nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-020-0740-8

Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra

Kai Dührkop[1], Louis-Félix Nothias[2], Markus Fleischauer[1], Raphael Reher[3], Marcus Ludwig[1], Martin A. Hoffmann[1,4], Daniel Petras[2,5], William H. Gerwick[3,6], Juho Rousu[7], Pieter C. Dorrestein[2] and Sebastian Böcker[1]✉

Metabolomics using nontargeted tandem mass spectrometry can detect thousands of molecules in a biological sample. However, structural molecule annotation is limited to structures present in libraries or databases, restricting analysis and interpretation of experimental data. Here we describe CANOPUS (class assignment and ontology prediction using mass spectrometry), a computational tool for systematic compound class annotation. CANOPUS uses a deep neural network to predict 2,497 compound classes from fragmentation spectra, including all biologically relevant classes. CANOPUS explicitly targets compounds for which neither spectral nor structural reference data are available and predicts classes lacking tandem mass spectrometry training data. In evaluation using reference data, CANOPUS reached very high prediction performance (average accuracy of 99.7% in cross-validation) and outperformed four baseline methods. We demonstrate the broad utility of CANOPUS by investigating the effect of microbial colonization in the mouse digestive system, through analysis of the chemodiversity of different *Euphorbia* plants and regarding the discovery of a marine natural product, revealing biological insights at the compound class level.

Liquid chromatography mass spectrometry (LC–MS) can detect large numbers of small molecules from fractional amounts of samples, and has been widely adopted by the metabolomics community. LC–MS, when performed in an untargeted fashion, enables detection of hundreds to thousands of metabolites from a single analysis. However, the structural annotation of these metabolites remains highly challenging. Fragmentation spectra (tandem mass spectrometry (MS/MS)) collected in nontargeted mode can be annotated by matching against reference spectra in libraries[1–4] or structure databases using in silico tools[5–12]. However, only a fraction of metabolites can be annotated in this fashion[13]. Spectral libraries are limited in size and have been created largely with commercially available compounds[4,14–16]; even molecular structure databases, which can be orders of magnitude larger, may not cover all biomolecular structures from a particular sample[17,18]. Although libraries and structure databases are constantly growing, the general landscape is unlikely to change substantially over the next decade. As such, for most of the data acquired in metabolomics experiments, little structural insight can be obtained. In particular, it is currently not possible to obtain a comprehensive structural picture of which metabolites are present in a sample, and which are up- or down-regulated between two experimental conditions.

The problem of predicting the presence or absence of certain substructures has been considered since the 1960s, predominantly for gas chtomatography–MS data[19–21]. FingerID[22] and CSI:FingerID[7] predict molecular fingerprints that encode several hundreds or thousands of substructures in a molecule, respectively. Substructures describe local parts of the molecule, such as the presence of a hydroxy group. In contrast, compound classes are usually substantially more complex. Compound classes have been defined—for example, in the ChEBI ontology[23] or the MeSH thesaurus[24]—but class annotations are available for only a small fraction of molecular structures. In contrast, ClassyFire[25] enables deterministic assignment of classes solely from structure, facilitating classification of all molecular structures. ClassyFire definitions involve the use of logical expressions, substructures of variable length and substructure count constraints.

Previous studies on compound class regulation usually require that metabolites are first structurally annotated[26–29], but class assignment remains challenging. First, prediction of classes for which limited or no MS/MS data are available in spectral libraries is difficult; libraries are notoriously incomplete and most classes are sparse. Second, even if sufficient MS/MS data appear to be available for certain classes, it is possible that compounds for which we have reference spectra are not distributed evenly (or not at all) among subclasses. As an extreme case, assume that our training data contain reference spectra for pregnane steroids but no other steroids. Training a model for steroids using such data will predict only the subclass pregnane steroids, whereas all compounds from other subclasses (for example, ergostane steroids) will be misclassified as 'not a steroid' with high confidence. Third, prediction of classes for compounds of unknown molecular structure is particularly challenging and difficult to evaluate.

At present, three strategies for structural classification exist: (1) cluster compounds based on spectral similarity, then propagate compound class annotations from the database search in a

[1]Chair for Bioinformatics, Friedrich-Schiller University, Jena, Germany. [2]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, USA. [3]Center for Marine Biotechnology and Biomedicine, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA. [4]International Max Planck Research School 'Exploration of Ecological Interactions with Molecular and Chemical Techniques', Max Planck Institute for Chemical Ecology, Jena, Germany. [5]Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA. [6]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, USA. [7]Helsinki Institute for Information Technology, Department of Computer Science, Aalto University, Espoo, Finland. ✉e-mail: sebastian.boecker@uni-jena.de

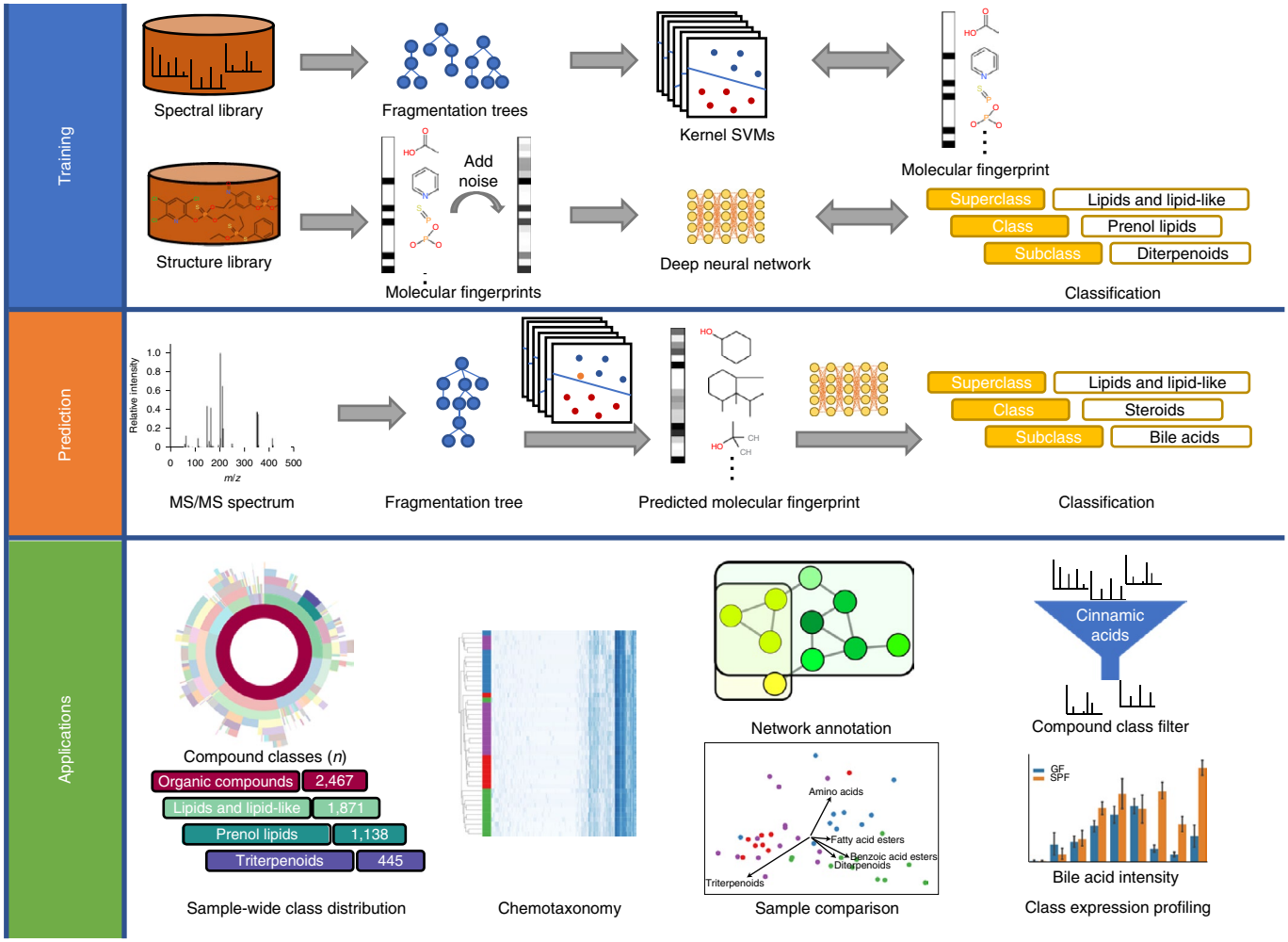


Fig. 1 | CANOPUS workflow. In the training phase, a battery of kernel SVMs is trained for prediction of molecular fingerprints from fragmentation spectra, and a DNN is trained to predict compound classes from molecular fingerprints (multilabel classification). In the prediction phase, we classify the query compound from its MS/MS spectrum by computing the fragmentation tree, predicting its molecular fingerprint and predicting compound classes from the fingerprint using the DNN. Numerous applications of compound classification exist, some of which are highlighted throughout this paper. These include compound class annotation for individual compounds as part of structural elucidation; visualizing and profiling distributions of compound classes in or between samples; clustering samples based on their compound class distribution; annotating molecular networks with compound classes; using principal component analysis to visualize compound class distributions; and filtering compounds based on compound classes of interest.

semiautomated manner[30–32]; (2) search for the query compound in a spectral library[33,34] or a structure database[27,28] and consider the top *k* hits for assignment of compound classes; or (3) use machine learning methods for direct prediction of compound classes from the MS/MS spectrum[28,35].

Here we present CANOPUS, a computational method that addresses these problems by assigning compound classes to every metabolite MS/MS feature in a LC–MS/MS run, including metabolites with structures not recorded in any database or publication.

Results

CANOPUS evaluation. The workflow of CANOPUS is depicted in Fig. 1. Given an MS/MS spectrum as input, we use a battery of support vector machines (SVMs) to predict a probabilistic fingerprint of the query compound[7,22]. This probabilistic fingerprint is used as input of a deep neural network (DNN)[36], which then predicts all compound classes simultaneously. SVMs are trained using reference MS/MS spectra, whereas the DNN is trained on 4.10 million compound structures and does not require any MS/MS data. To train the DNN we simulate a 'realistic' probabilistic fingerprint for any given molecular structure, although no MS/MS data for this structure are available. This integration of two machine learning techniques allows CANOPUS to reach high-quality predictions for 2,497 ClassyFire compound classes; this includes all biologically relevant classes (subclasses of organic compounds with 400 or more examples in PubChem); see Supplementary Table 1 for the remaining classes not predicted by CANOPUS. Because predictions are now independent from the availability of MS/MS reference data, CANOPUS can predict compound classes even when there are no MS/MS spectra available for training the method. CANOPUS can also predict classes for which MS/MS training data are missing for a complete subclass. CANOPUS predicts all 2,497 classes for every query MS/MS, facilitating not only a global overview of the compound classes measured in a biological sample but also of the differences between cohorts at the compound class level. CANOPUS does not require the user to choose compound classes of interest or to retrain it for individual datasets.

To evaluate CANOPUS we used reference MS/MS libraries, because we need to know the true answer for evaluation. The kernel SVMs used for prediction of molecular fingerprint were trained

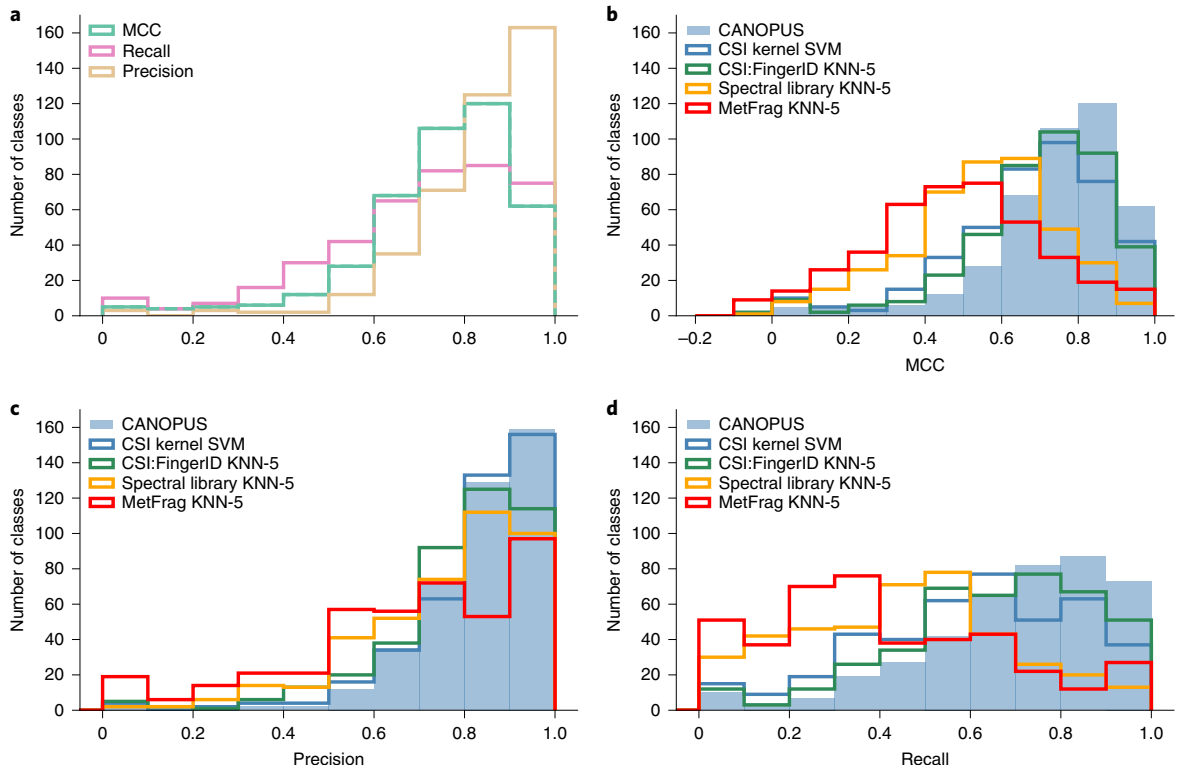


Fig. 2 | Method evaluation: number of ClassyFire compound classes predicted with a particular performance measure. a, Histogram of MCC, precision and recall of CANOPUS predictions for individual compound classes; SVM training dataset, compound classes with at least 20 positive examples. See Supplementary Table 2 for MCC, precision and recall results of individual compound classes. **b**–**d**, Histograms of MCC (**b**), precision (**c**) and recall (**d**) for all methods; independent dataset, compound classes with at least 20 positive examples. Note that MCC = 0 corresponds to random predictions and MCC = 1 to error-free predictions for a compound class. CSI kernel SVM is the direct prediction from MS/MS spectra using a kernel SVM. MetFrag KNN-5 and CSI:FingerID KNN-5 search in PubChem using MetFrag or CSI:FingerID, respectively; spectral library KNN-5 searches in the SVM training dataset using cosine similarity. All KNN-5 methods use the majority vote of the top-five search results for each compound class.

on positive ion mode MS/MS spectra from 24,539 compounds. We used structure-disjoint tenfold cross-validation on the training data, and structure-disjoint evaluation on an independent MS/MS dataset of 3,387 compounds measured in positive ion mode. The DNN was trained on 4.10 million structures with compound classes assigned by ClassyFire[25]. All structures for which MS/MS spectra are available were removed from the DNN training data. These steps ensure that all compounds are truly novel in evaluation, meaning that both the MS/MS and the structure are unknown to the method.

