New Phytologist

Methods

Quantifying chemodiversity considering biochemical and structural properties of compounds with the R package CHEMODIV

Hampus Petrén[1], Tobias G. Köllner[2] and Robert R. Junker[1,3]

[1]Evolutionary Ecology of Plants, Department of Biology, Philipps-University Marburg, 35043 Marburg, Germany; [2]Department of Natural Product Biosynthesis, Max Planck Institute for Chemical Ecology, 07745 Jena, Germany; [3]Department of Environment and Biodiversity, University of Salzburg, 5020 Salzburg, Austria

Author for correspondence:
Hampus Petrén
Email: hampus.petren@biologie.uni-marburg.de

Received: *9 June 2022*
Accepted: *11 December 2022*

New Phytologist (2023) **237:** 2478–2492
doi: 10.1111/nph.18685

Key words: chemical communication, chemical ecology, chemodiversity, phytochemicals, plant defence, R package, secondary metabolites.

Summary

- Plants produce large numbers of phytochemical compounds affecting plant physiology and interactions with their biotic and abiotic environment. Recently, chemodiversity has attracted considerable attention as an ecologically and evolutionary meaningful way to characterize the phenotype of a mixture of phytochemical compounds.
- Currently used measures of phytochemical diversity, and related measures of phytochemical dissimilarity, generally do not take structural or biosynthetic properties of compounds into account. Such properties can be indicative of the compounds' function and inform about their biosynthetic (in)dependence, and should therefore be included in calculations of these measures.
- We introduce the R package CHEMODIV, which retrieves biochemical and structural properties of compounds from databases and provides functions for calculating and visualizing chemical diversity and dissimilarity for phytochemicals and other types of compounds. Our package enables calculations of diversity that takes the richness, relative abundance and – most importantly – structural and/or biosynthetic dissimilarity of compounds into account. We illustrate the use of the package with examples on simulated and real datasets.
- By providing the R package CHEMODIV for quantifying multiple aspects of chemodiversity, we hope to facilitate investigations of how chemodiversity varies across levels of biological organization, and its importance for the ecology and evolution of plants and other organisms.

Introduction

Plants produce an astonishing diversity of phytochemical compounds (Kessler & Kalske, 2018; Wang *et al.*, 2019). With functions such as chemical defence, attractant or repellent signalling and protection against abiotic stressors, phytochemicals (also referred to as secondary metabolites) are crucial for mediating mutualistic and antagonistic interactions between plants and other organisms and the abiotic environment (Hartmann, 2007; Junker & Tholl, 2013; Kessler & Kalske, 2018; Whitehead *et al.*, 2021b). Understanding the evolutionary processes generating this phytochemical diversity and the ecological functions of it is the central goal in the field of chemical ecology (Fraenkel, 1959; Ehrlich & Raven, 1964; Hartmann, 2007; Raguso *et al.*, 2015).

Traditionally, research has mostly focused on understanding the function (e.g. herbivore protection or pollinator attraction) of individual phytochemical compounds (Richards *et al.*, 2016; Dyer *et al.*, 2018). However, phytochemicals occur in multicompound mixtures, the composition of which represents a complex phenotype that may vary along multiple dimensions (Marion *et al.*, 2015). Recently, the concept of chemodiversity has received increased attention as a way to quantify this phenotype (Junker *et al.*, 2018; Müller *et al.*, 2020; Wetzel & Whitehead, 2020). Multiple studies have found that function may depend on a diverse mixture of compounds (e.g. Bruce *et al.*, 2005; Iason *et al.*, 2005; Richards *et al.*, 2015; Junker *et al.*, 2018; Tewes *et al.*, 2018; Cosmo *et al.*, 2021; Whitehead *et al.*, 2021a). Less appreciated is the fact that phytochemical compounds are produced by a limited number of biosynthetic pathways and are characterized by different chemical structures (Wink, 2010; Wang *et al.*, 2019). Considering such properties of compounds as a part of the phytochemical phenotype can be important to account for interdependences due to shared biosynthetic pathways (Junker, 2018; Junker *et al.*, 2018), and crucially, a factor contributing to explaining the function of phytochemicals (Wetzel & Whitehead, 2020; Cosmo *et al.*, 2021).

Chemodiversity is often measured using diversity indices (Doyle, 2009; Hilker, 2014; Marion *et al.*, 2015; Kessler & Kalske, 2018; Wetzel & Whitehead, 2020), such as Shannon's diversity index. Numerous studies have used such indices to quantify phytochemical diversity at different levels of biological organization and explored its effects on ecological interactions and evolutionary processes. This includes examples where phytochemical diversity influences insect performance (Tewes *et al.*, 2018; Glassmire *et al.*, 2020), shapes patterns of herbivory and insect diversity (Richards *et al.*, 2015; Salazar *et al.*, 2016), and where it evolves over time in different plant genera (Becerra *et al.*, 2009; Cacho *et al.*, 2015; Volf *et al.*, 2018). Mechanistically, a high diversity of compounds might be selected for and enhance function in different ways (Wetzel & Whitehead, 2020). Synergistic effects may cause the effect of a mixture of compounds to be larger than the sum of the effects of individual compounds (Richards *et al.*, 2016). Alternatively, a diverse set of phytochemicals may result from the multitude of interactions plants experience, each imposing selection on different compounds with different functions (Berenbaum & Zangerl, 1996; Iason *et al.*, 2011; Junker, 2016). Regardless of mechanism, under each scenario, an increased diversity of phytochemical compounds within a plant may increase its fitness.

