
Protein of unknown function (DUF2378)	PF09536	x			Pyrrolidines	CHEMONTID:0000218	3.340		
Domain of unknown function (DUF4267)	PF14087	x			Coumarins and derivatives	CHEMONTID:0000145	2.774		*i*

Table 1: Select subset of highly weighted protein domain / chemical class associations learnt by CHAMOIS. The InterPro, EC and GO columns show whether the Pfam domain is part of InterPro 98.0, and if applicable whether the InterPro entry has GO terms and EC number annotations, respectively. Literature references consistent with CHAMOIS-inferred domain-class associations are included. Corresponding insets in Fig. 3 are indicated in the last column.

Figures

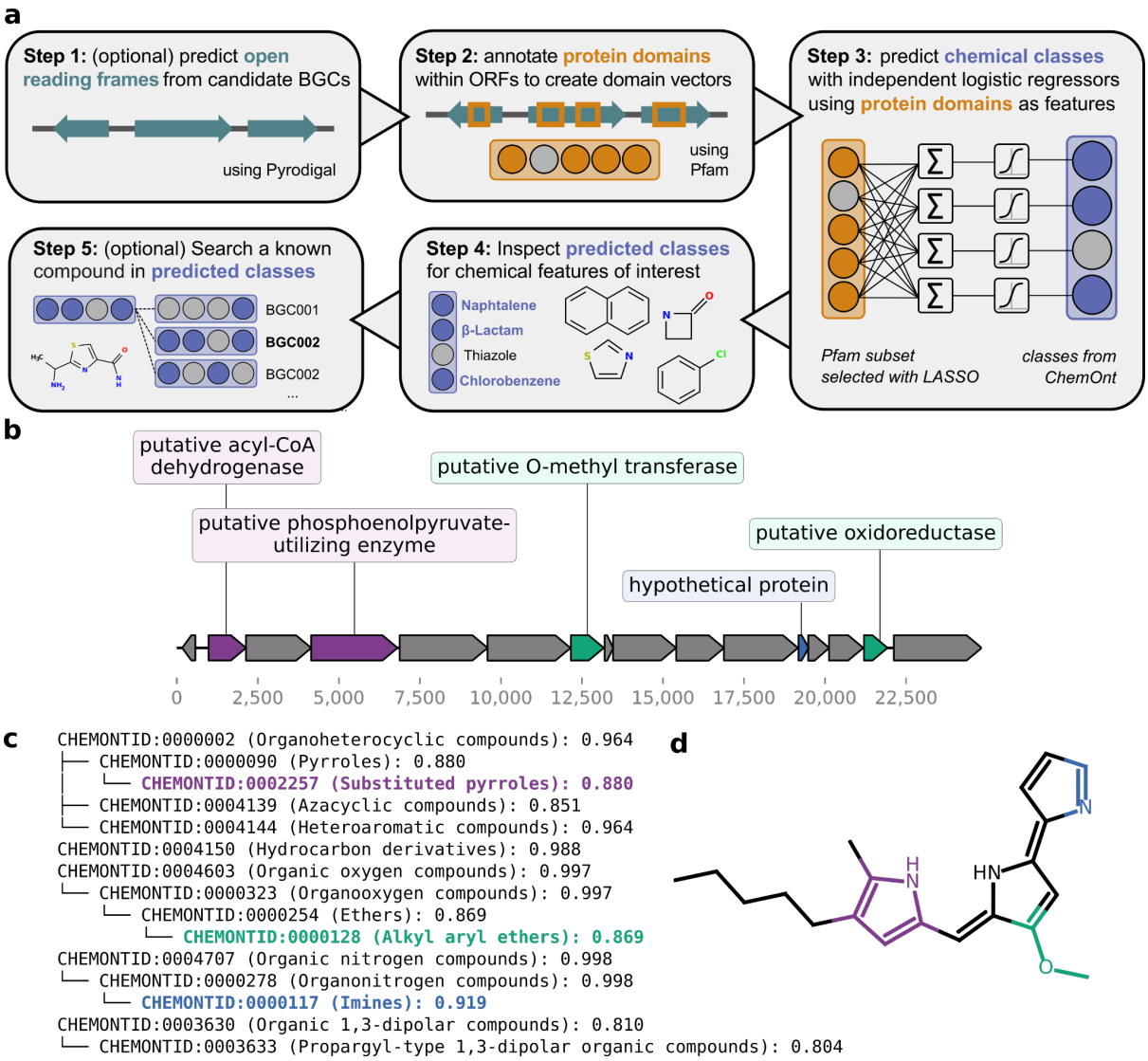

Figure 1: CHAMOIS, a novel method for predicting chemical properties of secondary metabolites from biosynthetic gene clusters (BGCs). **(a)** Workflow depiction of the chemical ontology prediction as implemented in CHAMOIS. First, CHAMOIS identifies the open reading frames in the input BGCs. Then, protein domains are annotated in the resulting ORFs using a subset of profile Hidden Markov Models from Pfam (N=1,617) 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=485). 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)[53]. 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[54] library. Genes containing domains that contributed to these class predictions are highlighted in the corresponding colour.