For compounds from the SVM training dataset the average accuracy of CANOPUS predictions was 99.7%, and 2,313 of 2,497 classes were predicted with accuracy of at least 99%. However, because compound classes are often sparse, the Matthews correlation coefficient (MCC) is a more suitable performance measure[37]; see Extended Data Fig. 1 for the MCC of all 782 compound classes with at least 50 examples. CANOPUS predicted 607 compound classes with MCC ≥ 0.8 (Fig. 2a and Supplementary Table 2), including phosphocholines (MCC = 0.972), flavonoid *O*-glycosides (MCC = 0.922) and pregnane steroids (MCC = 0.875). For 453 classes we do not have any example spectrum in our dataset; 1,405 of the remaining 2,044 classes can be predicted with MCC ≥ 0.5. CANOPUS is also applicable to negative ion mode spectra; evaluation results are slightly inferior (average MCC = 0.713) (Supplementary Note 1 and Supplementary Table 2).

For the independent dataset, CANOPUS reached an average prediction accuracy of 99.7% and 2,345 classes were predicted with accuracy of at least 99% (Supplementary Table 3). The independent dataset is substantially smaller than the SVM training dataset, and MCC estimation can be inaccurate and misleading if too few positive examples are available. We therefore distinguished between rich classes with at least 20 positive examples (416 classes) and sparse classes with fewer examples (2,081 classes, 1,156 with no example). For rich classes, average MCC was 0.744, precision was 0.837 and recall was 0.702 (Fig. 2). For sparse classes we computed the micro-averaged MCC, adding together true positives, true negatives, false positives and false negatives over all sparse classes. The microaveraged MCC for all sparse classes was 0.603.

We evaluated CANOPUS against four 'baseline methods' (Fig. 2b–d; Methods). All of these methods are nontrivial and usually hard to better using more advanced classifiers. Two baseline methods (MetFrag KNN-5 and CSI:FingerID KNN-5) search in a structure database using either MetFrag[9,38] or CSI:FingerID[7,39], then perform a top-five majority vote[40]. The third method involves searching in a spectral library with no precursor mass filtering (spectral library KNN-5). For the last of these four, we evaluated direct class prediction from MS/MS spectra using the kernel support vector machine setup of CSI:FingerID. We again ensured structure-disjoint evaluation. We cannot evaluate against propagation of class annotations through clustering or molecular networking[41–43], because such propagation is semiautomated at best and still requires manual interpretation.

CANOPUS substantially outperformed all baseline methods, as we demonstrate for the independent dataset (Fig. 2b–d and Supplementary Table 3): the average MCC for rich classes was 0.692 for CSI:FingerID KNN-5, 0.542 for spectral library KNN-5, 0.482 for MetFrag KNN-5 and 0.679 for direct prediction from the

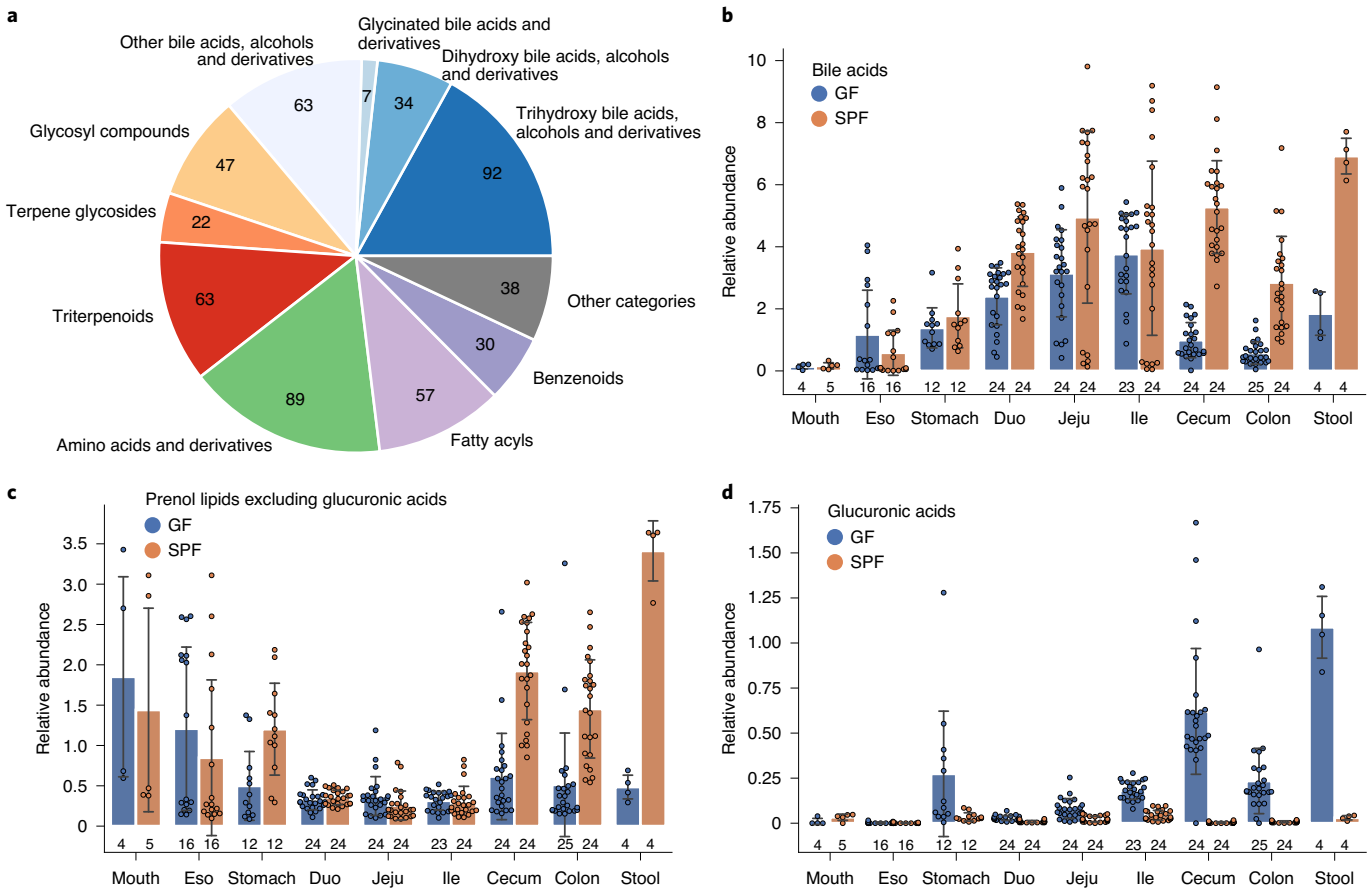


Fig. 3 | Comparing the digestive system of GF and SPF mice. a, Compound classes of all 542 compounds with fold change > 10 between the digestive systems of GF and SPF mice. **b**–**d**, Summed intensity of all compounds belonging to the classes bile acids (**b**), prenol lipids excluding glucuronic acids (**c**) and glucuronic acids (**d**) in the digestive system of GF (blue) and SPF (orange). Bars show averages (mean), s.d. shown as error bars and sample sizes shown below bars. Eso, esophagus; duo, duodenum; jeju, jejunum; ile, ileum.

MS/MS spectrum (0.744 for CANOPUS). The microaveraged MCC for sparse classes was 0.519 for CSI:FingerID KNN-5, 0.347 for spectral library KNN-5, 0.341 for MetFrag KNN-5 and 0.506 for direct prediction (0.603 for CANOPUS). Of note, the two baseline methods using CSI:FingerID performed substantially better than the other two.

Prediction of compound classes without MS/MS training data. CANOPUS can predict compound classes for which no training MS/MS spectra exist. To demonstrate this, we retrained the SVM battery such that all 491 MS/MS spectra of flavonoid glycosides were removed (Extended Data Fig. 2 and Supplementary Table 2). CANOPUS was still able to predict this compound class with MCC = 0.662, compared to 0.922 when flavonoid glycosides were included in the MS/MS training data. Similarly, the MCC for the parent class flavonoid was 0.782 when flavonoid glycosides were absent from training data compared to 0.885 otherwise. In both cases the drop in performance is surprisingly small, and the classifiers lacking training MS/MS spectra still gave an excellent performance.

By concept, direct prediction is not able to predict compound classes without any MS/MS training spectra. Beyond that, the lack of training spectra for one particular subclass can also substantially affect the performance of the parent class predictor. To demonstrate this, we trained a kernel SVM for prediction of flavonoids from MS/MS data but, again, omitted all 491 spectra of flavonoid glycosides. We then evaluated the flavonoid predictor on flavonoid glycoside MS/MS spectra: only 8% of the compounds were correctly classified as flavonoids. For comparison, CANOPUS correctly recognized 51% of flavonoid glycosides as flavonoids although the CANOPUS MS/MS training data did not contain a single flavonoid glycoside spectrum (Extended Data Fig. 2 and Supplementary Table 2).

We repeated the above analysis on a second compound class: bile acids, alcohols and derivatives. We trained CANOPUS and an SVM classifier without the MS/MS data of 904 bile acids and found that CANOPUS still gave a very high prediction performance for bile acids (MCC = 0.764, compared to 0.942 when bile acids were included in the MS/MS training data). The MCC of the superclasses remained almost constant (0.883 and 0.919, respectively). CANOPUS correctly predicted 83% of the omitted bile acid spectra as steroids; in comparison, direct prediction recognized only 51% of those spectra as steroids.

CANOPUS and metabolomics data analysis. Metabolomics aims to establish changes in the metabolite profile among different experimental conditions, time points and so on. These changes are usually monitored at a 'per feature' level but doing so cannot uncover complex changes in metabolite profiles, just as a complex trait cannot usually be attributed to a single genetic variant. We now demonstrate how monitoring of differences on a compound class level allows for a comprehensive view of the biological system with no previous knowledge. For that, we reanalyzed the data from Quinn et al.[44] where tissue samples from different organs of germ-free (GF)

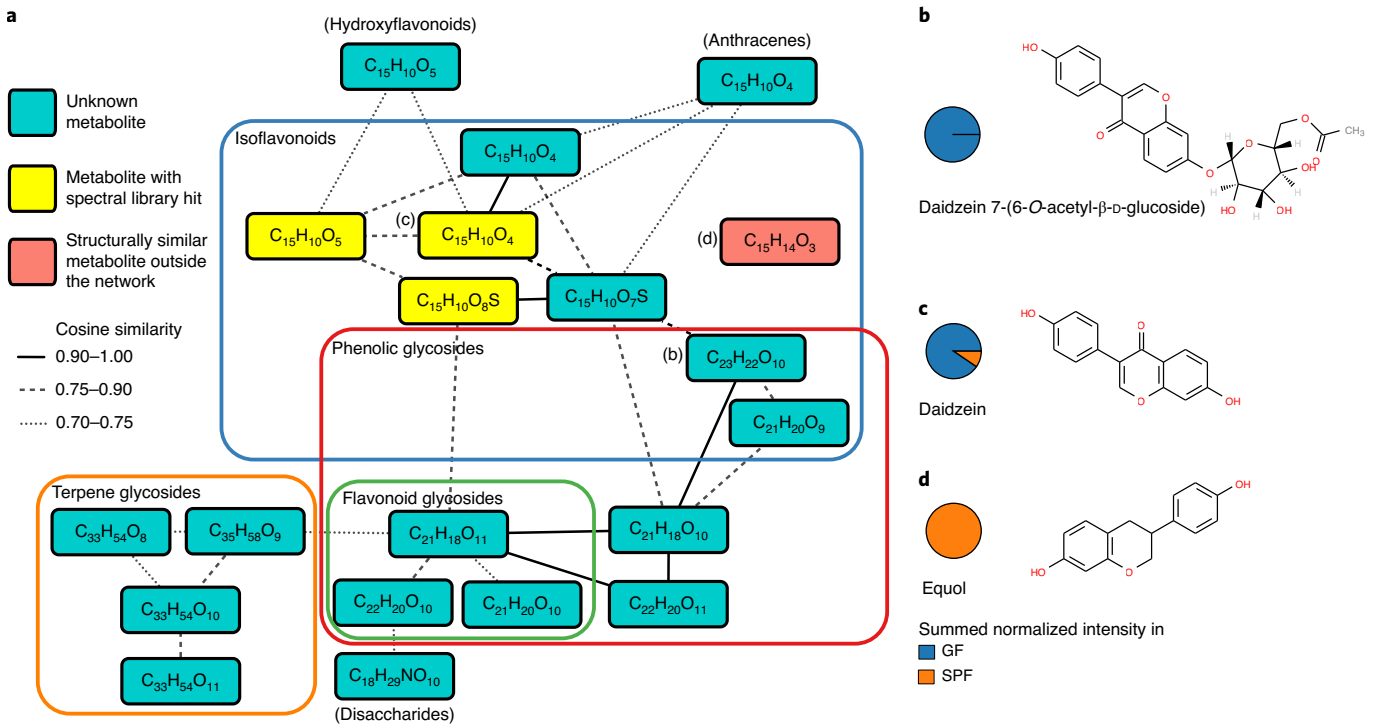


Fig. 4 | Molecular network of daidzein. a, Spectral library hits in the network are colored yellow, and spectra with no spectral library annotation are colored turquoise. Nodes are labeled with the SIRIUS molecular formula annotation. Solid borders are edges in the molecular network with high cosine similarity (0.9–1.0) while dashed and dotted lines are edges with low cosine similarity (0.75–0.9 and 0.7–0.75, respectively). CANOPUS compound classifications are indicated as boundaries in the network or written in parentheses next to the node. The compound equol (red) is not part of the network but was classified correctly as isoflavonoid by CANOPUS. **b**–**d**, Three compounds were annotated by CSI:FingerID as daidzein 7-(6-*O*-acetyl-β-D-glucoside) (**b**), daidzein (**c**) and equol (**d**). Pie charts show the relative intensity of the compounds in GF samples (blue) and SPF samples (orange) for large intestine and stool.

and specific-pathogen-free (SPF) mice were measured by nontargeted LC–MS/MS.