Using indices such as Shannon's diversity, most studies consider compound richness and evenness, but ignore disparity, the third component of diversity (Daly *et al.*, 2018). Analogous to measures of functional diversity, where species' traits are included in calculations of indices such as Rao's quadratic entropy index (Petchey & Gaston, 2006), the biosynthetic and/or structural disparity of phytochemicals (hereafter referred to as compound dissimilarity) can and should be included in calculations of phytochemical diversity (Bakhtiari *et al.*, 2021). All else equal, a phytochemical mixture of structurally dissimilar compounds produced by different biosynthetic pathways is arguably more diverse than a mixture of less dissimilar compounds from a single biosynthetic pathway. How dissimilar the compounds in a phytochemical mixture are, is thus a crucial component of the mixture's overall diversity. A higher structural diversity among compounds might mediate interactions with or increase effects against a broader set of interacting organisms, or influence synergies between compounds, thereby affecting function and shaping ecological interactions (Becerra *et al.*, 2009; Richards *et al.*, 2015; Sedio, 2017; Glassmire *et al.*, 2019; Sedio *et al.*, 2020; Cosmo *et al.*, 2021; Whitehead *et al.*, 2021a; Philbin *et al.*, 2022). Quantifications of compound dissimilarity based on tandem (MS/MS) mass spectra (Wang *et al.*, 2016) have been used in metabolomics (Tripathi *et al.*, 2021) and ecology (Sedio *et al.*, 2017) to calculate sample dissimilarities and construct molecular networks. In addition, Richards *et al.* (2015) pioneered quantifying phytochemical diversity using ^{1}H-NMR spectra with a measure reflective of both inter-molecular and intra-molecular diversity. Phytochemical data, however, are often analysed using standard GC–MS or LC–MS methods, where individual compounds are identified and quantified. We propose using similar methods to quantify compound dissimilarity for such datasets. By quantifying compound dissimilarities for datasets with identified compounds (Box 2), and calculating phytochemical diversity and dissimilarity of samples using measures of functional Hill diversity and Generalized UniFrac dissimilarities (Chen *et al.*, 2012; Chao *et al.*, 2014) (Box 1), we aim to enable chemical ecologists to quantify all components, including the richness, evenness and disparity, of phytochemical diversity.

We introduce CHEMODIV, a package for analyses of chemodiversity in the statistical software R (R Core Team, 2022). The package allows users, with data on relative abundances of identified phytochemical compounds in different samples, or any other type of chemical composition data, to quantify chemical diversity and dissimilarity in novel ways, where the richness, evenness and, importantly, the biosynthetic and/or structural properties of the compounds are considered. With these new measures, implemented in the R package, we enable researchers to, in a more comprehensive way, test what dimensions of phytochemical diversity are most important in shaping interactions between plants and their biotic and abiotic environment.

Materials and Methods

CHEMODIV is available as an R package on the Comprehensive R Archive Network, CRAN (https://CRAN.R-project.org/package=chemodiv). It contains a number of functions to easily calculate and visualize different types of phytochemical diversity and dissimilarity. In this section, we describe the functions of the package and provide examples of analyses on real and simulated datasets. Details on calculations of diversity and quantification of compound dissimilarity are described in Boxes 1 and 2, respectively, and are jointly summarized in Fig. 1.

Data requirements

Two sets of data are required to fully utilize the functions in the CHEMODIV package. First, a dataset on the relative abundances of phytochemical compounds in different samples is needed, as commonly obtained from GC–MS and LC–MS analyses. Second, a list with the common name, SMILES and InChIKey for all the compounds in the first dataset is needed. SMILES and InChIKey are chemical identifiers and are readily compiled by searching for compounds in chemical databases such as PubChem (S. Kim *et al.*, 2021), or using its automated Identifier Exchange Service tool. These identifiers are used to download data on biosynthetic and structural properties of the phytochemical compounds from different databases.

Description of functions in the R package

The CHEMODIV package functions are summarized in Table 1. A full analysis of the diversity and dissimilarity of a set of phytochemical samples includes a number of largely sequential steps. First, the function *chemoDivCheck* can be used to check that datasets are correctly formatted. Second, the function *NPCTable* enables the use of the *NPClassifier* tool (H. W. Kim *et al.*, 2021) directly within R, to classify compounds into three hierarchical levels largely corresponding to biosynthetic pathways. Third, the function *compDis* uses the list of compounds with their chemical identifiers to generate a dissimilarity matrix with dissimilarities between compounds, calculated based on the biosynthetic classification by *NPClassifier*, the structural properties of the

Box 1 Measures of diversity and dissimilarity

Diversity can be divided into components of richness, evenness and disparity (Daly *et al*., 2018). The most simple diversity measure is simply the richness, in this case the number of phytochemical compounds detected in a sample. Studies on phytochemical diversity often use Shannon's diversity index, calculated as

$$H = -\sum_{i=1}^{S} p_i \log p_i$$

where S is the total number of compounds in the sample and p_i is the relative abundance (proportion) of compound i. This index takes evenness into account, such that for a given number of compounds, diversity is maximized when they occur at equal proportions. For diversity measures also considering disparity, functional diversity indices such as Rao's Q can be used. For phytochemical diversity, Rao's Q measure the average dissimilarity between two randomly drawn compounds, weighted by their abundance, from a sample. It is calculated as

$$Q = \sum_{i=1}^{S} \sum_{j=1}^{S} d_{ij} p_i p_j$$

where p_i and p_j are the relative abundances of compounds i and j, and d_{ij} is the dissimilarity between compounds i and j. In this way, a dissimilarity matrix containing pairwise dissimilarities between phytochemical compounds, calculated based on biosynthetic or structural properties of the molecules (Box 2), can be included in measures of phytochemical diversity.