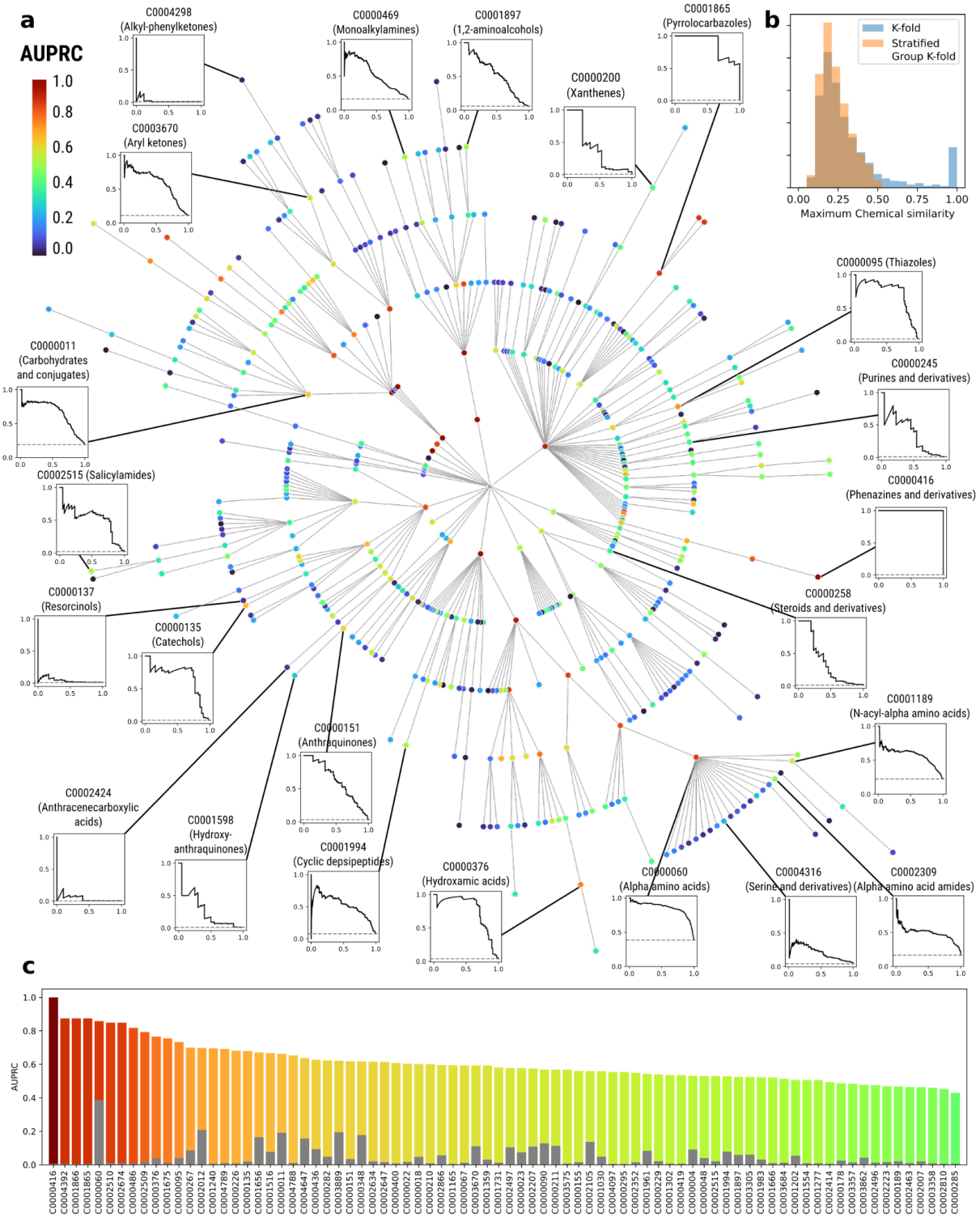

Figure 2: Evaluation of CHAMOIS against N=2,068 BGCs with known metabolites from MIBiG 3.1. **(a)** Model performance on every predicted ChemOnt class (N=485) using stratified 5-fold cross-validation. Classes are displayed as nodes within the ChemOnt hierarchy and coloured according to CHAMOIS's performance 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). **(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 for the 80 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 111 ChemOnt classes (Supplementary Table 1).