We sorted metabolites by fold change in intensity between GF and SPF samples. MS/MS spectra of each metabolite were classified using CANOPUS; 91.22% of the compounds were annotated at class level 5 and 60.00% at class level 6 (Extended Data Fig. 3). Next, classes were statistically tested for over-representation (one-tailed Mann–Whitney *U*-test, $n = 4{,}109$). The most significant compound classis 'bile acids, alcohols and derivatives' ($P = 6.58 \times 10^{-15}$), and also its subclasses and parent classes. Other highly significant classes include triterpenoids ($P = 2.85 \times 10^{-74}$) and glucuronic acid derivatives ($P = 3.04 \times 10^{-24}$) (Supplementary Table 4). See Fig. 3a for the classes of all 542 compounds with an intensity fold change >10 between GF and SPF mice.

We found that the abundance of bile acids is similar for GF and SPF in the small intestine but substantially different in the large intestine (cecum, colon) and in stool (Fig. 3b), consistent with the findings of Quinn et al.[44]. Our results also showed that glucuronic acids (including related saccharides) were among the most discriminating compounds; most of these were also classified as prenol lipids and isoflavonoids. In SPF, non-glycosylated prenol lipids appeared to be increasing through the digestive system from stomach to stool (Fig. 3c). In contrast, the abundance of prenol lipids in GF did not change notably through the digestive system. For the glucuronic acid derivatives class we observed an opposite trend: these had a relatively lower abundance in SPF and did not show a noticeable trend, but accumulated through the digestive system of GF with the highest abundance in stool (Fig. 3d). These results suggest the involvement of microbiota in the metabolization of glycosylated prenol lipids by cleaving off sugar acids. To verify this hypothesis, we considered two glycosylated compounds detected in GF but undetectable in SPF: we searched for deglycosylated derivatives in both GF and SPF. Using CSI:FingerID, we annotated the first of these compounds as a glycosylated derivative of the isoflavone genistein; this glycosylated derivative was detected only in the GF samples whereas genistein was detected in both SPF and GF samples. The second compound was annotated as glycosylated daidzein. We found that both daidzein and its glycosylated derivative, daidzein 7-(6-*O*-acetyl-β-D-glucoside), are abundant in GF but undetectable or low abundant in SPF (Fig. 4b–d), which appears to contradict the above hypothesis. However, daidzein is known to be metabolized into equol[45] and, indeed, equol was detected only in the SPF samples (Fig. 4d). Unfortunately, such in-depth verification is not possible in general because, for many compounds, the molecular structure cannot be confidently annotated.

Molecular networks are a popular method for pushing the boundaries of database search, by propagating annotations via spectral network similarity[46]. The underlying idea is that spectral similarity often implies structural similarity, so that annotations from a spectral library search can be propagated through connected subnetworks[42]. For the mouse samples we found 344 molecular subnetworks with at least three compounds, but only 92 of these subnetworks had at least one spectral library match enabling subnetwork annotation; an additional 376 compounds were singletons and did not cluster with any other compound in the samples. Propagation of compound classes to the full subnetwork, as proposed in ref. [32], functions only when there is an annotation within the connected subnetwork; furthermore, it may result in partial or imprecise classification of some spectra because it classifies based on consensus of subnetwork annotations derived from database-dependent annotation

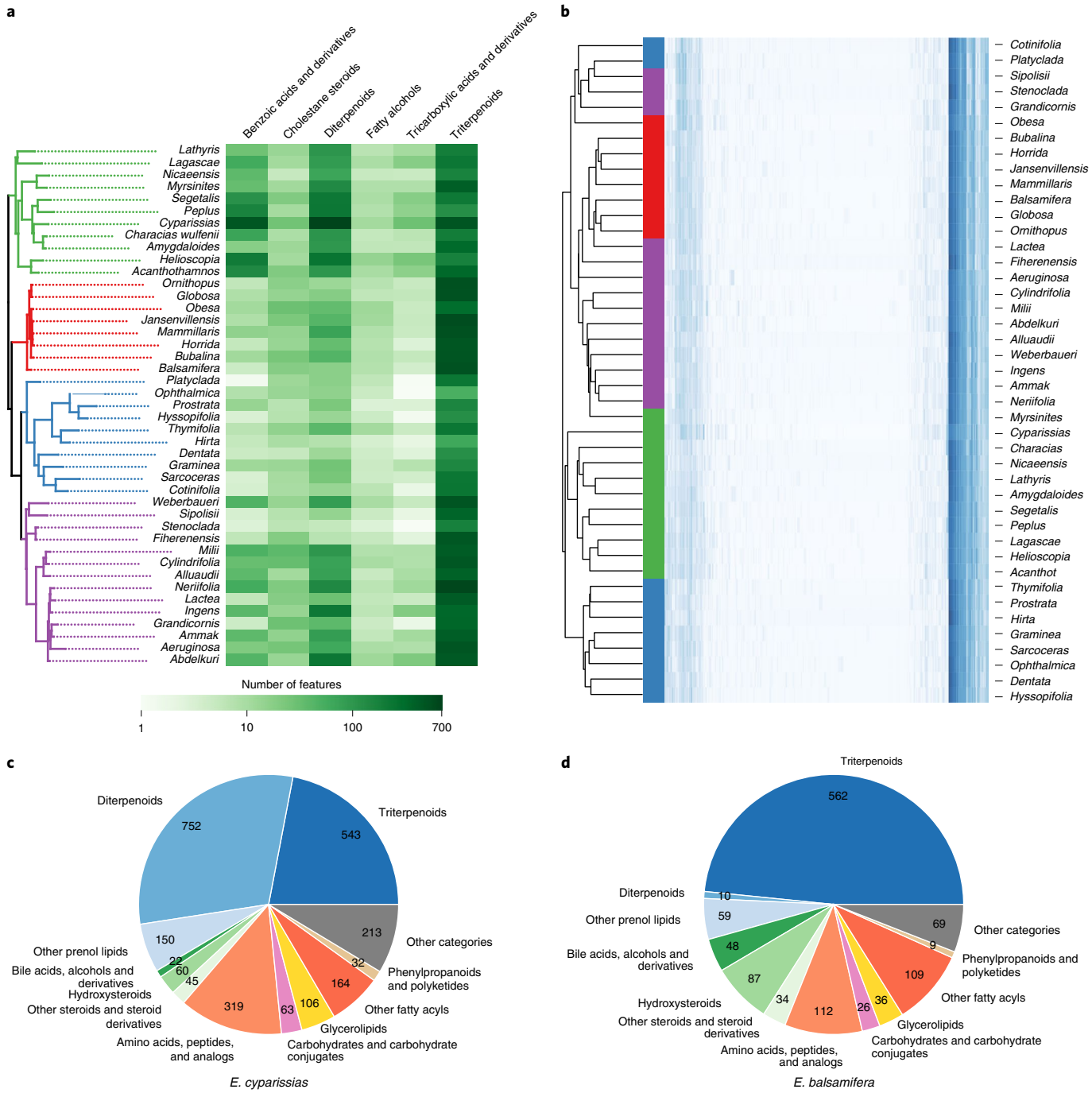


Fig. 5 | Compound class distribution in *Euphorbia* species. a, Heatmap of the number of compounds for six compound classes in *Euphorbia* species, grouped in a phylogenetic tree computed from genomic data. **b**, Clustered heatmap with dendrogram showing hierarchical clustering on compound class distributions. **a**,**b**, The color code on the left indicates the clade to which a row belongs: *Chamaesyce* (blue), *Euphorbia* (violet), *Athymalus* (red) and *Esula* (green). **c**,**d**, Distribution of compound classes in *E. cyparissias* (**c**) and *E. balsamifera* (**d**). Numbers within the pie charts are the absolute number of compounds of the class annotated in this species.

tools. CANOPUS, on the other hand, classifies each spectrum irrespectively of whether there is an annotation in the subnetwork. As an example, see the molecular subnetwork containing the compound daidzein shown in Fig. 4a. A spectral library search allowed us to annotate structures for four nodes of the molecular network, all isoflavonoids. However, inferring that all other compounds in this subnetwork are also isoflavonoids is most probably incorrect: CANOPUS annotated flavonoids and terpene glycosides in this subnetwork. Furthermore, most of the compounds in the network were annotated as glycosylated compounds (either phenolic glycosides or terpene glycosides) by CANOPUS, whereas none of the spectral annotations belong to these classes. While daidzein 7-(6-*O*-acetyl-β-D-glucoside) and daidzein are part of the same network, equol, their metabolic product, is a singleton and is not contained in the network because the MS/MS spectrum itself is quite different and therefore does not align (Fig. 4a). CANOPUS correctly annotated two additional isoflavonoids that form singleton subnetworks and, hence, are missing from the daidzein network;

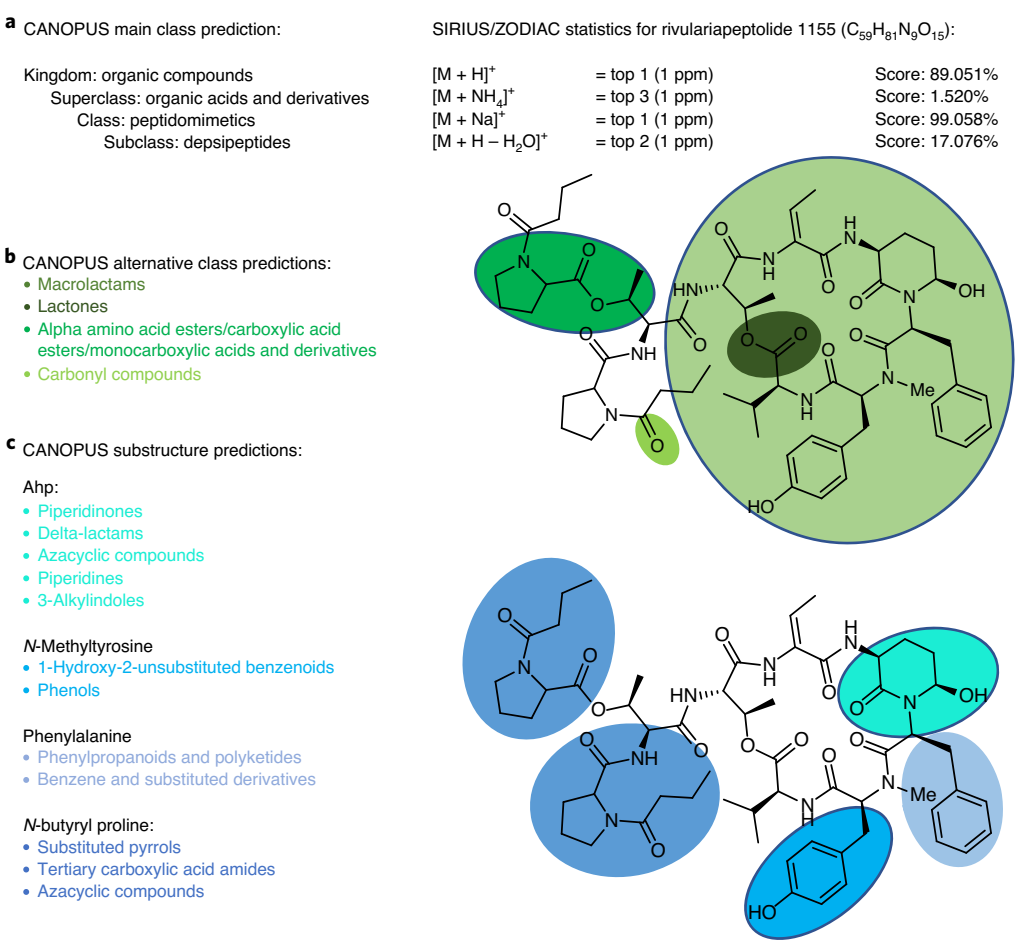


Fig. 6 | Structural analysis of rivulariapeptolide 1155 using CANOPUS. a, CANOPUS main class predictions based on ClassyFire ontology (left) and SIRIUS analysis enhanced with ZODIAC (right) to determine the correct molecular formula of rivulariapeptolide 1155. **b**, CANOPUS alternative class predictions confirming the macrocyclic depsipeptide structure of rivulariapeptolide 1155. **c**, CANOPUS substructure predictions (posterior probability >50%) that facilitated and accelerated the structural elucidation of rivulariapeptolide 1155.

these isoflavonoids were annotated by CSI:FingerID as acacetin and formononetin, which are structurally similar to daidzein. On a larger scale, we find that CANOPUS compound class annotations and molecular networks usually agree well (see the Cytoscape[47] visualization of the network in Extended Data Figs. 4 and 5).

Comparative metabolomics of *Euphorbia* species. We used CANOPUS to study chemical diversity among representative plant species of *Euphorbia* subgenera[27]. This time we were not considering the fold change of compounds or compound classes, but rather the change in metabolite diversity: how many compounds of a certain class do we find in each sample? To answer this question, it is crucial to annotate the entire set of molecules detected, as done by CANOPUS. Annotating only a small subset by, say, spectral library search will, at best, limit the findings and, at worst, produce misleading results: neither is the annotated subset an unbiased subsample of the complete metabolome, nor is the spectral library unbiased for compound classes. In a previous study[27], around 30% of the detected compounds were classified using the consensus of annotations from a spectral library search with GNPS[2], in silico structure annotation with network propagation[42], CSI:FingerID and spectral motifs from MS2LDA[48] that were manually annotated. The study showed that the diversity of *Euphorbia* diterpenoids, a type of bioactive compound studied for their antiviral and drug-resistance reversal properties[49], was larger in the subgenera *Esula* and *Euphorbia* than in *Chamaesyce* and *Athymalus*. This change in diversity agrees with the geographic colocation of these plants with certain herbivores, given that *Euphorbia* diterpenoids are a known feeding deterrent.