While these traditional diversity indices are frequently used, a consensus has developed that Hill numbers represent a more suitable way of quantifying diversity (Ellison, 2010). Hill numbers, also referred to as Hill diversity or effective number of species (Hill, 1973; Jost, 2006; Chao *et al*., 2014), are related to the traditional indices, and defined as

$$^{q}D = \left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)}, \quad q \geq 0, \quad q \neq 1.$$

This measure is undefined for $q = 1$, but this can still be calculated because its limit as q approaches 1 equals

$$^{1}D = \lim_{q \to 1} {^{q}D} = \exp\left(-\sum_{i=1}^{S} p_i \log p_i\right).$$

The parameter q is the diversity order and controls the sensitivity of the measure to the relative abundances of the compounds. For $q = 0$, the measure is simply equal to the number of compounds so that $^{0}D = S$. For $q = 1$, compounds are weighed in proportion to their abundance, and ^{1}D is equal to the exponential of Shannon's diversity. For $q > 1$, more weight is put on abundant compounds, and at $q = 2$, ^{2}D is equal to the inverse Simpson diversity. Using Hill numbers to measure diversity has several advantages (Chao *et al*., 2014). First, the parameter q controls the sensitivity of the measure to the relative abundances of compounds. Adjusting q, the behaviour of the index can be controlled to enable a more nuanced measure of diversity. Second, Hill numbers are expressed in units of effective numbers, which is the number of equally abundant compounds required to obtain the same value of diversity. In this way, the units behave intuitively, facilitating comparisons between groups. Third, partitions of Hill numbers into α-, β- and γ-diversity are straightforward (Jost, 2007). Finally, Hill numbers can be generalized to a measure of functional diversity so that compound dissimilarity can also be taken into account (Chao *et al*., 2014; Chiu & Chao, 2014). In this way, it is possible to measure several types of functional diversity in the Hill numbers framework. The most central of these is (total) functional diversity, which can be calculated as

$$^{q}\mathrm{FD}(Q) = \left[\sum_{i=1}^{S} \sum_{j=1}^{S} d_{ij} \left(\frac{p_i p_j}{Q}\right)^{q}\right]^{1/(1-q)}$$

where Q is Rao's Q (Chiu & Chao, 2014). This measure is also undefined for $q = 1$, but its limit as q approaches 1 equals

$$^{1}\mathrm{FD}(Q) = \lim_{q \to 1} {^{q}\mathrm{FD}(Q)} = \exp\left[-\sum_{i=1}^{S} \sum_{j=1}^{S} d_{ij} \left(\frac{p_i p_j}{Q}\right) \log\left(\frac{p_i p_j}{Q}\right)\right].$$

The index $^{q}\mathrm{FD}(Q)$ is a function of all three diversity components. This functional diversity quantifies the effective total dissimilarity between compounds in a sample (Chiu & Chao, 2014). It can therefore be used as a comprehensive measure of phytochemical diversity, sensitive to variation in richness, evenness and disparity. Overall, Hill numbers provide a unified approach to quantifying phytochemical diversity, and to our knowledge, the non-functional version has been used in a few studies (e.g. Marion *et al*., 2015; Cosmo *et al*., 2021; Philbin *et al*., 2022).

Diversity measures combining richness, evenness and disparity into a single metric may obscure independent variation in each component. However, Hill numbers enable separate and combined quantification of all three components. As mentioned, Hill diversity at $q = 0$ simply equals the richness, while at $q = 1$ it is dependent on richness and evenness. Functional Hill diversity adds a layer of data by also considering disparity. At $q = 0$, it is equal to the sum of the pairwise dissimilarities in the dissimilarity matrix, a measure known as functional attribute diversity (Walker *et al*., 1999). At $q = 1$, it is a measure sensitive to all three components of diversity. For a given number of compounds, functional Hill diversity increases with increasing compound dissimilarities, and, in contrast to Rao's Q (Shimatani, 2001), is always maximised at complete evenness. Evenness can also be

Box 1 (Continued).

calculated in this framework (Tuomisto, 2012). Thus, the Hill numbers framework can quantify all components of diversity. Overall, it is crucial to understand the indices' behaviour, and additional ways of calculating diversity exist (Petchey & Gaston, 2006; Chao *et al*., 2019).

Hill numbers measure α-diversity, quantifying the diversity within a single sampling unit. Quantifying differences between samples can be done by calculating β-diversity from measures of α- and γ-diversity (Jost, 2007). Alternatively, and more common in chemical ecology, Bray–Curtis dissimilarities can be calculated between samples, and visualized with a non-metric multidimensional scaling (NMDS) plot (Brückner & Heethoff, 2017). Bray–Curtis dissimilarities measure the compositional dissimilarity between samples, but do not take compound dissimilarity into account. A method to do so was developed by Junker (2018), who calculated a biosynthetically informed dissimilarity measure using Generalized UniFrac dissimilarities (Chen *et al*., 2012). For datasets of compounds with known biosynthetic pathways, compound dissimilarities can be calculated based on the proportion of shared enzymes. These are then incorporated in calculations of sample dissimilarities as Generalized UniFrac dissimilarities such that two samples containing more biosynthetically different compounds have a higher dissimilarity.

Collectively, with Hill numbers and Generalized UniFracs, it is possible to quantify both phytochemical diversity within sampling units and phytochemical dissimilarity between sampling units, in a way that considers compound dissimilarities (Fig. 1). To make such quantifications possible for any phytochemical dataset, a generalized way of quantifying compound dissimilarities is needed (Box 2).