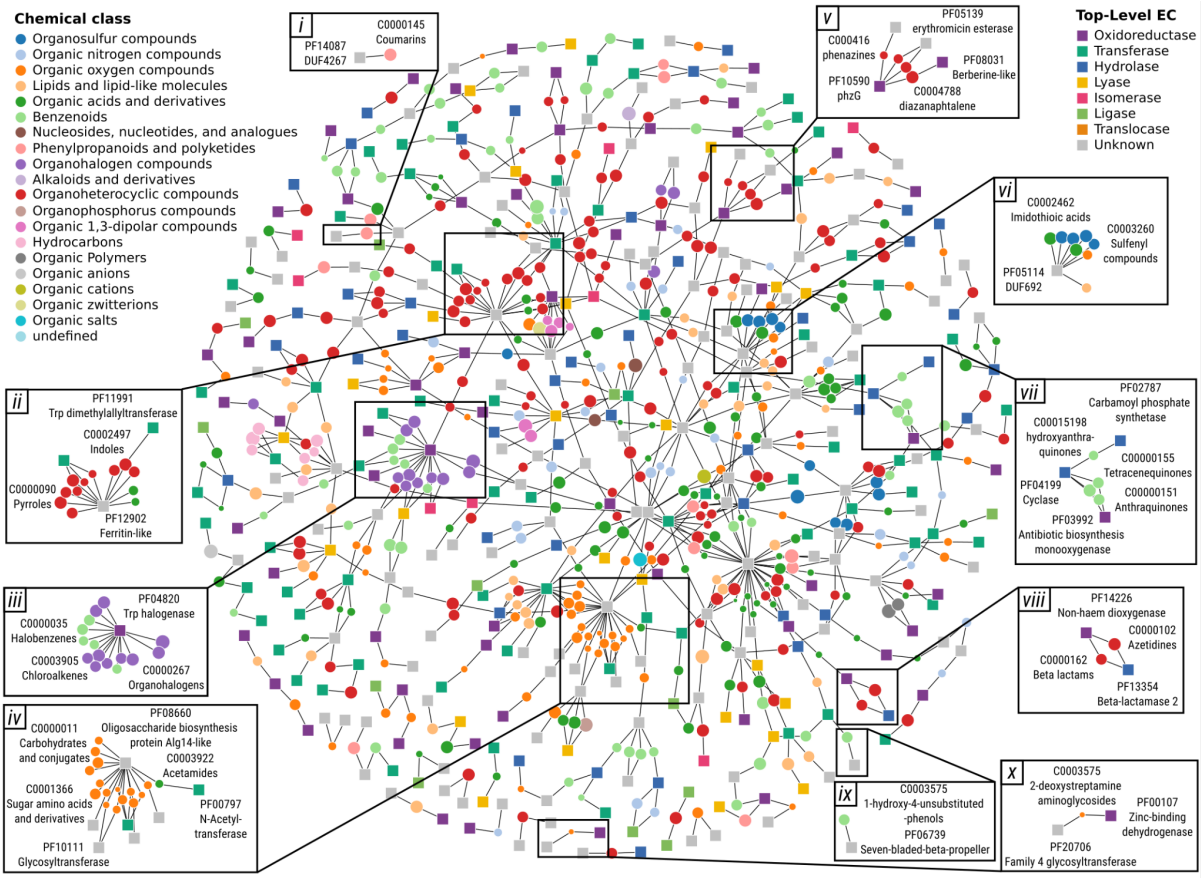

Figure 3: Network visualisation of associations between protein domains and chemical classes extracted from CHAMOIS's logistic regression model coefficients (Supplementary Table 2). Each class (circle) is linked to the two Pfam domains (square) with the highest weights. Classes are coloured according to their highest-level ancestor in the ChemOnt taxonomy. Pfam domains are coloured by their top-level EC number, obtained with the ECDomainMiner tool[55]. Several subgraphs feature associations supported by the literature (insets *ii, iii, iv, v, vi, vii, viii, x),* such as the involvement of DUF692 (PF05114) in the formation of imidothioic acid moieties in natural product compounds[50], but also associations suggesting novel functions of uncharacterized domains (insets *i* and *ix*, Table 1*)*.