Reanalysis of this dataset with CANOPUS allowed us to reproduce the above biological findings, but also to derive new findings. Different from the method used previously[27], compounds were not aligned across *Euphorbia* species LC–MS/MS data but each species data were annotated individually. Using CANOPUS, we assigned compound classes to each compound; 94.27% of compounds were annotated at class level 5 and 43.04% at class level 6 (Extended Data Fig. 3). For each species, we counted the number of compounds belonging to each class but ignored compound intensities. First, we considered six compound classes investigated manually in a previous study[27] (Fig. 5a). For each class, our automated workflow detected and annotated substantially more compounds than the original study. This is due to the preprocessing method used in the original study, which led to fewer compounds being detected (Extended Data Fig. 6), but also to the unique ability of CANOPUS to systematically annotate fragmentation spectra with compound classes. For diterpenoids and triterpenoids we observed a similar class distribution pattern: the subgenera *Esula* and *Euphorbia* have a larger diversity of diterpenoids than the other two subgenera

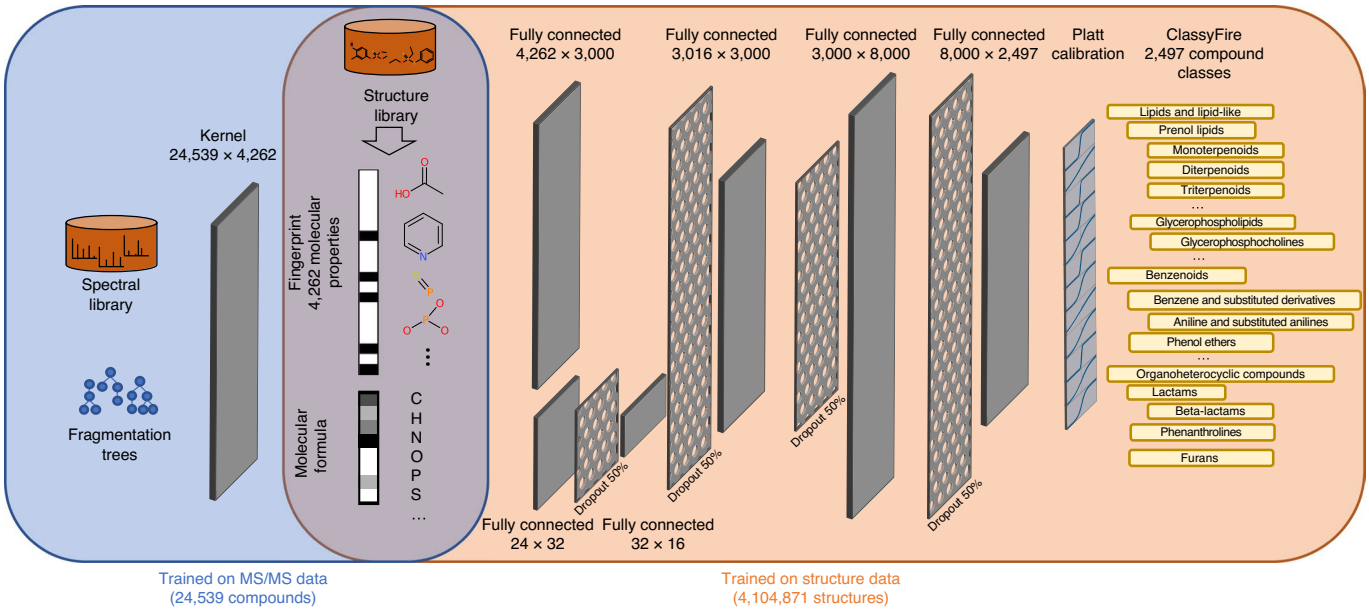


Fig. 7 | Heterogeneous training for compound class prediction. First, a battery of kernel SVMs is trained on spectral training data (blue area, left) to predict the molecular fingerprint. Second, a DNN predicts compound classes from the molecular fingerprint (orange area, right) and molecular formula. The network consists of several fully connected layers with ReLu activation function, and dropout layers between. Platt calibration is used to transform the network output to posterior probabilities. The key feature is that the molecular fingerprints used for training the DNN are computed directly from the structure database, without the need for spectral training data. Numbers of compounds shown are for positive ion mode.

(Fig. 5a and Extended Data Fig. 7), while species from the subgenera *Euphorbia* and *Athymalus* have a slightly higher diversity of triterpenoids (Extended Data Fig. 8). Cholestane steroids did not show a notable distribution pattern here, or in a previous study[27].

For benzoic acid esters we observed a notable distribution pattern: these show large and medium prevalence for subgenera *Esula* and *Euphorbia*, respectively, and no notable prevalence for the other two. This is in contrast to previous findings[27] where one to three benzoic acids were found across all species, corresponding to an absence of specific distribution. Many compounds annotated as diterpenoids were also annotated as benzoic acid esters, fatty acid esters or dicarboxylic acids (Extended Data Fig. 9). This is characteristic of *Euphorbia* 'lower' diterpenes that are often esterified with various acyls[49]. The subgenus *Esula* shows the highest diversity of benzoic acid esters, followed by *Euphorbia*. For the species *E. cyparissias*, 462 of 745 compounds annotated as diterpenoids were also annotated as benzoic acid esters. This agrees well with observations by Yang et al.[50] on *E. esula*, a species very closely related to *E. cyparissias*[51]. Although diterpenoids in *Euphorbia* species are frequently observed with benzoyloxy substituents[49], it was not known that the occurrence of this acylation differs to this extent for different subgenera of the genus *Euphorbia*. Exploring the distribution of compound classes allowed us to spot differences and common features among samples (Supplementary Data). Consider *E. cyparissias* (Fig. 5c) and *E. balsamifera* (Fig. 5d): the variety of diterpenoids observed in the former (752 diterpenoids) was higher than in the latter (ten diterpenoids). CANOPUS annotated 22 and 48 compounds as bile acids in *E. cyparissias* and *E. balsamifera*, respectively. It is noteworthy that these are no annotation errors: the ClassyFire ChemOnt ontology does not distinguish between bile acids and phytosteroids.

Next, we compared the entire set of chemical class annotations for these *Euphorbia* species using hierarchical clustering based on CANOPUS class annotations (Fig. 5b). The resulting chemodendogram shows relatively good agreement with a phylogenetic tree of *Euphorbia* computed from genomic data[27,52,53]: the normalized quartet distance between the chemodendogram and phylogenetic tree was 0.396, compared to 0.666 for two random trees with $n=43$ taxa (s.d. = 0.0160, $z=-16.95$). Results show that, apart from a few species not grouping with their monophyletic counterparts, the chemodendogram matches monophyletic clades. Differing from the phylogenetic tree, the chemodendogram regrouped the subgenera *Athymalus*[54] and *Euphorbia*[55], both largely composed of succulent species and with a high degree of Crassulacean acid metabolism (CAM) photorespiration, a metabolism adaptation beneficial to plants facing environmental pressure in a dry climate[53]. In the same fashion, the chemodendrogram regrouped the subgenera *Esula* and *Chamaesyce*, which consist for the most part of herbaceous/leafy species[51,56], using the C3 pathway to fix carbon dioxide[52]. A few species from the subgenera *Euphorbia* and *Chamaesyce* were not placed in their respective phylogenetic clade, and instead were grouped with the mostly succulent subgenus, *Athymalus*. Examination of their morphology and distribution shows that all of these, except *E. cotinifolia*, are highly succulent and adapted to dry areas, suggesting a high degree of CAM metabolism. Regarding the two other species not grouped in their phylogenetic clade, *E. cotinifolia* (subgenus *Chamaesyce*) and *E. myrsinites* (subgenus *Esula*), the chemodendogram from CANOPUS indicates that the metabolome of succulent species resembles that of other succulent *Euphorbia* species. These drought-resistant species, despite not having a highly succulent morphology, may have some level of CAM metabolism. A previous study analyzing the same dataset found a relationship between *Euphorbia* diterpenes and herbivore distribution pattern[27]. In the present study, the comprehensive annotation of CANOPUS allowed us to observe that the metabolome of *Euphorbia* species appears to be driven by both the phylogenic distance, and the dominant photorespiration metabolism used by the species, which relates to its lifestyle and habitat. These results are supported by the fact that CAM adaptation occurred independently on many occasions in the genus *Euphorbia*[53].

CANOPUS accelerates the annotation of a novel cyclodepsipeptide, rivulariapeptolide 1155. The structural elucidation of natural products is a time-consuming manual task that remains a bottleneck

in the discovery of drug leads. It is most frequently done by the interpretation of nuclear magnetic resonance (NMR) experiments that require compounds isolated purely. Because CANOPUS can rapidly annotate structural classes for any compound from crude mixtures, it has the potential to accelerate the recognition of unique structural features. CANOPUS was applied for the structural elucidation of a novel Ahp (3-amino-6-hydroxy-2-piperidone)-containing cyclodepsipeptide isolated from an extract of *Rivularia* sp., a marine filamentous cyanobacterium. Crude fractions were analyzed by LC–MS/MS and the data were processed with MZmine2 (ref. [57]). CANOPUS indicated that the major compound isolated was a depsipeptide (Fig. 6a,b), along with at least 20 other related compounds in the same range (*m*/*z* 900–1,250). Interestingly, for the isolated compound, none of the top candidates predicted by CSI:FingerID were depsipeptides, suggesting that this compound was unknown; notably, no structures exist in PubChem with the molecular formula $C_{59}H_{81}N_9O_{15}$, the formula of the major compound. Further, examination of CANOPUS results offered additional structural insights into several conserved amino acids that were very probably part of all of these related molecules, such as tyrosine (CANOPUS classification as 1-hydroxy-2-unsubstituted benzenoids and phenols), phenylalanine (CANOPUS classification as 'phenylpropanoids and polyketides' and 'benzene and substituted derivatives') and, most intriguingly, piperidinone moieties (CANOPUS classification as piperidinone and delta-lactams). To confirm these predictions and ultimately to resolve the planar structure of the compound, one- and two-dimensional (2D) NMR experiments were performed (Supplementary Table 5 and Supplementary Figs. 1–10). The NMR data confirmed the class annotations from CANOPUS and, with the combined data, the planar structure of the isolated compound was rapidly assigned as a novel Ahp-containing cyclodepsipeptide which we named rivulariapeptolide 1155. The identified structure agrees with CANOPUS predictions, including the correctly annotated Ahp-moiety (piperidinone), the *N*-methyltyrosine and the phenylalanine residues, along with a novel modified *N*-acylated proline residue (CANOPUS: substituted pyrrols; Fig. 6c); see Supplementary Note 2 for details.

Discussion

CANOPUS is an automated method for the systematic classification of compounds from fragmentation spectra. CANOPUS comprehensively predicted 2,497 compound classes for every query compound, and was best of class for this task. Clearly, CANOPUS is comprehensive only within the limits of the technology employed (for example, compounds that do not ionize, separate or fragment well cannot be annotated) and the available training data and ClassyFire classes. CANOPUS requires high-resolution MS/MS data: from our experience, a mass accuracy of 10 ppm or better is required. Surprisingly, CANOPUS can reliably predict compound classes for which few or no MS/MS training data are available and is not distracted if MS/MS data are available for only a subclass. Its integration into SIRIUS allowed us to perform the entire workflow from feature detection to compound classification with a single tool and on complete datasets; it is also available via recent GNPS workflows[58,59]. CANOPUS can also import data from popular mass spectrometry frameworks such as MZmine[57], OpenMS[60] and XCMS[61].

We have demonstrated how to use CANOPUS for comparative metabolomics as applied to microbiome research and chemotaxonomical investigations in plants. Class annotations allowed us to infer new biological findings without the need to annotate all MS/MS spectra with a spectral library or structure database. We found that gut microbiota are probably involved in the metabolization of glycosylated lipids such as plant saponins; similarly, CANOPUS can be applied to human microbiome data in a clinical setting. For the genus *Euphorbia*, CANOPUS allowed us to observe that it is shaped by both the phylogenetic relatedness and dominating photorespiration metabolism, and this despite the diversity of habitat and geological distribution. We also found a characteristic distribution pattern for benzoic acid esters in diterpenoids across *Euphorbia* subgenera. CANOPUS classifications can be used as part of semiquantitative and qualitative metabolomics data analysis; in particular, class distributions allow us to compare samples or species with little or no overlap in compounds. We anticipate that, in the foreseeable future, a large fraction of compounds detected by nontargeted mass spectrometry will remain without structural annotation; CANOPUS can classify these unannotated fragmentation spectra and it allows us directly to deduce information from compound class distribution. CANOPUS can also accelerate the traditional structure elucidation process, by automatically providing structural insights into new natural products from crude fractions. Beyond the examples given in this article, potential fields of application include natural products, foodomics, environmental research, drug degradation research and pathology. With further development and testing, CANOPUS may predict classes outside the ClassyFire ontology, and other chemical properties for which few or no MS/MS training data are available; examples include the prediction of antibiotic activity or toxicity of compounds.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-020-0740-8.

Received: 15 April 2020; Accepted: 16 October 2020;
Published online: 23 November 2020

References

1. Horai, H. et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
2. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
3. Guijas, C. et al. METLIN: a technology platform for identifying knowns and unknowns. *Anal. Chem.* **90**, 3156–3164 (2018).
4. Kind, T. et al. Identification of small molecules using accurate mass MS/MS search. *Mass Spectrom. Rev.* **37**, 513–532 (2018).
5. Allen, F., Greiner, R. & Wishart, D. Competitive fragmentation modeling of ESI–MS/MS spectra for putative metabolite identification. *Metabolomics* **11**, 98–110 (2015).
6. Brouard, C. et al. Fast metabolite identification with Input Output Kernel Regression. *Bioinformatics* **32**, i28–i36 (2016).
7. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl Acad. Sci. USA* **112**, 12580–12585 (2015).
8. Ridder, L. et al. Automatic chemical structure annotation of an LC-MS^n based metabolic profile from green tea. *Anal. Chem.* **85**, 6033–6040 (2013).
9. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J. & Neumann, S. MetFrag relaunched: incorporating strategies beyond in silico fragmentation. *J. Cheminform.* **8**, 3 (2016).
10. Tsugawa, H. et al. Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER Software. *Anal. Chem.* **88**, 7946–7958 (2016).
11. Schymanski, E. L. et al. Critical assessment of small molecule identification 2016: automated methods. *J. Cheminf.* **9**, 22 (2017).
12. Blaženović, I., Kind, T., Ji, J. & Fiehn, O. Software tools and approaches for compound identification of LC-MS/MS data in metabolomics. *Metabolites* **8**, 31 (2018).
13. Silva, R. R., Dorrestein, P. C. & Quinn, R. A. Illuminating the dark matter in metabolomics. *Proc. Natl Acad. Sci. USA* **112**, 12549–12550 (2015).
14. Tsugawa, H. Advances in computational metabolomics and databases deepen the understanding of metabolisms. *Curr. Opin. Biotechnol.* **54**, 10–17 (2018).
15. Montenegro-Burke, J. R., Guijas, C. & Siuzdak, G. METLIN: a tandem mass spectral library of standards. *Methods Mol. Biol.* **2104**, 149–163 (2020).