Box 2 Quantifying compound dissimilarities

Calculating functional diversity in the Hill numbers framework, and dissimilarity with Generalized UniFracs, requires a way to quantify dissimilarities between phytochemical compounds. We utilize three different complementary methods to quantify compound dissimilarity that only require knowing compound identities. For a given set of compounds, all pairwise dissimilarities are calculated, and a dissimilarity matrix is then constructed. The first method is based on a hierarchical classification of phytochemicals. H. W. Kim *et al*. (2021) developed a deep-learning tool called *NPClassifier*. This tool automatically classifies natural products into three hierarchical levels: pathway, superclass and class, which at the time of publication consisted of 7, 70 and 672 categories each. Categories are based on expert knowledge and largely correspond to the biosynthetic pathways that synthesize the phytochemicals. As an example, the common volatile linalool is classified in the pathway 'terpenoids', superclass 'monoterpenoids' and class 'acyclic monoterpenoids'. Using a similar approach as in Junker (2018) (see Box 1), we use the classification for each compound to calculate Jaccard dissimilarities between pairs of compounds, as a measure of their biosynthetic dissimilarity. The second method uses molecular fingerprints to quantify compound dissimilarities based on structural properties of the molecules (Cereto-Massagué *et al*., 2015). We use the *PubChem Fingerprint*, which consists of 881 binary variables representing the presence or absence of different features in the molecule, including specific elements, bonds and ring structures (Bolton *et al*., 2008; Cereto-Massagué *et al*., 2015). We chose this specific fingerprint as it is readily acquired from the PubChem database, and has been used for phytochemicals in other studies (e.g. Sorokina *et al*., 2021; Whitehead *et al*., 2021a). The fingerprints are then used to calculate Jaccard dissimilarities between compounds, as a measure of their structural dissimilarity. The third method is a graph-based *flexible Maximum Common Substructure* (*fMCS*) method (Cao *et al*., 2008b; Wang *et al*., 2013). The *fMCS* of two compounds is the largest substructure that occurs in both of them, allowing for a set number atom/bond mismatches in the identified substructures. By comparing the number of atoms in the common substructure to the total number of atoms in the molecules, Jaccard dissimilarities can be calculated based on *fMCS*, as a measure of their structural dissimilarity. Using *fMCS* is more computationally intensive than *PubChem Fingerprints*, but may have increased performance (Wang *et al*., 2013). Using three different methods to quantify compound dissimilarity provides a choice upon which properties (biosynthetic: *NPClassifier*; structural: *PubChem Fingerprints*, *fMCS*) to compare phytochemicals, depending on research questions addressed (see the Results and Discussion section). Data needed for dissimilarity calculations is accessed by the *NPClassifier* tool (H. W. Kim *et al*., 2021), and the PubChem database (S. Kim *et al*., 2021) via functions in the CHEMODIV package.

compounds (*PubChem fingerprints*, *fMCS*; Box 2), and/or a combination of the methods. Fourth, three different functions can be used to calculate different types of diversity for the samples. Function *calcDiv* calculates diversity within samples using the most common indices of α-diversity and evenness, including Shannon's diversity, inverse Simpson diversity, Rao's *Q*, two types of evenness and both types of Hill diversity (Box 1). Functional Hill diversity and Rao's *Q* use the dissimilarity matrix generated by *compDis* in the diversity calculations. Function *calcDivProf* can be used to generate a diversity profile, where both types of Hill diversity are calculated for a range of *q*-values. When plotted, a diversity profile can provide a more nuanced view of the diversity. Function *calcBetaDiv* calculates β-diversity as both types of Hill diversity. Fifth, the function *sampDis* generates a dissimilarity matrix with phytochemical dissimilarities between samples, calculating either Bray–Curtis or Generalized UniFrac dissimilarities, the latter of which uses the compound dissimilarity matrix generated by *compDis*. Sixth, functions *molNet* and *molNetPlot* generate and plot molecular networks, where nodes represent compounds and edges (links) represent similarities between compounds. Such networks can illustrate dissimilarities between compounds, calculated by *compDis*, and simultaneously visualize their abundances. Finally, the function *chemoDivPlot* can be used to conveniently create basic plots of the calculated measures of compound dissimilarity, sample diversity and sample dissimilarity, for different groups of samples that may represent treatments, populations, species or similar. In addition, the function *quickChemoDiv* is a shortcut function that uses the other functions to calculate or visualize phytochemical diversity for a dataset in a single step. The central parts of the workflow are