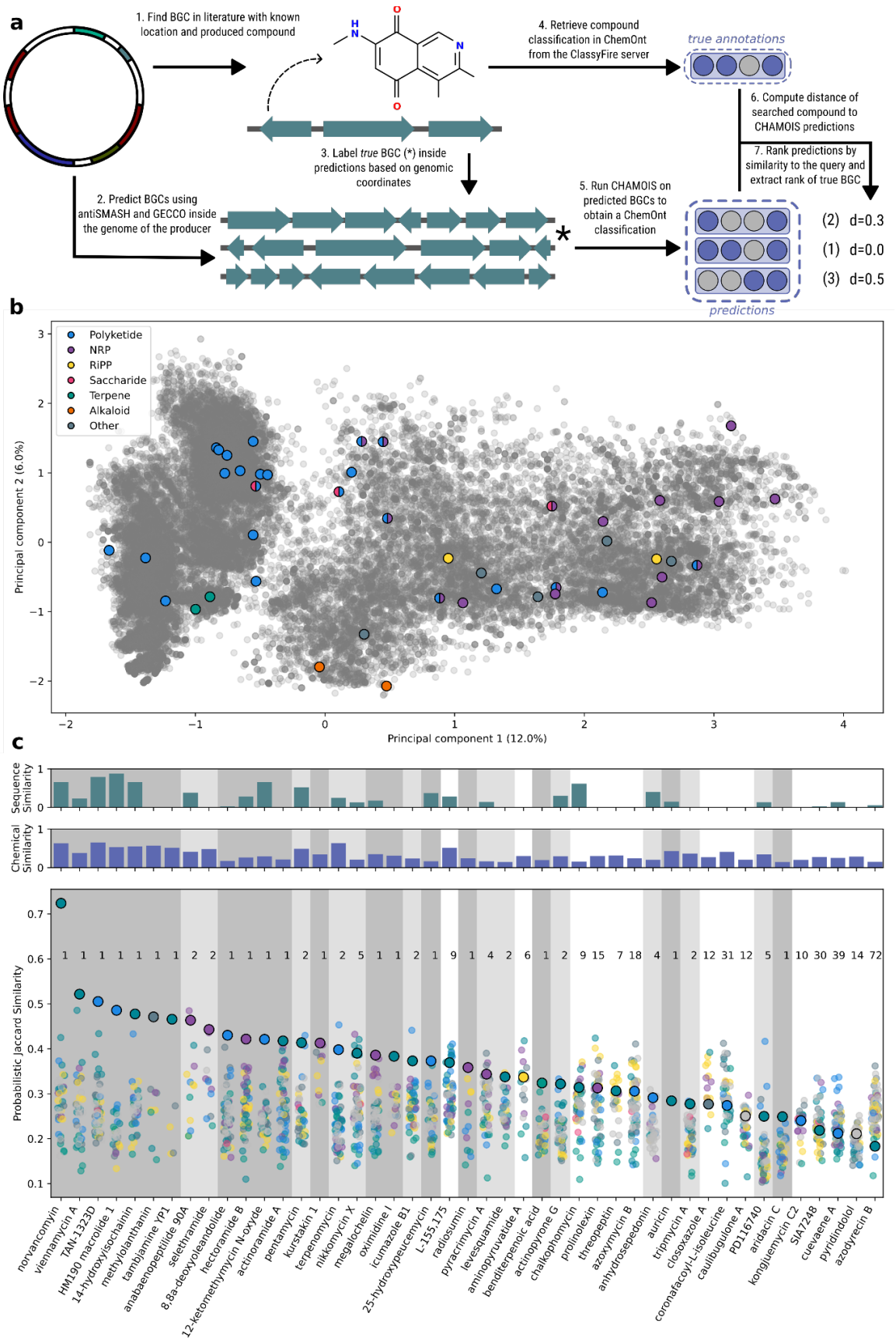

Figure 4: CHAMOIS can screen BGCs from within a given producer genome to pinpoint which one produces a given metabolite. **(a)** Graphical depiction of the prediction and evaluation framework. Shortly, as ground truth, experimentally validated BGCs with a known compound

and found in a complete genome were extracted from literature (N=45). The known compounds were annotated with ClassyFire to obtain their reference ChemOnt classification. The source genomes were passed through antiSMASH (v7.0) and GECCO (v0.9.10) to predict putative BGCs (see Methods). BGC predictions (N=1,795) were then given to CHAMOIS to predict their ChemOnt classification. For each true compound, all predictions for a given genome were ranked using Probabilistic Jaccard similarity[38] 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 Online Methods). NPAtlas compounds[56] 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=19), light-grey backgrounds highlight instances where the true compound ranked among the top five predictions. 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 Online 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 Online Methods).

Online Methods

Sequence dataset preparation. BGC sequences for MIBiG 3.1[7] were downloaded from the MIBiG website (https://mibig.secondarymetabolites.org) as GenBank files. 73 BGCs were filtered based on a manually curated blocklist (Supplementary Table 4). The cluster sequences were trimmed to the coordinates of the outermost genes of the clusters. The coordinates of 21 clusters were manually 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] v3.0 wrapping Prodigal[27] v2.6.3 in *meta* mode with closed ends (`-c -p meta`). The genes were then mapped against Pfam[29] v36.0 using `hmmsearch` as implemented in PyHMMER[30] v0.10.11 wrapping HMMER v3.4 (http://hmmer.org). 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=3,585) was constructed, which was stored in an HDF5 file using the `anndata` library[57] v0.10.5.

Compound dataset preparation. The BGC metadata for MIBiG 3.1[7] were downloaded from the MIBiG website (https://mibig.secondarymetabolites.org) as JSON files, containing the structures of the compounds as SMILES. 