16. Vinaixa, M. et al. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: state of the field and future prospects. *Trends Anal. Chem.* **78**, 23–35 (2016).
17. Aksenov, A. A., Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology by mass spectrometry. *Nat. Rev. Chem.* **1**, 0054 (2017).
18. Frainay, C. et al. Mind the gap: mapping mass spectral databases in genome-scale metabolic networks reveals poorly covered areas. *Metabolites* **8**, 51 (2018).
19. Venkataraghavan, R., McLafferty, F. W. & Lear, G. E. Computer-aided interpretation of mass spectra. *Org. Mass Spectrom.* **2**, 1–15 (1969).
20. Curry, B. & Rumelhart, D. E. MSnet: a neural network that classifies mass spectra. *Tetrahedron Comput. Methodol.* **3**, 213–237 (1990).
21. Werther, W., Lohninger, H., Stancl, F. & Varmuza, K. Classification of mass spectra: a comparison of yes/no classification methods for the recognition of simple structural properties. *Chemom. Intell. Lab. Syst.* **22**, 63–76 (1994).
22. Heinonen, M., Shen, H., Zamboni, N. & Rousu, J. Metabolite identification and molecular fingerprint prediction via machine learning. *Bioinformatics* **28**, 2333–2341 (2012).
23. Hastings, J. et al. The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* **41**, D456–D463 (2013).
24. Rogers, F. B. Communications to the editor. *Bull. Med. Libr. Assoc.* **51**, 114–116 (1963).
25. Djoumbou Feunang, Y. et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *J. Cheminf.* **8**, 61 (2016).
26. Blaženović, I. et al. Structure annotation of all mass spectra in untargeted metabolomics. *Anal. Chem.* **91**, 2155–2162 (2019).
27. Ernst, M. et al. Assessing specialized metabolite diversity in the cosmopolitan plant genus *Euphorbia* L. *Front. Plant Sci.* **10**, 846 (2019).
28. Tsugawa, H. et al. A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nat. Methods* **16**, 295–298 (2019).
29. Barupal, D. K. & Fiehn, O. Chemical Similarity Enrichment Analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. *Sci. Rep.* **7**, 14567 (2017).
30. Rasche, F. et al. Identifying the unknowns by aligning fragmentation trees. *Anal. Chem.* **84**, 3417–3426 (2012).
31. Treutler, H. et al. Discovering regulated metabolite families in untargeted metabolomics studies. *Anal. Chem.* **88**, 8082–8090 (2016).
32. Ernst, M. et al. MolNetEnhancer: enhanced molecular networks by integrating metabolome mining and annotation tools. *Metabolites* **9**, 144 (2019).
33. Lowry, S. R. et al. Comparison of various K-nearest neighbor voting schemes with the self-training interpretive and retrieval system for identifying molecular substructures from mass spectral data. *Anal. Chem.* **49**, 1720–1722 (1977).
34. Askenazi, M. & Linial, M. ARISTO: ontological classification of small molecules by electron ionization-mass spectrometry. *Nucleic Acids Res.* **39**, W505–W510 (2011).
35. Peters, K. et al. Chemical diversity and classification of secondary metabolites in nine bryophyte species. *Metabolites* **9**, 222 (2019).
36. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
37. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**, 442–451 (1975).
38. Wolf, S., Schmidt, S., Müller-Hannemann, M. & Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **11**, 148 (2010).
39. Dührkop, K. et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nat. Methods* **16**, 299–302 (2019).
40. Cooper, B. T. et al. Hybrid search: a method for identifying metabolites absent from tandem mass spectrometry libraries. *Anal. Chem.* **91**, 13924–13932 (2019).
41. Allard, P.-M. et al. Integration of molecular networking and in-silico MS/MS fragmentation for natural products dereplication. *Anal. Chem.* **88**, 3317–3323 (2016).
42. Silva, R. R. et al. Propagating annotations of molecular networks using in silico fragmentation. *PLoS Comput. Biol.* **14**, e1006089 (2018).
43. Fox Ramos, A. E. et al. CANPA: computer-assisted natural products anticipation. *Anal. Chem.* **91**, 11247–11252 (2019).
44. Quinn, R. A. et al. Global chemical effects of the microbiome include new bile-acid conjugations. *Nature* **579**, 123–129 (2020).
45. Minamida, K. et al. Production of equol from daidzein by Gram-positive rod-shaped bacterium isolated from rat intestine. *J. Biosci. Bioeng.* **102**, 247–250 (2006).
46. Quinn, R. A. et al. Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy. *Trends Pharmacol. Sci.* **38**, 143–154 (2017).
47. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
48. Hooft, J. J. J., Wandy, J., Barrett, M. P., Burgess, K. E. V. & Rogers, S. Topic modeling for untargeted substructure exploration in metabolomics. *Proc. Natl Acad. Sci. USA* **113**, 13738–13743 (2016).
49. Vasas, A. & Hohmann, J. *Euphorbia* diterpenes: isolation, structure, biological activity, and synthesis (2008–2012). *Chem. Rev.* **114**, 8579–8612 (2014).
50. Yang, M. et al. Studies on the fragmentation pathways of ingenol esters isolated from *Euphorbia esula* using IT-MSn and Q-TOF-MS/MS methods in electrospray ionization mode. *Int. J. Mass Spectrom.* **323-324**, 55–62 (2012).
51. Riina, R. et al. A worldwide molecular phylogeny and classification of the leafy spurges, *Euphorbia* subgenus *Esula* (Euphorbiaceae). *TAXON* **62**, 316–342 (2013).
52. Horn, J. W. et al. Phylogenetics and the evolution of major structural characters in the giant genus *Euphorbia* L. (Euphorbiaceae). *Mol. Phylogenet. Evol.* **63**, 305–326 (2012).
53. Horn, J. W. et al. Evolutionary bursts in *Euphorbia* (Euphorbiaceae) are linked with photosynthetic pathway. *Evolution* **68**, 3485–3504 (2014).
54. Peirson, J. A., Bruyns, P. V., Riina, R., Morawetz, J. J. & Berry, P. E. A molecular phylogeny and classification of the largely succulent and mainly African *Euphorbia* subg. *Athymalus* (Euphorbiaceae). *TAXON* **62**, 1178–1199 (2013).
55. Dorsey, B. L. et al. Phylogenetics, morphological evolution, and classification of *Euphorbia* subgenus *Euphorbia*. *TAXON* **62**, 291–315 (2013).
56. Yang, Y. et al. Molecular phylogenetics and classification of *Euphorbia* subgenus *Chamaesyce* (Euphorbiaceae). *TAXON* **61**, 764–789 (2012).
57. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).
58. Nothias, L.-F. et al. Feature-based molecular networking in the GNPS analysis environment. *Nat. Methods* **17**, 905–908 (2020).
59. Schmid, R. et al. Ion identity molecular networking in the GNPS Environment. Preprint at *bioRxiv* https://doi.org/10.1101/2020.05.11.088948 (2020).
60. Röst, H. L. et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. *Nat. Methods* **13**, 741–748 (2016).
61. Benton, H. P., Wong, D. M., Trauger, S. A. & Siuzdak, G. XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization. *Anal. Chem.* **80**, 6382–6389 (2008).

Methods

Training and evaluation data. To train the kernel SVMs for positive ion mode spectra, we used mass spectral data from 24,539 compounds with 16,710 unique 2D structures, 13,708 from NIST 2017 compounds (commercial; NIST, v.17), 8,573 from GNPS[2] and 2,277 from MassBank[1]. This MS/MS dataset is referred to as the SVM training dataset, to differentiate it from the structure dataset used to train CANOPUS (Supplementary Table 6). As an independent dataset, we used mass spectral data from 3,387 compounds in the MassHunter Forensics/Toxicology PCDL library (Agilent Technologies).

For negative ion mode, we trained the SVMs on spectral data from 16,785 compounds with 8,079 unique 2D structures (5,517 from NIST 2017, 4,493 from GNPS and 6,775 from MassBank).

The training set for the DNN consisted of molecular structures from numerous public databases including KNApSAcK[62], HMDB[63], KEGG[64], UNDP and others. In addition, we added structures from PubChem until we had obtained at least 400 example structures per compound class. We excluded all structures from the DNN training dataset contained in either the SVM training or independent dataset. In total, the DNN training set consisted of 4,104,871 structures along with their ClassyFire compound classes and molecular fingerprints.

Chemical classes. The ChemOnt ontology of ClassyFire consists of 4,825 classes[25], which are organized as a tree. In practice, every compound is assigned to several classes. For many classes we did not find a single positive example in any biological structure database. We used 2,497 classes for which at least 400 example structures were present in PubChem. We excluded inorganic compounds and all their subclasses; see Supplementary Table 1 for a list of compound classes not predicted by CANOPUS.

Historically, biomolecules have been classified based on either their common biosynthetical origin or chemical characteristics. However, by so doing, automated classification of compounds is a non-trivial problem even when the compound structure is known[65]. In contrast, ClassyFire classes are not dependent on biological precursors or characteristics but, instead, can be deterministically computed from the molecular structure. This allows us to assign each class for every compound in a structure database. Certain, often rather complex, ClassyFire classes (for example, leukotrienes[25] and bile acids discussed above) can deviate from what an expert might expect for the corresponding class names.

Molecular fingerprints and fingerprint prediction. As described in refs. [7,39], we used molecular properties from several known molecular fingerprints, namely CDK Substructure, PubChem CACTVS, Klekota-Roth[66], FP3, MACCS and Extended Connectivity[67]. Molecular properties were computed from molecular structure using the Chemistry Development Kit (CDK) v.2.1.1 (ref. [68]). In addition, we used 490 molecular properties that describe larger substructures that will be added to SIRIUS and CSI:FingerID in an upcoming release. These additional molecular properties did not result in improved prediction performance of CANOPUS, so we have omitted details. We discarded molecular properties with <20 positive examples and that could not be predicted with reasonable quality ($F_1 > 0.25$) during cross-validation.The F_1 is the harmonic mean of precision and recall. In total, CSI:FingerID predicted 4,262 molecular properties.

We found that building molecular structures from IUPAC International Chemical Identifiers resulted in inconsistent representations of molecules because structures were not standardized. Hence, we now build molecular structures from PubChem canonical simplified molecular input line entry specifications (SMILES) in all cases[69].

Training the CSI:FingerID kernel SVMs was performed as described in ref. [39]. We trained ten models in tenfold cross-validation such that every model was trained on 90% of the structures. Cross-validation was performed structure-disjoint: whenever we predicted the molecular fingerprint of a query compound in evaluation, we used the cross-validation model that had not seen the query structure during training. We did so for both the SVM training and independent datasets. To this end, all compounds are novel in the sense that none could be identified by dereplication (spectral library searching).

Prediction of compound classes from molecular fingerprints. Assuming that we know the exact molecular fingerprint and molecular formula of a query compound, but not its structure, can we predict whether it belongs to a certain compound class? We used a DNN[36] for this purpose, simultaneously predicting all compound classes. DNNs can be trained with millions of training examples within a reasonable timeframe. Molecular formulae are encoded as feature vectors containing the number of atoms for each element, the mass, the ring double-bond equivalent (RDBE) value and the ratios between certain elements (see Supplementary Table 7 for a description of all features). The binary molecular fingerprint and molecular formula features constitute the input of the DNN.

The above DNN appears to be of no practical use: to compute the exact fingerprint of a compound, we have to know its molecular structure; however, if we know the molecular structure there is no need to apply the DNN because we can deterministically find the correct answer using ClassyFire. The answer is that we can use a predicted fingerprint as input of the DNN: we use the probabilistic molecular fingerprint predicted by CSI:FingerID as well as the molecular formula computed by SIRIUS (see below) as input. The predicted fingerprint can either be transformed to a binary fingerprint, or we can directly use the probabilistic fingerprint as input of the DNN; for CANOPUS, we use the latter option.

Simulation of probabilistic fingerprints. It turns out that we can improve the performance of the above DNN as follows. During training of the DNN, we use binary fingerprints (computed from structure) whereas in application we use a probabilistic fingerprint predicted from the query MS/MS spectrum; the latter fingerprint contains errors and uncertainties. To improve DNN performance, we present two probabilistic methods to introduce errors and uncertainties into the training fingerprints. Due to the probabilistic nature of the methods, one molecular structure will result in different probabilistic fingerprints, all of which we can use to train the DNN.

The first method of sampling probabilistic fingerprints considers all molecular properties individually: for each molecular property i, we have trained an individual SVM to predict the property from MS/MS data. We record all Platt probabilities that were estimated for positive property i: let $\mathcal{P}_i$ be the set of all Platt probability estimates in cross-validation for all MS/MS spectra in the training set where the SVM should have predicted a positive outcome for property i. Analogously, we record Platt probability estimates for negative property i in $\mathcal{N}_i$. Given a compound from the structure database, we know its exact binary fingerprint. Consider a molecular property i: if the property is present in the compound, we can uniformly sample from $\mathcal{P}_i$ to simulate a Platt probability; if this is not present, we uniformly sample from $\mathcal{N}_i$. However, this may result in overfitting of the DNN, as we are using exactly the same real numbers for training that are later used for evaluation. To this end, we also sample from the 'holes' between values in $\mathcal{P}_i$: we sort the set $\mathcal{P}_i := \{p_1, \ldots, p_n\}$ such that $p_1 \leq \ldots \leq p_n$, and assume $p_0 := 0$ and $p_{n+1} := 1$. We uniformly draw a random number $x \in (0, n+1)$ and then interpolate between values p_k and p_{k+1}, with $k := \lfloor x \rfloor$:

$$y \leftarrow (1 - x + k) \times p_k + (x - k) \times p_{k+1}$$

is the simulated Platt probability. Analogously, we can do so for $\mathcal{N}_i$. This can be interpreted as drawing random numbers using a kernel density estimate of the observed Platt probabilities; we avoid elaborate kernel estimates to guarantee swift running times.

The disadvantage of drawing every position independently is that we completely ignore correlations between fingerprint positions. If two molecular properties are highly correlated, we can assume that prediction errors also correlate. To this end, our second sampling method draws multiple positions at once. First, we define further sets of Platt probabilities for each position in the fingerprint, namely TP_i, FP_i, TN_i and FN_i. These are defined analogously to $\mathcal{P}_i$ and $\mathcal{N}_i$, but contain Platt estimates for true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) for molecular property i. Let $\mathcal{F}$ be a binary fingerprint from the structure database for which we want to simulate a probabilistic fingerprint. We measure the similarity of two fingerprints using the Tanimoto coefficient (Jaccard index)[70]. We first sort all structures from the SVM training dataset in descending order of their Tanimoto coefficient to $\mathcal{F}$. We then pick the kth structure and its binary fingerprint B, together with its predicted probabilistic fingerprint $\tilde{B}$. Here, k is a random number drawn from a geometric distribution with small parameter p; we then use $p = 0.2$. For each molecular property i with $B_i = \mathcal{F}_i$, we randomly draw a Platt probability from the appropriate set TP_i, FP_i, TN_i and FN_i, using the procedure described above. For example, if $\tilde{B}_i \geq 0.5$ and $B_i = 1$, this is a true positive prediction and we sample from TP_i; if $\tilde{B}_i \geq 0.5$ and $B_i = 0$, this is a false positive prediction and we sample from FP_i, and so on. The remaining positions differ between $\mathcal{F}$ and B; for these, we repeat the above procedure: This time, we use only the subset of remaining positions to calculate the Tanimoto between $\mathcal{F}$ and the structures from the SVM training data, and for sampling the probabilities.

We can think of the second sampling method as combining the simulated fingerprint using those parts of probabilistic fingerprints available for training. One may assume that this sampling method yields 'more realistic' probabilistic fingerprints. On the other hand, it may lead to probabilistic fingerprints that are 'too similar' to the training data and, thus, may result in overfitting of the DNN. As both sampling strategies have advantages and disadvantages, we do not select one only but simulate fingerprints using the first and the second sampling strategies alternately.