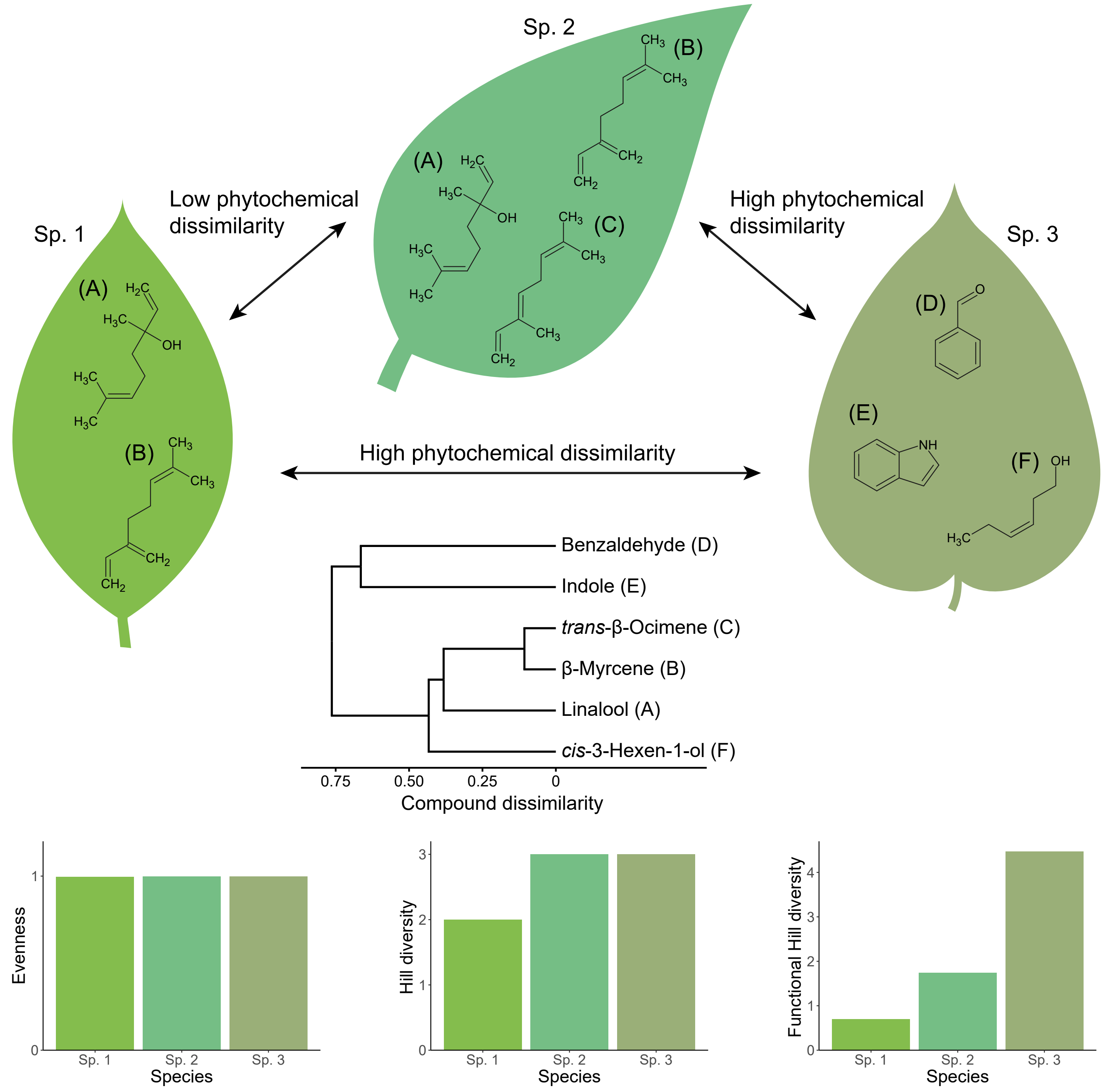

Fig. 1 A conceptual illustration of how phytochemical diversity and dissimilarity can be quantified. Leaves from three different plant species contain different phytochemicals, at equal abundances. Species 1 and 2 contain two and three structurally similar monoterpenes (linalool, β-myrcene, *trans*-β-ocimene), respectively. Species 3 contains three structurally more dissimilar compounds produced in different biosynthetic pathways (indole, an alkaloid; *cis*-3-hexen-1-ol, an aliphatic/fatty acid derivative; benzaldehyde, a benzenoid). The dendrogram illustrates structural dissimilarities between the compounds (calculated using *PubChem Fingerprints*). Species 1 and 2 contain similar compounds, and have a low phytochemical dissimilarity. Species 3 contains different compounds, and has a high phytochemical dissimilarity to the other species. The phytochemical diversity of the species depends on how it is quantified, indicated by the bar plots. All species have equal evenness. Hill diversity is lowest in species 1 because it contains only two compounds. Functional Hill diversity, taking compound dissimilarities into account, is higher in species 3 than in species 2, as an effect of the former having a set of more dissimilar phytochemicals.

shown in Fig. 2. A detailed demonstration of the functions is included in a vignette in the package. Functions in the package do not perform statistical tests, but diversity and dissimilarity calculations produce output in standard formats, enabling subsequent statistical tests by other R functions.

Examples on simulated and real datasets

To demonstrate the applicability of the CHEMODIV package for measuring phytochemical diversity and dissimilarity, we analysed a number of simulated and real datasets with it.

Table 1 Overview of the functions available in the CHEMODIV package, and a description of what they do.

Function call with default parameters	Description
`chemoDivCheck(sampleData, compoundData)`	Checks that datasets are appropriately formatted
`NPCTable(compoundData)`	Classifies phytochemical compounds into a hierarchy of biosynthetic groups using *NPClassifier*
`compDis(compoundData, type = "PubChemFingerprint", npcTable = NULL, unknownCompoundsMean = FALSE)`	Calculates dissimilarities between compounds using biosynthetic (*NPClassifier*) and/or structural (*PubChem Fingerprints*, *fMCS*) properties
`calcDiv(sampleData, compDisMat = NULL, type = "HillDiv", q = 1)`	Calculates selected types of α-diversity and evenness measures, in both traditional and Hill numbers frameworks
`calcDivProf(sampleData, compDisMat = NULL, type = "HillDiv", qMin = 0, qMax = 3, step = 0.1)`	Calculates diversity profiles in the Hill numbers framework
`calcBetaDiv(sampleData, compDisMat = NULL, type = "HillDiv", q = 1)`	Calculates β-diversity in the Hill numbers framework
`sampDis(sampleData, compDisMat = NULL, type = "BrayCurtis", alpha = 1)`	Calculates Bray–Curtis and/or Generalized UniFrac dissimilarities between samples
`molNet(compDisMat, npcTable = NULL, cutOff = "median")`	Creates a molecular network based on compound dissimilarities, and calculates some network properties
`molNetPlot(sampleData, networkObject, groupData = NULL, npcTable = NULL, plotNames = FALSE, layout = "kk")`	Uses the output of *molNet* to create a plot of the molecular network
`chemoDivPlot(compDisMat = NULL, divData = NULL, divProfData = NULL, sampDisMat = NULL, groupData = NULL)`	Creates selected plots to visualize phytochemical diversity, compound dissimilarity and sample dissimilarity
`quickChemoDiv(sampleData, compoundData = NULL, groupData = NULL, outputType = "plots")`	Makes use of other functions in the package to calculate or visualize phytochemical diversity for a dataset in a single step

The first example includes a semi-simulated dataset on plant defence compounds. Glucosinolates is a class of phytochemicals produced by most species in the Brassicaceae family. They provide protection against generalist herbivores (after being hydrolysed by myrosinase), but are less efficient against specialist herbivores (Hopkins *et al*., 2009). Plants in the *Erysimum* (Brassicaceae) genus have additionally gained a novel chemical defence in the form of cardenolides, a class of phytochemicals that provides protection against some glucosinolate-adapted specialist herbivores. A diverse mixture of phytochemicals from both groups could therefore maximise herbivore protection (Züst *et al*., 2018, 2020). However, quantifying such diversity is not straightforward, as, for example, Shannon's diversity do not consider that these two groups of compounds contain structurally different molecules produced in different biosynthetic pathways. To demonstrate the applicability of calculating measures of phytochemical diversity and dissimilarity that take such difference into account, we used a randomly selected set of eight glucosinolates and eight cardenolides that have been found in *Erysimum cheiranthoides* L. in previous studies (Mirzaei *et al*., 2020; Züst *et al*., 2020). Thereafter, by sampling from different normal distributions, we simulated data on the relative concentration of these compounds in three groups of 16 individual plants each: (1) plants with a high concentration of glucosinolates and a low concentration of cardenolides; (2) plants with a low concentration of glucosinolates and a high concentration of cardenolides; and (3) plants with a high concentration of four glucosinolates and four cardenolides, and a low concentration of the remaining glucosinolates and cardenolides. Using the functions in the package, we quantified compound dissimilarity based on the structure of the compounds using *fMCS*, calculated functional Hill diversity and a corresponding diversity profile, calculated sample dissimilarities using Generalized UniFracs, and visualized results. To further demonstrate the applicability of these analyses on real datasets, we additionally compared glucosinolate diversity and dissimilarity with and without taking compound dissimilarities into account for a dataset of 48 *Erysimum* species from Züst *et al*. (2020).