108 cluster compounds were manually 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[58], or, failing that, against PubChem[59] (https://pubchem.ncbi.nlm.nih.gov). All obtained SMILES strings were then converted into InChi using RDKit[60] 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, 2,068 could be successfully classified this way. The resulting labels were stored in an HDF5 file using the `anndata` library[57] v0.10.5. For further stratification, we grouped BGCs using a pairwise Hamming distance[33] cutoff of 0.5 between their MHFP6 fingerprints[25] computed using RDKit[60] v2023.9.6 (http://www.rdkit.org) resulting in 1,536 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 10 positive or negative occurrences were excluded from the label matrix. A LASSO logistic regression model was trained on each target label independently (N=485) with the `LogisticRegression` classifier implemented in `scikit-learn`[61] v1.3.0, using the `LIBLINEAR`[62] 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=1,617). 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 10 positive or negative occurrences 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`[61] v1.3.0 to split the data according to the molecular-similarity groups built from MHFP6 fingerprints[25]. 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.

Visualization of cross-validation performance across the ChemOnt hierarchy.. The `pronto` Python library (https://github.com/althonos/pronto) 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. Individual precision-recall curves were generated with `matplotlib`[63] v3.10.0 (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[55] 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[35] 98.0 in XML format (https://www.ebi.ac.uk/interpro/download/). We collected all Pfam[29] 36.0 domains, and retained those as "uncharacterized" which did not have a corresponding InterPro entry with an Enzyme Commission (EC) number or a Gene Ontology[64] term. We further analyzed domains that were assigned a weight of at least 2.0 by any classifier within CHAMOIS (Supplementary Tables 2 and 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 45 experimentally-validated BGCs not included in MIBiG 3.1[7] that were contained in 42 (near-)complete bacterial genomes. Each of these genomes was newly annotated with the antiSMASH[65] web server and GECCO[10]. For regions where both tools predicted a BGC, their intersection was kept. All predicted BGCs were then annotated with CHAMOIS to predict a ChemOnt classification. For each source genome, the distance between the true compound and all predictions was computed using Probabilistic Jaccard similarity[38] 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 for this genome.