DNN architecture and training. Both input layers are centered feature-wise: to center the molecular fingerprint, we calculate the mean for every predicted molecular property in the SVM training dataset. We stress that no structures from the SVM training dataset are used for training the DNN, but only statistics about these structures such as the mean of each molecular property, as this is needed for centering the feature vectors; we also use the SVM training dataset to determine the early stopping of DNN training. Recall that no structures from the independent dataset are used for training the DNN, either. The molecular formula feature vectors are normalized such that every feature has unit variance; we use all molecular formulae from the DNN structure training database to calculate the empirical mean and variance.

Rather than concatenating both input layers directly, we first connect the fingerprint input layer with a fully connected inner layer of 3,000 neurons, and the molecular formula input layer with a fully connected inner layer of 16 neurons (Fig. 7). The outputs of both inner layers are concatenated and fed into two additional inner layers of 3,000 and 8,000 neurons. All inner layers use the rectified linear activation function (ReLu), a learnable bias, an L_2 regularization (ridge regression) as well as a dropout of 50% of the neurons[71]. The output layer is linear and predicts 2,497 compound categories. We use the sigmoid cross-entropy as loss function and Adam as optimizer[72]. The DNN training is performed using the TensorFlow library[73]. Training is done in minibatches of roughly 12,000 structures; we draw compounds from the pool of all available training compounds such that every compound class is present in each minibatch at least once. The sign of each output neuron encodes whether a compound belongs to the respective compound class. We stop training after 12,000 iterations and do not report epochs, because our training data are randomized.

The following is not performed as part of our method evaluation, but only when applying the method to biological data: We transform the linear outputs of the DNN into posterior probabilities using Platt calibration[74]. For learning the parameters of the logistic function, we use the predicted fingerprints from the SVM training dataset. If we do not have sufficient positive examples (<30), we add simulated probabilistic fingerprints. After calibration we update the weights of the network one final time, using the SVM training dataset as input.

Assignment of molecular formulae. Using the molecular formula as input of the DNN requires us first to identify the correct molecular formula. However, this is necessary anyway because the kernel support vector machines[7] used here operate on fragmentation trees, which are an outcome of molecular formula identification with SIRIUS[75].

Because our method is targeting of novel compounds, we must not assume that the molecular formula of the query compound is recorded in any molecular structure database. We initially use both isotope and fragmentation patterns to determine the molecular formula, using SIRIUS 4 (ref. [39]). Unfortunately, molecular formula identification rates drop markedly for compounds >500 Da (see Fig. 5 in ref. [75]). ZODIAC improves molecular formula annotations of complete LC–MS runs using a network-based approach, where compatible molecular formula assignments support each other and assignments for the complete dataset are estimated by Gibbs sampling; in evaluations, ZODIAC incurs substantially smaller error rates than SIRIUS[76]. We add all reference compounds from GNPS, MassBank and NIST as anchors into the ZODIAC network.

Method evaluation. We evaluate against four other methods for compound class assignment: direct prediction, *k*-nearest neighbor (KNN) using either MetFrag[38] or CSI:FingerID[7] as the underlying search engine, and KNN searching in a spectral library. All of these baseline methods are nontrivial: KNN methods are often employed in practice to transform similarity search results to a binary classification, considering the top *k* of the similarity search by majority vote. This is an 'archetypical' baseline method against which to evaluate, and is usually hard to better using more advanced classifiers.

Direct prediction. We can think of the compound classes as additional molecular properties of a compound that directly predict the corresponding fingerprint from MS/MS data. This approach has been suggested repeatedly in the literature, in particular for gas chromatography–MS data, but usually for targeting of compound classes defined by the presence or absence of a certain substructure. Here, we employ the kernel SVM machinery behind CSI:FingerID for direct prediction, following the usual CSI:FingerID training and evaluation setup. We argue that this setup is currently best in class for direct prediction. Consequently, this baseline method inherits the intricate setup of multiple kernel learning on fragmentation trees, which is responsible for the excellent performance of CSI:FingerID. By design, direct prediction cannot predict compound classes for which there are no positive MS/MS training data; it cannot predict compound classes if training data are available for a subclass only, as the resulting predictor is in fact targeting the subclass; and, as we do not distinguish between compounds with known structure and compounds with MS/MS data, it is not possible to evaluate whether direct prediction will show reasonable performance for truly novel compounds of unknown structure.

KNN. MetFrag and CSI:FingerID are methods employed for compound structure identification by searching MS/MS spectra in a structure database such as PubChem. Given an MS/MS spectrum, both tools report a ranked list of structure candidates. This approach must fail for novel structures that are not included in structure databases. However, for the task of compound classification it is sufficient that the top-ranking candidates belong to the same compound class as the query compound. The compound classes for any given candidate structure are computed using the ClassyFire webservice (http://classyfire.wishartlab.com/). A KNN uses the compound class annotations of the top *k* candidates and, for each compound class, decides via majority vote (yes/no) of the top *k* candidates whether that class is assigned to the query. We search in the PubChem structure database, considering all structures with the same molecular formula as the query, and use MetFrag or CSI:FingerID as the search engine. We remove structures in the independent dataset from PubChem and the SVM training dataset before searching, as was done for training the DNN. This corresponds to the situation where the measured compound is a novel structure. We evaluated parameters $k=1$ and $k=5$, and found that $k=5$ performs slightly better for both MetFrag and CSI:FingerID (data not shown).

It is noteworthy that this approach has a number of conceptual shortcomings. In particular, if there are no molecular structures with the query molecular formula then the method cannot assign any compound classes; this is the case for the novel compound rivulariapeptolide 1155. This shortcoming is the most obvious, but indicates only an underlying issue resulting in numerous similar problems: as an example, for $k=5$ we need at least three positive examples with this molecular formula so that a compound class can be positively assigned. Presented with an outlier structure substantially different from all other structures with the same molecular formula, we will incorrectly predict all corresponding compound classes even if the correct structure is part of the structure database searched. More generally, if a certain compound class is over- or under-represented for a particular molecular formula, then it is more likely to be predicted 'yes' or 'no', respectively, independent of the actual data. These shortcomings do not necessarily result in bad evaluation statistics: considering any reference compound used for evaluation, not only is its structure present in PubChem but also several structures with high similarity and identical molecular formula. To this end, evaluation statistics of KNN classifiers must be interpreted with care. It is also noteworthy that the above-mentioned issues are not those usually attributed to KNN classifiers (curse of dimensionality, label noise).

Rather than searching in a structure database, we can also compare the query spectrum against spectra in a spectral library and thereby obtain a ranked list of reference spectrum candidates. Considering the small size of the spectral library, we do not filter by the precursor mass of the query. For comparison of two spectra, we use the cosine similarity of the two spectra as well as of the inverse spectra (obtained by subtracting each peak mass from the precursor mass), and sum both values. Using the spectrum and its inverse allows comparison of spectra of compounds differing in mass but structurally very similar[30,77]. We search in the SVM training dataset as our spectral library. For each query, we remove all spectra from the spectral library corresponding to the query structure. We found that hybrid search works better than simply comparing spectra by their cosine (data not shown). All of the above-mentioned issues also apply to the spectral library KNN, considering its size; in addition, it suffers from all the issues of direct prediction.

MCC. The MCC is defined as:

$$\mathrm{MCC} = \frac{\mathrm{TP}\times\mathrm{TN} - \mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}$$

Where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives, respectively, of a binary classification[37]. The MCC lies between −1 and +1, where +1 corresponds to a perfect classification, 0 to a random classification and −1 to a 'perfectly wrong' classification. The MCC is considered more informative than the F_1 score and accuracy, because it takes into account the balance ratios of these four values; in particular, the F_1 score of a random classification is nonzero and depends on the actual classification task, rendering predictor performance incomparable. The MCC is equivalent to the Pearson correlation coefficient between observations and predictions.

LC–MS data processing. For processing in SIRIUS v.4.4, one or more LC–MS/MS runs must be provided in mzML or mzXML format. Feature detection in SIRIUS v.4.4 is similar in essence to a targeted analysis: rather than searching for all features in a run, SIRIUS first collects all MS/MS spectra and their precursor information. It then searches for features that are associated with those MS/MS spectra—precursor ions, adduct ions and isotope peaks. The precursor information reveals the retention time and mass range where precursor ions can be detected. Adducts and isotopes can be found using predefined lists of mass differences. Fragmentation spectra assigned to the same feature (precursor ion) are merged. SIRIUS 4 uses a greedy feature alignment method similar to that of MZmine[57].

Mice dataset. We analyzed 834 LC–MS runs from MassIVE (id no. MSV000079949)[44]. The corresponding samples were taken from different organs of eight mice (four SPF and four GF). Feature detection and feature alignment were performed using SIRIUS v.4.4. We retained only those features for which we had MS/MS measurements in at least two samples and at least three explainable peaks in MS/MS. We used ZODIAC to improve molecular formula annotation[76]. To speed up running times, only compounds <860 Da were considered. All compound classes were predicted for all remaining MS/MS, and all of these were used in subsequent enrichment analysis.

This resulted in a feature quantification table of 5,763 compounds, where each row corresponds to a compound and each column to its intensity (maximum peak height) per sample. We subtracted the blank intensities from the table and performed quantile normalization to make the samples comparable: 4,109 compounds remained after blank subtraction. Next, we analyzed those

compound classes with different abundance between GF and SPF in the digestive system. In the following, we consider only those samples taken from mouth, esophagus, stomach, duodenum, jejunum, ileum, cecum, colon and stool.

Summing up all intensities over all compounds of a class already gives helpful insights (Fig. 3b–d). However, this assumes a change in the intensity of all compounds of a class in the same direction. A better indicator is fold change, which is calculated as the truncated mean (5% truncated at each end) of all intensities in GF samples divided by the truncated mean of all intensities in SPF samples. Truncated means are taken over all organs—that is, mouth to stool. A pseudo-count must be added beforehand, to avoid division by zero. We use the 1% percentile of nonzero intensities in the quantification table as a pseudo-count.

We sorted the 4,109 compounds by absolute logarithm (base 10) of their fold changes. Values <1 (fold change between 0.1 and 10) were set to zero, because we do not consider such fold changes informative. For each compound class, we checked whether compounds of that class appeared more often in the higher discriminative region; this was done using a one-tailed Mann–Whitney *U*-test[78]. We then sorted compound classes based on their *P* value. Compound classes with the lowest *P* value are those that triggered the metabolic difference between GF and SPF samples (Supplementary Table 4).

We uploaded quantification table and input MS/MS spectra to GNPS and performed feature-based molecular networking[2,58]. We downloaded the Cytoscape files for network visualization and mapped CANOPUS class annotations to each node in the network. For evaluation of results, we compared selected class annotations with both GNPS library hits and the CSI:FingerID database search results (searching in PubChem).

***Euphorbia* dataset.** We used LC–MS runs from the study by Ernst et al.[27] downloaded from MassIVE (id no. MSV000081082). In total, we analyzed samples from 43 different *Euphorbia* species. Feature detection was performed using SIRIUS v.4.4. In contrast to ref. [27] we did not align features, because samples were taken from different species and, therefore, have a different metabolic composition (see Fig. 3b in ref. [27]). For most samples, more compounds were detected with SIRIUS than in ref. [27] (Extended Data Fig. 6). For molecular formula annotation we used SIRIUS and ZODIAC[39,76]. ZODIAC processed the plant samples of all species within one network. Compound classes were predicted for all compounds using the best-scoring ZODIAC molecular formula. To count the number of compounds per class, we considered all compound class assignments with probability >0.5.

Chemotaxonomy. We used WPGMA hierarchical clustering to compute the dendrogram in Fig. 5b. As the distance metric, we chose the Euclidean distance over normalized compound class distribution; this is the vector with the logarithmized number of compounds per compound class plus one, centered to zero mean and unit variance. It transpires that this is critical, because the number of detected compounds differs substantially between species (Extended Data Fig. 6). We used the scipy library for clustering[79].

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Input mzML/mzXML files are available at MassIVE (https://massive.ucsd.edu/) with the accession nos. MSV000079949 (mice data) and MSV000081082 (*Euphorbia* data). The mass spectrometry data for *Rivularia* sp. cyanobacteria were deposited at MassIVE (accession no. MSV000085578). The spectra for rivulariapeptolide 1155 were annotated in the GNPS spectral library (accession nos. CCMSLIB00005723986 and CCMSLIB00005723388). The structure database with ClassyFire annotations, the publicly available part of the evaluation data and the Cytoscape files for network visualization can be downloaded from https://bio.informatik.uni-jena.de/data/ and https://doi.org/10.6084/m9.figshare.13073051. Source data are provided with this paper.

Code availability

CANOPUS is part of SIRIUS software and can be downloaded from https://bio.informatik.uni-jena.de/software/canopus/. The source code of CANOPUS is available at https://github.com/boecker-lab/sirius-libs. The scripts for analysis and visualization of CANOPUS results are available at https://github.com/kaibioinfo/canopus_treemap.

References

62. Shinbo, Y. et al. in *Plant Metabolomics* Vol. 57 (eds Saito, K. et al.) 165–181 (Springer, 2006).
63. Wishart, D. S. et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
64. Kanehisa, M., Goto, S., Kawashima, S. & Nakaya, A. The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42–46 (2002).
65. Bobach, C., Böhme, T., Laube, U., Püschel, A. & Weber, L. Automated compound classification using a chemical ontology. *J. Cheminform.* **4**, 40 (2012).
66. Klekota, J. & Roth, F. P. Chemical substructures that enrich for biological activity. *Bioinformatics* **24**, 2518–2525 (2008).
67. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
68. Willighagen, E. L. et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminf.* **9**, 33 (2017).
69. Hähnke, V. D., Kim, S. & Bolton, E. E. PubChem chemical structure standardization. *J. Cheminf.* **10**, 36 (2018).
70. Rogers, D. J. & Tanimoto, T. T. A computer program for classifying plants. *Science* **132**, 1115–1118 (1960).
71. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
72. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).
73. Abadi, M. N. et al. in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (eds Keeton, K. & Roscoe, T.) 265–283 (USENIX, 2016).
74. Platt, J. C. *Advances in Large Margin Classifiers* (MIT Press, 2000).
75. Böcker, S. & Dührkop, K. Fragmentation trees reloaded. *J. Cheminform.* **8**, 5 (2016).
76. Ludwig, M. et al. Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nat. Mach. Intell.* **2**, 629–641 (2020).
77. Moorthy, A. S., Wallace, W. E., Kearsley, A. J., Tchekhovskoi, D. V. & Stein, S. E. Combining fragment-ion and neutral-loss matching during mass spectral library searching: a new general purpose algorithm applicable to illicit drug identification. *Anal Chem.* **89**, 13261–13268 (2017).
78. Mann, H. B. & Whitney, D. R. On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat.* **18**, 50–60 (1947).
79. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Meth.* **17**, 261–272 (2020).