The second example is a fully simulated dataset designed to further examine the use of functional Hill diversity as a measure of phytochemical diversity. This example includes a base dataset with relative concentrations of eight phytochemicals (four at high concentration and four at low concentration) simulated in a similar way as example one. This time, also compound dissimilarities were simulated by sampling from a binomial distribution. We then created additional groups of samples that had an increased richness, evenness and/or compound dissimilarity for a total of eight groups with all combinations of high and low values for the three diversity components. We then calculated and plotted compound richness, evenness and both versions of Hill diversity for the different groups.

The third example is a dataset on floral scent from Larue *et al*. (2016), where floral volatiles were collected from *Achillea millefolium* L. and *Cirsium arvense* L. in a scent manipulation experiment. Here we only include plants in the control treatment. Using the package functions, we calculated and created plots of the compound dissimilarities (based on *PubChem Fingerprints*), functional Hill diversity and sample dissimilarities (using Generalized UniFracs), and created molecular networks to compare the two species.

Finally, we compared the three methods (*NPClassifier*, *PubChem fingerprints* and *fMCS*) for quantifying compound dissimilarities. For this, we used a collection of 2855 phytochemical compounds from the KEGG database (Kanehisa & Goto, 2000). The heaviest 20% of molecules were excluded to reduce computation times.

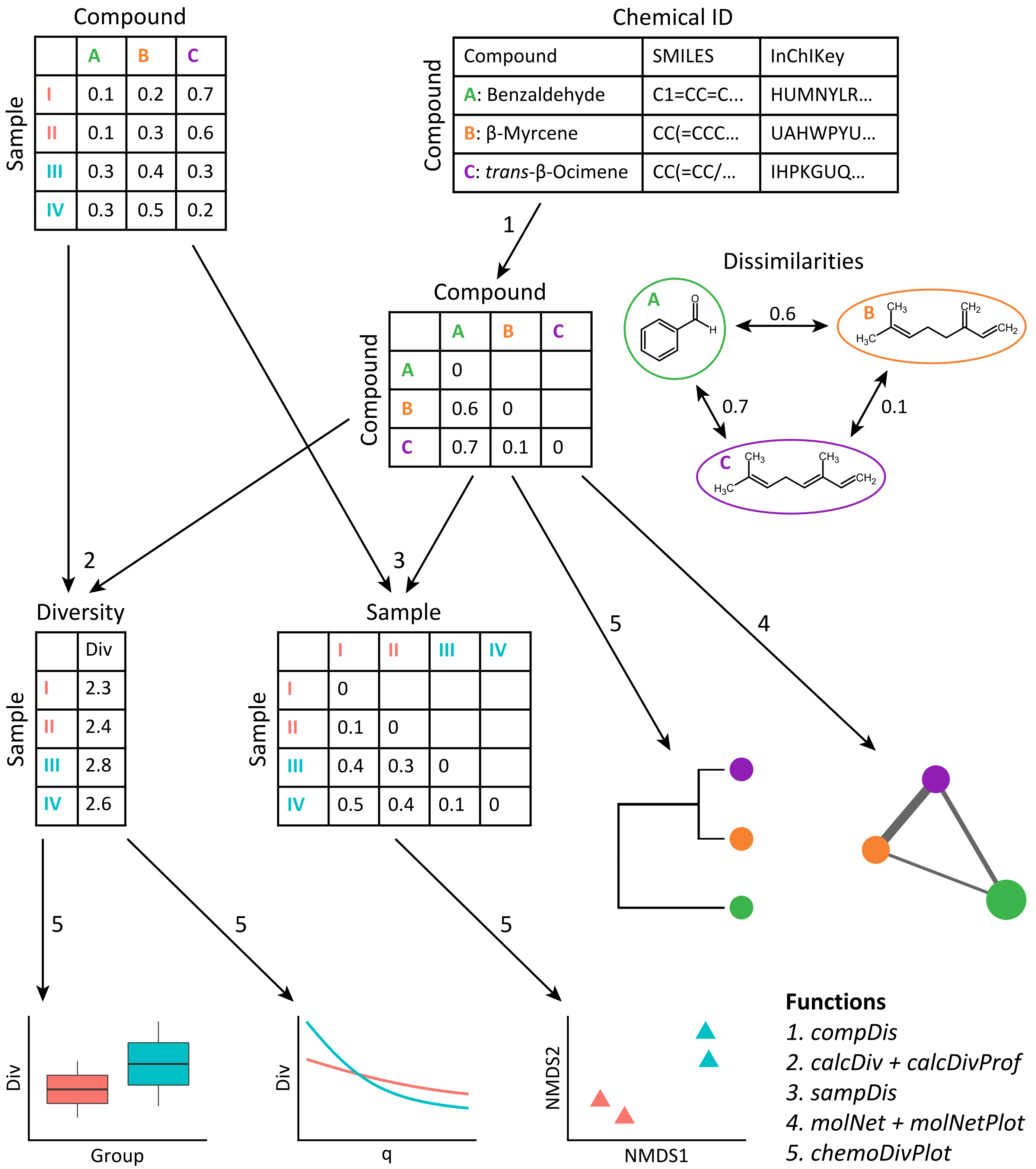

Fig. 2 An illustration of the workflow of the main functions in the CHEMODIV package. A dataset with relative abundances of phytochemical compounds in four samples belonging to two different groups (red and blue), and a list with the common name, SMILES and InChIKey for the compounds in the dataset are required. The *compDis* function uses the list of compounds to generate a dissimilarity matrix with dissimilarities between compounds (1). β-Myrcene (B; orange) and *trans*-β-ocimene (C; purple) are two linear monoterpenes that have a low structural dissimilarity, while benzaldehyde (A; green), a benzenoid, is more dissimilar to the other compounds. In combination with the sample dataset, the compound dissimilarity matrix is used to calculate phytochemical diversity within samples (2; functions *calcDiv* and *calcDivProf*) and phytochemical dissimilarity between samples (3; function *sampDis*). Functions *molNet* and *molNetPlot* are used to create a molecular network (4), while *chemoDivPlot* is used to create multiple plots of compound dissimilarity, sample diversity and sample dissimilarity (5).