Contextualization of benchmark data. The NPAtlas v2024_03 entries[66] were downloaded from the NPAtlas[58] 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`[61] v1.3.0 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`[63] 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[69] 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`[67] v0.2.2 (with parameters `-c10 -m50`) wrapped in Pyskani[68] v0.1.3. For each hit, the query ANI was computed, where $query\ ANI = identity * query\ fraction$ with $identity$ and $query\ fraction$ 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`[63] v3.10.0 (Fig. 4c).

Availability

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.

Data availability. All training and test data can be downloaded from https://zenodo.org/records/15009032. CHAMOIS weights are available in Supplementary Table 4.

References

1. Fischbach, M. A., Walsh, C. T. & Clardy, J. The evolution of gene collectives: How natural selection drives chemical innovation. *Proc. Natl. Acad. Sci.* **105**, 4601–4608 (2008).
2. Walsh, C. T. & Tang, Y. *Natural Product Biosynthesis: Chemical Logic and Enzymatic Machinery*. (Royal Society of Chemistry, London, 2023).
3. Crits-Christoph, A., Bhattacharya, N., Olm, M. R., Song, Y. S. & Banfield, J. F. Transporter genes in biosynthetic gene clusters predict metabolite characteristics and siderophore activity. *Genome Res.* **31**, 239–250 (2021).
4. Brady, S. F. Construction of soil environmental DNA cosmid libraries and screening for clones that produce biologically active small molecules. *Nat. Protoc.* **2**, 1297–1305 (2007).
5. Stohl, E. A., Milner, J. L. & Handelsman, J. Zwittermicin A biosynthetic cluster. *Gene* **237**, 403–411 (1999).
6. Blin, K. *et al.* antiSMASH 7.0: new and improved predictions for detection, regulation, chemical structures and visualisation. *Nucleic Acids Res.* **51**, W46–W50 (2023).
7. Terlouw, B. R. *et al.* MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.* **51**, D603–D610 (2023).
8. Cimermancic, P. *et al.* Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
9. Hannigan, G. D. *et al.* A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, e110–e110 (2019).
10. Carroll, L. M. *et al.* Accurate de novo identification of biosynthetic gene clusters with GECCO. *BioRxiv Prepr. Serv. Biol.* (2021) doi:10.1101/2021.05.03.442509.
11. Sanchez, S. *et al.* Expansion of novel biosynthetic gene clusters from diverse environments using SanntiS. 2023.05.23.540769 Preprint at https://doi.org/10.1101/2023.05.23.540769 (2023).
12. Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y. & Weber, T. The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. *Nucleic Acids Res.* **45**, D555–D559 (2017).
13. Fullam, A. *et al.* proGenomes3: approaching one million accurately and consistently annotated high-quality prokaryotic genomes. *Nucleic Acids Res.* **51**, D760–D766 (2023).
14. Liu, T. *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.* **10**, 278–286 (2009).
15. Huo, L. *et al.* Heterologous expression of bacterial natural product biosynthetic pathways. *Nat. Prod. Rep.* **36**, 1412–1436 (2019).
16. Nivina, A., Yuet, K. P., Hsu, J. & Khosla, C. Evolution and Diversity of Assembly-Line Polyketide Synthases: Focus Review. *Chem. Rev.* **119**, 12524–12547 (2019).
17. Marahiel, M. A., Stachelhaus, T. & Mootz, H. D. Modular Peptide Synthetases Involved in Nonribosomal Peptide Synthesis. *Chem. Rev.* **97**, 2651–2674 (1997).
18. Stachelhaus, T., Mootz, H. D. & Marahiel, M. A. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem. Biol.* **6**, 493–505 (1999).
19. Röttig, M. *et al.* NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. *Nucleic Acids Res.* **39**, W362–W367 (2011).
20. Kunyavskaya, O. *et al.* Nerpa: A Tool for Discovering Biosynthetic Gene Clusters of Bacterial Nonribosomal Peptides. *Metabolites* **11**, 693 (2021).