Acknowledgements

We thank Deutsche Forschungsgemeinschaft for providing financial support (no. BO 1910/20 to S.B., K.D. and M.L. and no. PE 2600/1 to D.P.), and the Academy of Finland (no. 310107/MACOME to J.R.). P.C.D., R.R. and W.H.G. were supported by the Gordon and Betty Moore Foundation (no. GBMF7622) and by the US National Institutes of Health (NIH; no. R01 GM107550). P.C.D. was supported by NIH grants nos. P41 GM103484 and R03 CA211211. L.-F.N. was supported by NIH grant no. R01 GM107550 and by the European Union's Horizon 2020 program (MSCA-GF, no. 704786). We thank F. Kuhlmann and Agilent Technologies, Inc. for providing data used in the evaluation of CANOPUS. We thank Y. Djoumbou Feunang, D. Arndt and D. Wishart for providing ClassyFire annotations for a database of molecular structures. We thank K. Alexander, E. Caro-Diaz and B. Naman for assistance with the collection of *Rivularia* sp. Further, we thank S. Whitner and K. Joosten for 16S recombinant DNA analysis. We thank M. Ernst for valuable discussions on the *Euphorbia* plant study, and J. van der Hooft and S. Rogers for feedback on the manuscript.

Author contributions

K.D., J.R. and S.B. designed the research. K.D. and S.B. developed the computational method. K.D. implemented the computational method with contributions from M.L., M.F. and M.A.H. M.F. integrated CANOPUS into SIRIUS v.4.4. K.D., L.-F.N. and P.C.D. applied and evaluated the method in the mouse and *Euphorbia* studies. R.R. isolated rivulariapeptolide 1155 and applied CANOPUS (on mass spectrometry data collected and analyzed by D.P. and R.R. and supervised by W.H.G.) and one-/two-dimensional NMR analysis for its structural elucidation. K.D., S.B., L.-F.N. and R.R. wrote the manuscript, in concert with all authors.

Competing interests

S.B., K.D., M.L., M.F. and M.A.H. are cofounders of Bright Giant GmbH. P.C.D. is scientific advisor for Sirenas LLC.

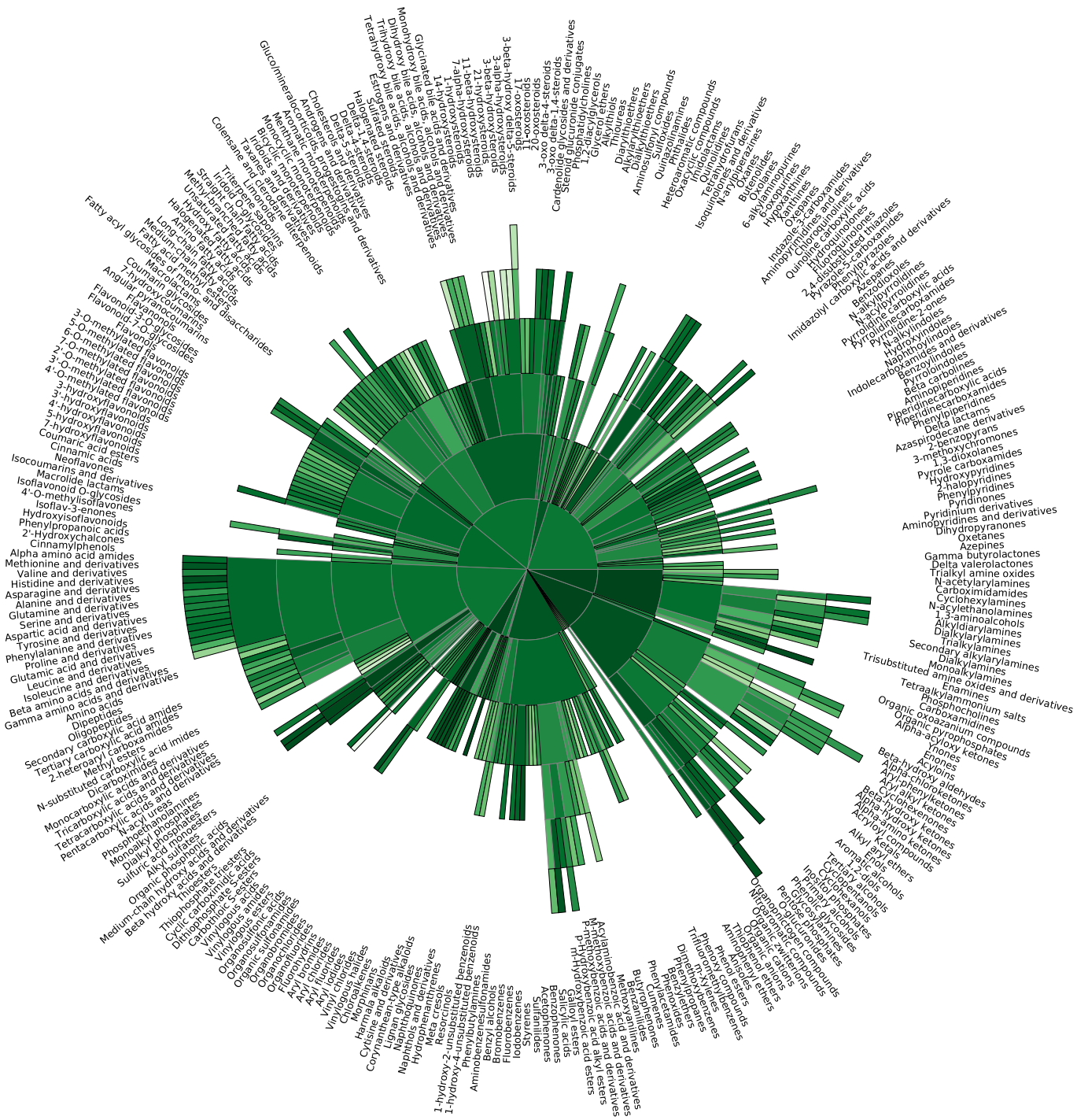
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-020-0740-8.

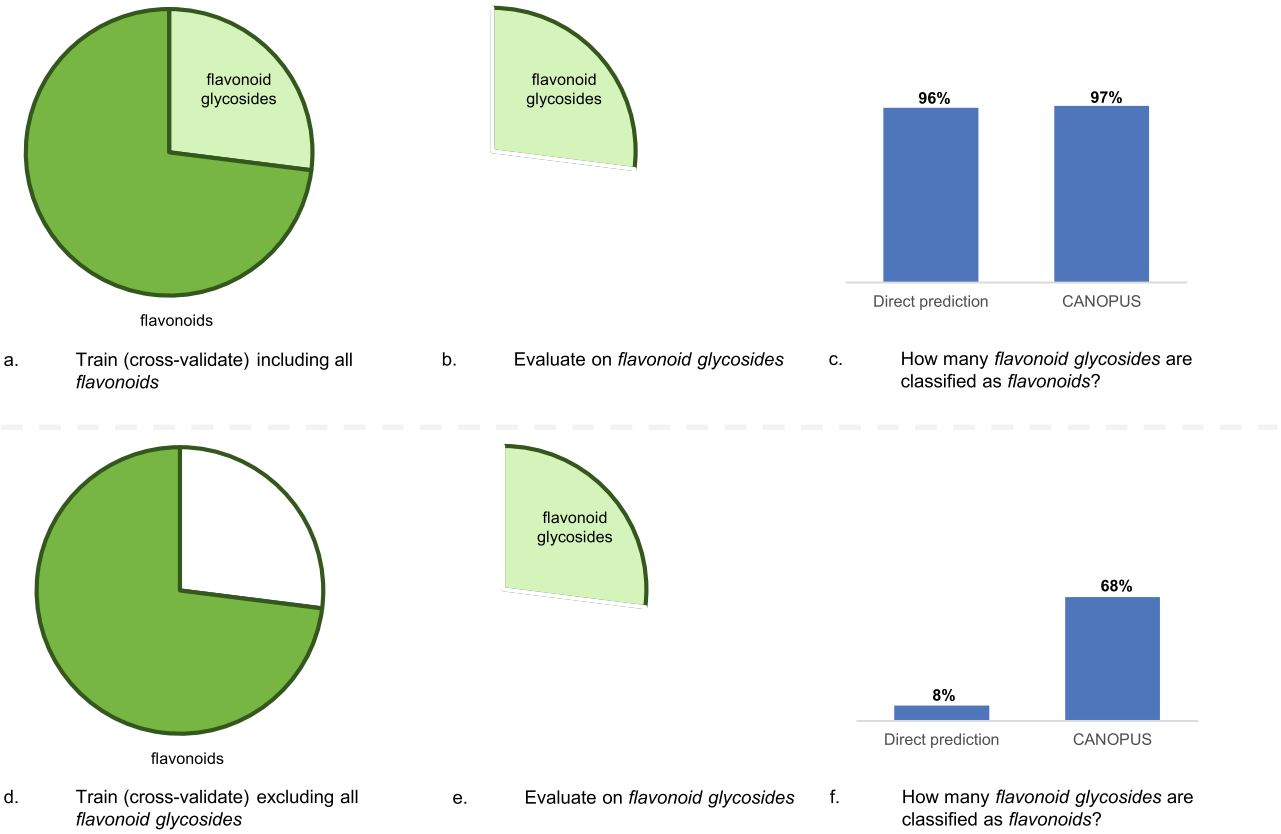
Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-0740-8.

Correspondence and requests for materials should be addressed to S.B.

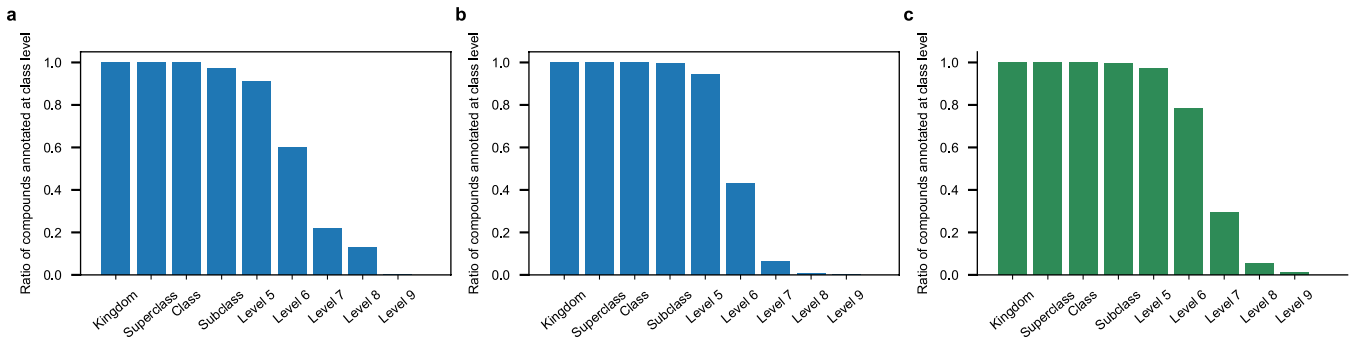
Reprints and permissions information is available at www.nature.com/reprints.



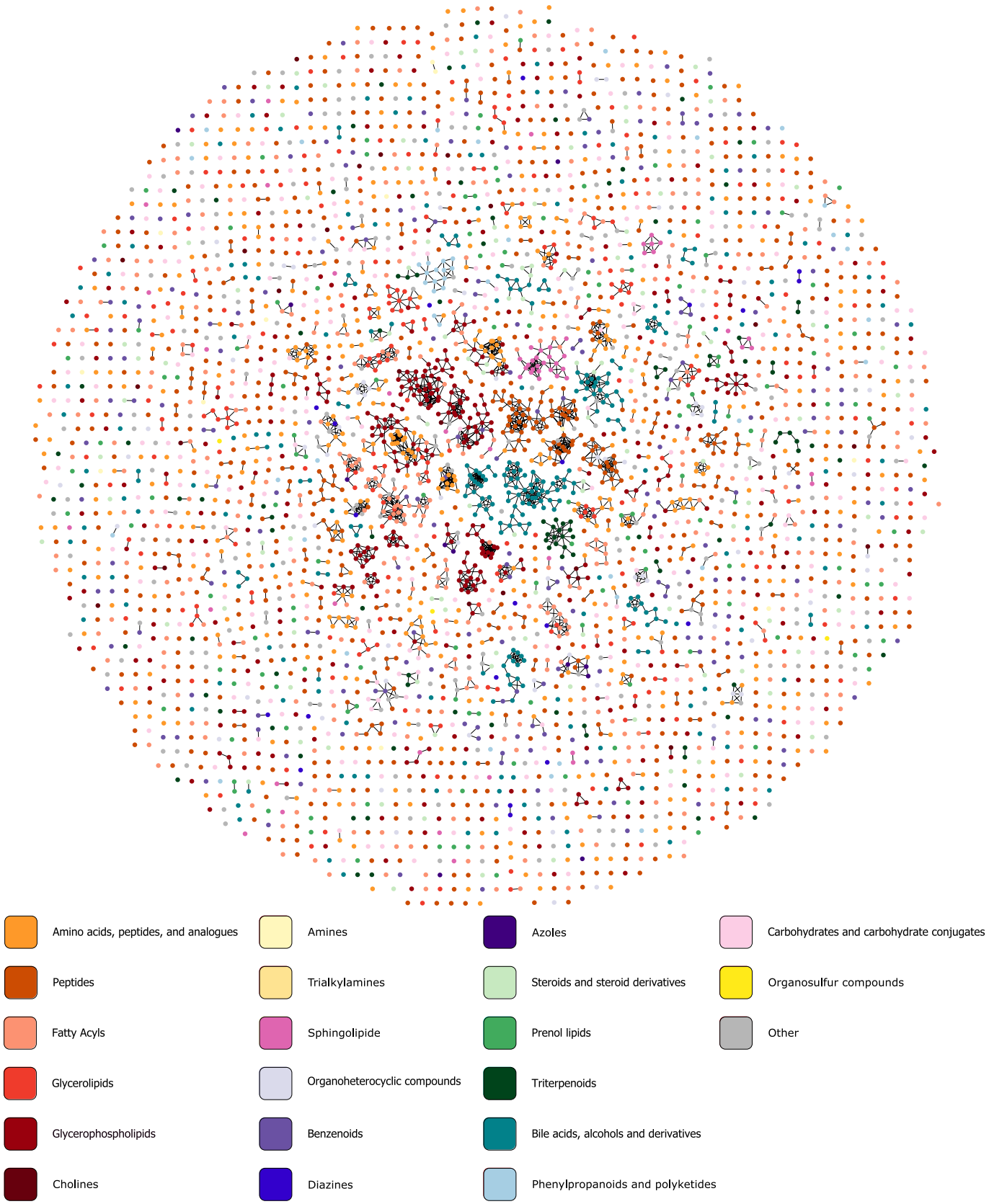
Extended Data Fig. 1 | CANOPUS performance sunburst plot. Matthews correlation coefficient (MCC) for the 782 of 2,497 compound classes with at least 50 positive examples. SVM training dataset. A darker green coloring corresponds to better prediction performance for the class. The size of each slice is chosen such that all classes fit into the figure and has no further meaning. Inner slices represent parent classes of outer slices.