Then, from this subset, 20–40 compounds were randomly selected, and dissimilarity matrices were calculated using the *compDis* function with the three different methods. Mantel tests were then used to calculate correlation coefficients between matrices. This was repeated 50 times, and results were plotted to examine how comparable compound dissimilarities generated with the different methods were to each other. With the same dataset, we also examined computation times of the *quickChemoDiv* function, which executes the other main functions in the package. This was done both with and without compound data as a comparison.

Results and Discussion

Evaluating examples on glucosinolates, cardenolides and simulated data

Analyses for the semi-simulated dataset with cardenolides and glucosinolates exemplify how the structural component of phytochemical diversity can be quantified. Compound dissimilarity, quantified using *fMCS*, is low among glucosinolates and among cardenolides, but higher when comparing glucosinolates to cardenolides, as evident by the dendrogram separating the two groups of compounds (Fig. 3a). These differences in compound dissimilarity influence the phytochemical diversity measured as functional Hill diversity (Fig. 3b), which is significantly different between groups (ANOVA, $F_{2,45} = 60.4$, $P < 0.001$). Even without variation in compound richness or evenness, there are clear differences in the diversity of samples from the different groups. Diversity is lowest for the group with high concentration of only glucosinolates, intermediate for the group with high concentration of only cardenolides (due to a somewhat higher average compound dissimilarity among cardenolides than among glucosinolates) and highest for the group containing a high concentration of compounds from both classes. The diversity profile displays functional Hill diversity for $q = 0$–3, varying how much weight is put on low-concentration compounds (Fig. 3c). At