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.* **45**, W80–W88 (2017).
22. Helfrich, E. J. N. *et al.* Automated structure prediction of trans-acyltransferase polyketide synthase products. *Nat. Chem. Biol.* **15**, 813–821 (2019).
23. Skinnider, M. A. *et al.* Comprehensive prediction of secondary metabolite structure and biological activity from microbial genome sequences. *Nat. Commun.* **11**, 6058 (2020).
24. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminformatics* **8**, 61 (2016).
25. Probst, D. & Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *J. Cheminformatics* **10**, 66 (2018).
26. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* **58**, 267–288 (1996).
27. Hyatt, D. *et al.* Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* **11**, 119 (2010).
28. Larralde, M. Pyrodigal: Python bindings and interface to Prodigal, an efficient method for gene prediction in prokaryotes. *J. Open Source Softw.* **7**, 4296 (2022).
29. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
30. Larralde, M. & Zeller, G. PyHMMER: A Python library binding to HMMER for efficient sequence analysis. *Bioinformatics* btad214 (2023) doi:10.1093/bioinformatics/btad214.
31. Walker, A. S. & Clardy, J. A Machine Learning Bioinformatics Method to Predict Biological Activity from Biosynthetic Gene Clusters. *J. Chem. Inf. Model.* **61**, 2560–2571 (2021).
32. Zhang, M. *et al.* Comparative Investigation into Formycin A and Pyrazofurin A Biosynthesis Reveals Branch Pathways for the Construction of C-Nucleoside Scaffolds. *Appl. Environ. Microbiol.* **86**, e01971-19 (2020).
33. Hamming, R. W. Error detecting and error correcting codes. *Bell Syst. Tech. J.* **29**, 147–160 (1950).
34. Hendry, S. *et al.* Functional Analysis of Phenazine Biosynthesis Genes in Burkholderia spp. *Appl. Environ. Microbiol.* **87**, e02348-20 (2021).
35. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2020).
36. Burkhart, B. J., Schwalen, C. J., Mann, G., Naismith, J. H. & Mitchell, D. A. YcaO-Dependent Posttranslational Amide Activation: Biosynthesis, Structure, and Function. *Chem. Rev.* **117**, 5389–5456 (2017).
37. Wang, H.-Q. *et al.* Structural Basis for the Friedel-Crafts Alkylation in Cylindrocyclophane Biosynthesis. Preprint at https://doi.org/10.26434/chemrxiv-2021-xdbzq-v2 (2021).
38. Martire, I., da Silva, P. N., Plastino, A., Fabris, F. & Freitas, A. A. A novel probabilistic Jaccard distance measure for classification of sparse and uncertain data. in (eds. Rebeiro de Faria Paiva, E., Merschmann, L. & Cerri, R.) 81–88 (Uberlandia, MG, Brazil, 2017).
39. Segal, M. R., Dahlquist, K. D. & Conklin, B. R. Regression approaches for microarray data analysis. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **10**, 961–980 (2003).
40. Averbeck, N., Keppler-Ross, S. & Dean, N. Membrane topology of the Alg14 endoplasmic reticulum UDP-GlcNAc transferase subunit. *J. Biol. Chem.* **282**, 29081–29088 (2007).
41. Chen, Y.-H. & Cheng, W.-H. Hexosamine biosynthesis and related pathways, protein N-glycosylation and O-GlcNAcylation: their interconnection and role in plants. *Front. Plant*