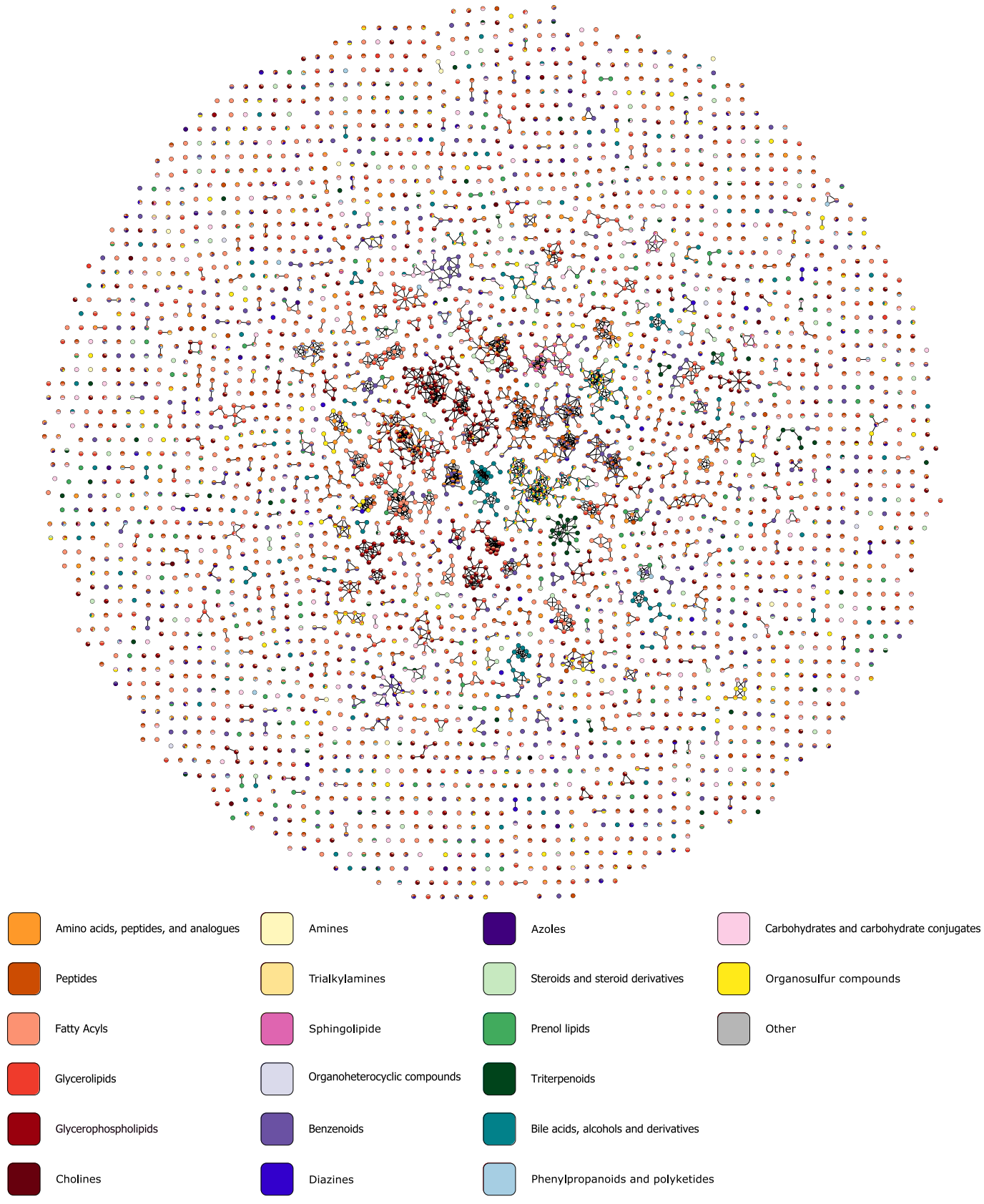
Extended Data Fig. 2 | Effect of removing a subclass from the MS/MS training data. a–c, Regular evaluation setup: classes and subclasses are distributed into cross-validation folds, ensuring that methods are never evaluated on the same MS/MS data or structures they were trained on. **d**-**f**, We remove all flavonoid glycosides (the subclass) from the MS/MS training data (**d**), and then evaluate the predictor for glycosides (the class) on these removed MS/MS spectra (**e**). A perfect method would still classify all flavonoid glycoside MS/MS spectra as glycosides (**f**). CANOPUS exhibits only a small drop (68% to 97%) in correct classifications (**c**,**f**). In contrast, direct prediction performed mostly on par with CANOPUS before removing flavonoid glycosides from the MS/MS training data (**c**), but misses almost all of them (8%) afterwards (**f**). We were able to attribute this to the presence of isoflavonoid glycosides in the training data; these do not belong to the flavonoid class, but have highly similar structures and MS/MS spectra, except for the presence of a sugar residue. We observed that direct prediction in (**d**-**f**) uses the presence of a sugar residue to infer that a MS/MS spectrum is not a glycoside. In contrast, CANOPUS does not fall for this 'bait'; heterogeneous training allows us to integrate the substantially more comprehensive structure data in its predictions.



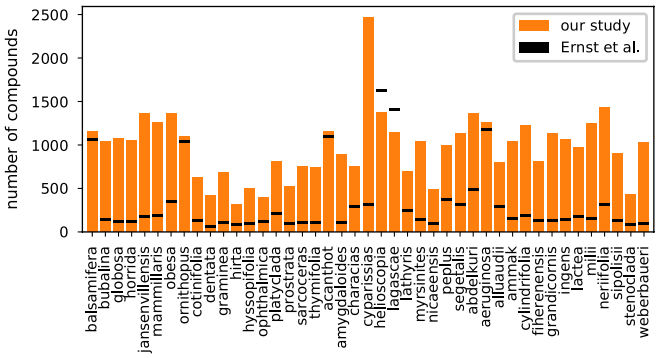
Extended Data Fig. 3 | Relative number of compounds annotated at varying ClassyFire class levels in the mice study (a) and the Euphorbia plant study (b). The ClassyFire ChemOnt ontology is organized as a tree, where the Kingdom is either Organic compounds or Inorganic compounds. Superclasses like Lipids and lipid like molecules, Benzenoids are children of Kingdom class. Flavonoids and Steroids and steroid derivatives are examples for the Class level, while Flavonoid glycosides and Bile acids, alcohols, and derivatives are examples for subclasses. There can be up to 11 levels in the ontology. **c**, ClassyFire classes of compounds in the biological databases. We observe a similar distribution of class levels as for the two biological datasets, indicating that CANOPUS is comprehensively classifying compounds at all possible compound class levels.



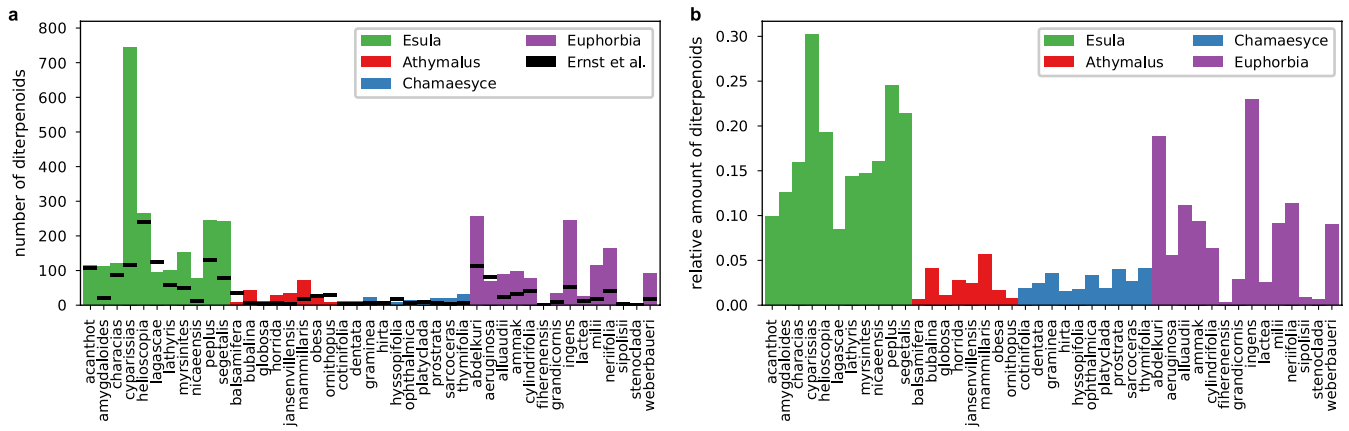
Extended Data Fig. 4 | Molecular network and compound class annotations (single class annotations) for the mice digestive system. Node colors indicate the compound class annotated by CANOPUS; displayed compound classes were manually selected. When a compound is annotated with multiple classes, the class with the larger structural pattern is selected. Nodes are connected by an edge if the spectral similarity is 0.7 or higher.



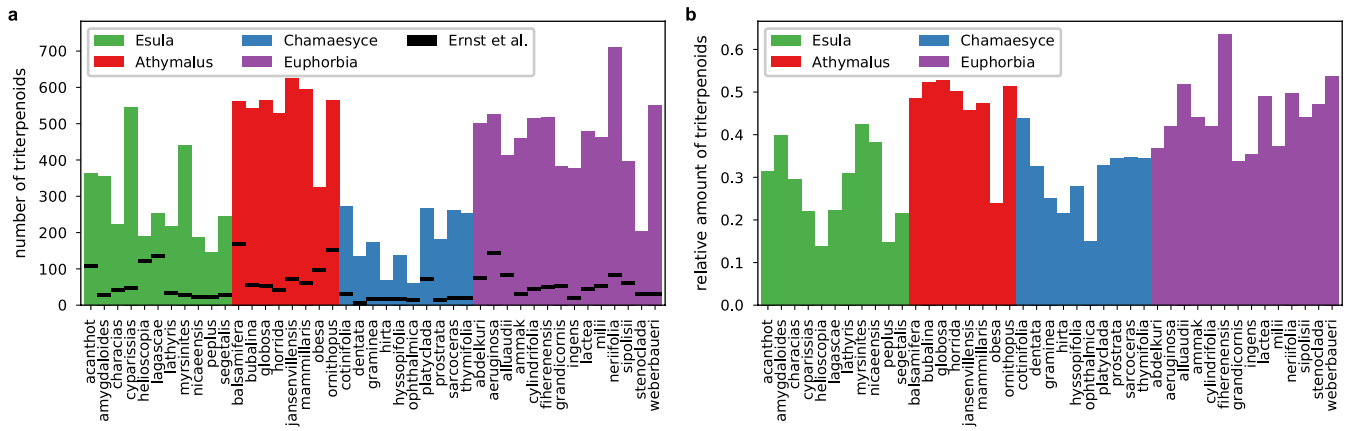
Extended Data Fig. 5 | Molecular network and compound class annotations (muliple class annotations) for the mice digestive system. Node colors indicate the compound class annotated by CANOPUS; compound classes are the same as in Supplementary Fig. 4 1. Compounds belonging to multiple classes displayed as multicolored nodes. Nodes are connected by an edge if the spectral similarity is 0.7 or higher.



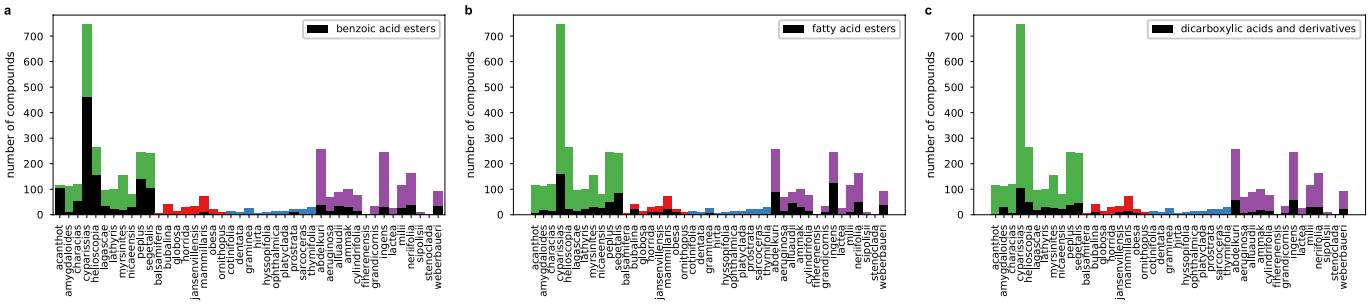
Extended Data Fig. 6 | Number of compounds detected for each Euphorbia subgenus. Orange bars indicate the number of compounds detected here, black ticks indicate the number of compounds reported in the original study. Higher numbers of detected features are not a measure of quality for the two methods, but depend mainly on the preprocessing executed before compound classification.



Extended Data Fig. 7 | Number of compounds annotated as diterpenoids in different species of *Euphorbia*. Left: absolute number of compounds. Right: relative number of compounds, that is, number of diterpenoids divided by total number of compounds in each species. Black ticks in the left figure mark the reported number of diterpenoids in the original study by Ernst et al.



Extended Data Fig. 8 | Number of compounds annotated as triterpenoids in different species of *Euphorbia*. Left: absolute number of compounds. Right: relative number of compounds, that is, number of triterpenoids divided by total number of compounds in each species. Black ticks in the left figure mark the reported number of triterpenoids in the original study by Ernst et al.



Extended Data Fig. 9 | Number of diterpenoids in different species of *Euphorbia*. Black bars show the amount of diterpenoids that have a benzoic acid ester (**a**), fatty acid ester (**b**) or two carboxylic acids (**c**).

nature research

Corresponding author(s): Sebastian Böcker

Last updated by author(s): Apr 23, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used.
Data analysis	*Provide a description of all commercial, open source and custom code used to analyse the data in this study, specifying the version used OR state that no software was used.*

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Provide your data availability statement here.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	N/A
Data exclusions	N/A
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Study did not involve laboratory animals
Wild animals	Study did not involve wild animals
Field-collected samples	Cyanobacterial Collection and Taxonomy: The marine cyanobacterium Rivularia sp. (voucher specimen available from W. H. G. as collection no. CUR6APR19-1) was found growing in 0.5 – 2.0 m of water at Carlos Rosario Beach in Culebra, Puerto Rico, U.S. The sample was hand collected on April 6th 2019, preserved in a 1:1 2-propanol – seawater solution, and stored in the laboratory at – 20 ° C until extraction. Microscopic examination indicated that this collection was morphologically consistent with the genus Rivularia and 16S rDNA analysis indicated that this collection was morphologically consistent with the genus confirmed the identity as Rivularia spp. PCC 7116.
Ethics oversight	No ethical guidance was required, since no animals were involved

Note that full information on the approval of the study protocol must also be provided in the manuscript.