Sci. **15**, 1349064 (2024).
42. Adak, S., Lukowski, A. L., Schäfer, R. J. B. & Moore, B. S. From Tryptophan to Toxin: Nature's Convergent Biosynthetic Strategy to Aetokthonotoxin. *J. Am. Chem. Soc.* **144**, 2861–2866 (2022).
43. Lingkon, K. & Bellizzi, J. J. Structure and Activity of the Thermophilic Tryptophan-6 Halogenase BorH. *Chembiochem Eur. J. Chem. Biol.* **21**, 1121–1128 (2020).
44. Juneja, P., Hubrich, F., Diederichs, K., Welte, W. & Andexer, J. N. Mechanistic Implications for the Chorismatase FkbO Based on the Crystal Structure. *J. Mol. Biol.* **426**, 105–115 (2014).
45. Barry, S. M. *et al.* Cytochrome P450–catalyzed L-tryptophan nitration in thaxtomin phytotoxin biosynthesis. *Nat. Chem. Biol.* **8**, 814–816 (2012).
46. Antônio, R. V. & Creczynski-Pasa, T. B. Genetic analysis of violacein biosynthesis by Chromobacterium violaceum. *Genet. Mol. Res. GMR* **3**, 85–91 (2004).
47. Liu, A. *et al.* Functional elucidation of TfuA in peptide backbone thioamidation. *Nat. Chem. Biol.* **17**, 585–592 (2021).
48. Meneely, K. M., Ronnebaum, T. A., Riley, A. P., Prisinzano, T. E. & Lamb, A. L. Holo Structure and Steady State Kinetics of the Thiazolinyl Imine Reductases for Siderophore Biosynthesis. *Biochemistry* **55**, 5423–5433 (2016).
49. Mori, T. *et al.* Structural Basis for β-Carboline Alkaloid Production by the Microbial Homodimeric Enzyme McbB. *Chem. Biol.* **22**, 898–906 (2015).
50. Ayikpoe, R. S., Zhu, L., Chen, J. Y., Ting, C. P. & van der Donk, W. A. Macrocyclization and Backbone Rearrangement During RiPP Biosynthesis by a SAM-Dependent Domain-of-Unknown-Function 692. *ACS Cent. Sci.* **9**, 1008–1018 (2023).
51. Dunbar, K. L. *et al.* Discovery of a new ATP-binding motif involved in peptidic azoline biosynthesis. *Nat. Chem. Biol.* **10**, 823–829 (2014).
52. 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* Tu494. *FEMS Microbiol. Lett.* **163**, 149–157 (1998).
53. Harris, A. K. P. *et al.* The Serratia gene cluster encoding biosynthesis of the red antibiotic, prodigiosin, shows species- and strain-dependent genome context variation. *Microbiol. Read. Engl.* **150**, 3547–3560 (2004).
54. Terlouw, B. R., Vromans, S. P. J. M. & Medema, M. H. PIKAChU: a Python-based informatics kit for analysing chemical units. *J. Cheminformatics* **14**, 34 (2022).
55. Alborzi, S. Z., Devignes, M.-D. & Ritchie, D. W. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. *BMC Bioinformatics* **18**, 107 (2017).
56. van Santen, J. A. *et al.* The Natural Products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Res.* **50**, D1317–D1323 (2022).
57. Virshup, I., Rybakov, S., Theis, F. J., Angerer, P. & Wolf, F. A. anndata: Access and store annotated data matrices. *J. Open Source Softw.* **9**, 4371 (2024).
58. van Santen, J. A. *et al.* The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Cent. Sci.* **5**, 1824–1833 (2019).
59. Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380 (2023).
60. Landrum, G. *et al.* rdkit/rdkit: 2024_09_6 (Q3 2024) Release. Zenodo https://doi.org/10.5281/zenodo.14943932 (2025).
61. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

62. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R. & Lin, C.-J. LIBLINEAR: A Library for Large Linear Classification. *J. Mach. Learn. Res.* **9**, 1871–1874 (2008).
63. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
64. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**, D330–D338 (2019).
65. Blin, K. *et al.* antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res.* **47**, W81–W87 (2019).
66. van Santen, J. A., Linington, R. G. & Contributors, N. A. The Natural Products Atlas - data download. Zenodo https://doi.org/10.5281/zenodo.13756408 (2024).
67. Shaw, J. & Yu, Y. W. Fast and robust metagenomic sequence comparison through sparse chaining with skani. *Nat. Methods* **20**, 1661–1665 (2023).
68. Larralde, M., Zeller, G. & Carroll, L. M. PyOrthoANI, PyFastANI, and Pyskani: a suite of Python libraries for computation of average nucleotide identity. 2025.02.13.638148 Preprint at https://doi.org/10.1101/2025.02.13.638148 (2025).
69. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

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.

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.

Ethics declarations

The authors declare no competing interests.