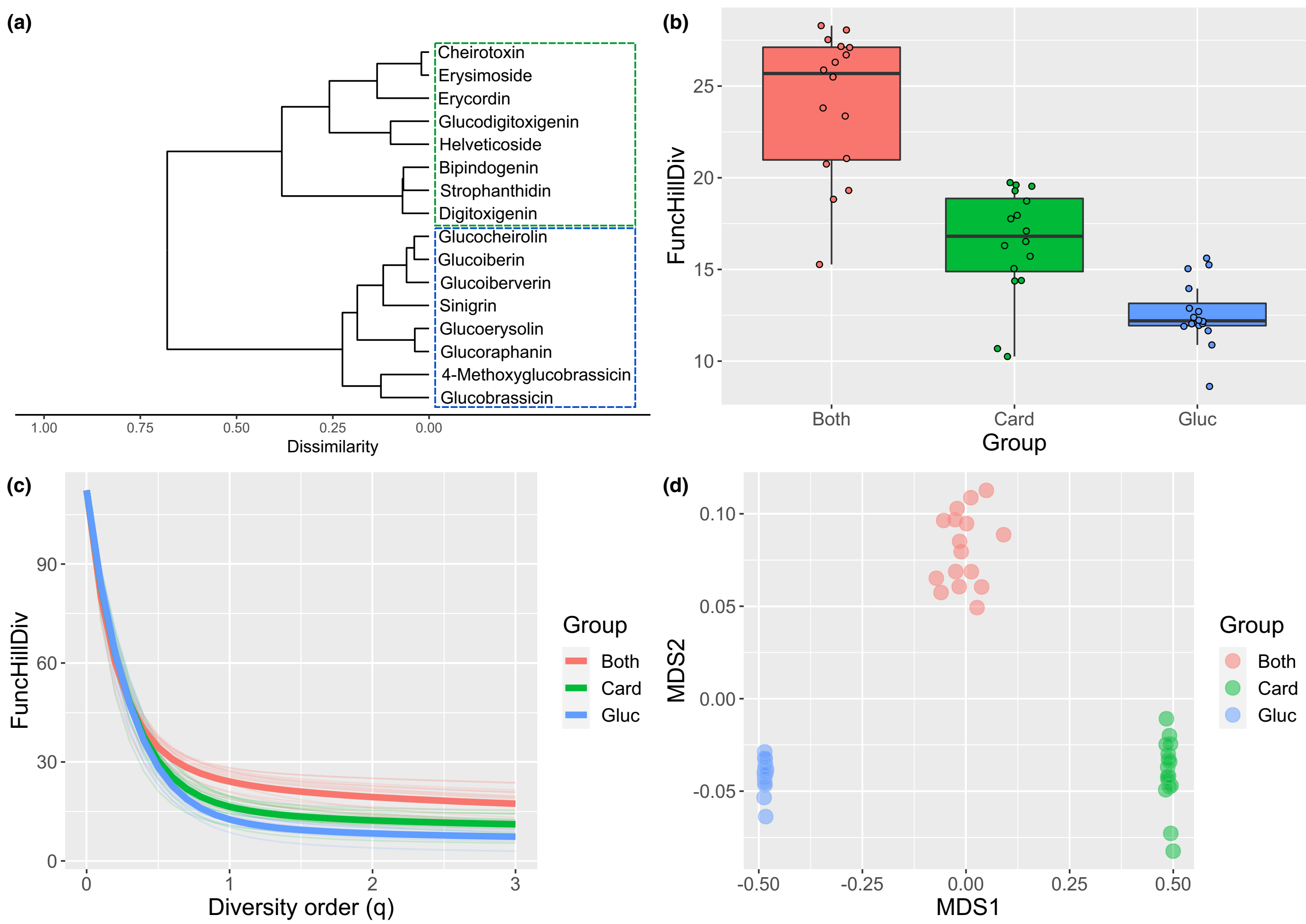

Fig. 3 Phytochemical diversity and dissimilarity for the semi-simulated dataset with glucosinolates and cardenolides, visualized by the *chemoDivPlot* function in the CHEMODIV package. (a) Dendrogram of compound dissimilarities based on *fMCS*, with a clear separation between cardenolides (upper branch) and glucosinolates (lower branch). For clarity, coloured borders have been added around compounds to indicate the two classes (green, cardenolides; blue, glucosinolates). (b) Functional Hill diversity ($q = 1$) for the groups containing a high concentration of cardenolides (Card), a high concentration of glucosinolates (Gluc) and both (Both). Boxes display median values and upper and lower quartiles, with whiskers extending up to 1.5 times the interquartile range. (c) Diversity profile showing the functional Hill diversity for $q = 0$–3. Thick lines represent group means while thin lines represent individual samples. (d) Non-metric multidimensional scaling (NMDS) plot visualizing sample dissimilarities (Generalized UniFracs) between the three groups.

$q = 1$ (also shown in Fig. 3b), equal weight is put on all compounds. At $q > 1$, more weight is put on abundant compounds (the upper limit plotted is set to $q = 3$, as little changes for larger values). At $q = 0$, compound proportions are not taken into account and the functional Hill diversity is thus equal for all three groups. Lastly, the NMDS illustrates Generalized UniFrac dissimilarities of samples (Fig. 3d). Samples cluster in groups, as an effect of being dominated by different compounds. Within-group dispersion is highest for the group containing high concentrations of both types of compounds, as a result of higher average compound dissimilarity. A comparison between the diversities and dissimilarities calculated here, and the traditionally used Shannon's diversity and Bray–Curtis dissimilarities is shown in Supporting Information Fig. S1. Results for glucosinolate diversity and dissimilarity comparisons are presented in Fig. S2.

Results from the second example, with the fully simulated dataset, are summarized in Fig. S3. In short, this example illustrates the behaviour of different diversity measures, and demonstrates the overall suitability of using functional Hill diversity as a measure of phytochemical diversity. By simulating samples with low and high richness, evenness and compound dissimilarity, as expected we found that functional Hill diversity is lowest when all three components have low values, intermediate when some components have high values and other have low values, and highest when all three components have high values.

While many studies have found that phytochemical diversity, measured as, for example, Shannon's diversity, can shape interactions between plants and other organisms (e.g. Iason *et al.*, 2005; Glassmire *et al.*, 2016; Tewes *et al.*, 2018), the structural dimension of phytochemical diversity may also be important for ecological interactions (Richards *et al.*, 2015; Junker *et al.*, 2018; Cosmo *et al.*, 2021). In the example with glucosinolates and cardenolides, the group with structurally dissimilar compounds from two different biosynthetic pathways had the highest diversity when measured as functional Hill diversity. On a general level, two structurally similar molecules can be expected to have a more similar biological activity than two structurally dissimilar molecules (Berenbaum & Zangerl, 1996; Martin *et al.*, 2002). Therefore, a set of structurally dissimilar phytochemicals from different biosynthetic pathways may be more diverse in regard to its function (Philbin *et al.*, 2022), with potential effects on plant fitness. For example, increased structural diversity of phytochemicals in leaves, quantified from ^{1}H-NMR spectra, has been found to decrease herbivory in multiple *Piper* species (Glassmire *et al.*, 2019; Cosmo *et al.*, 2021; Philbin *et al.*, 2022). In addition, Whitehead *et al.* (2021a) found that increasing the structural diversity of phenolics in the diet of eight insect and fungi plant consumers increased the proportion of those consumers negatively affected by the phenolics. However, there are also contrasting examples where similar compounds, for example, different enantiomers, have different function (He *et al.*, 2019). Overall, the molecular and physiological mechanisms by which phytochemicals function are so far often unknown (Richards *et al.*, 2016). However, although the association between dissimilarity in structure and dissimilarity in function for any given pair of phytochemicals may be uncertain, associations between structural diversity and function for multicompound mixtures of phytochemicals may be more prevalent. Structurally diverse mixtures of phytochemicals may have a stronger effect for specific interactions because of synergistic effects between compounds, or be more likely to affect a broader set of interactions (Philbin *et al.*, 2022). By enabling calculations of compound dissimilarity based on molecular structure (*PubChem Fingerprints*, *fMCS*), and subsequent measures of sample diversity or dissimilarity, the CHEMODIV package can help to test hypotheses about how structural diversity of phytochemicals may affect various functions and shape ecological interactions.

If compound dissimilarities are calculated with *NPClassifier*, this may help to account for non-independence of compounds due to shared biosynthetic pathways (Junker, 2018), and inform about biosynthesis differences between sets of compounds. In this respect, a high phytochemical diversity may represent a form of 'biochemical potential', where a plant species producing compounds from several different pathways may more successfully respond to changing selection pressures arising from, for example, new herbivores. Related to this, Becerra *et al.* (2009) found evidence of increasing biosynthetic diversity in the *Bursera* genus over macroevolutionary time, suggesting that production of compounds from different biosynthetic pathways makes it more difficult for herbivores to adapt to new compounds. How such diversity may vary across phylogenies or between plant communities is a potential avenue for future research (Junker, 2018).

While efficient, the *NPClassifier* method has a lower resolution compared to using manually collected data on enzymes (Junker, 2018), because the classification is limited to three hierarchical levels. It should also be noted that structural dissimilarity has been used as a proxy for biosynthetic similarity in other studies (Dowell & Mason, 2020; Cna'ani *et al.*, 2021), and in our simulations dissimilarities calculated with all three methods were correlated (Fig. S4), indicative of an overall consistency between methods. In addition, the structural and biosynthetic similarity of compounds may correlate with similarity of physicochemical properties such as volatility, reactivity and polarity, that may be ecologically important (Rasmann & Agrawal, 2011; Conchou *et al.*, 2019). Researchers should make a deliberate choice of how to quantify compound dissimilarities based on what questions are addressed. Overall, the structural and biosynthetic components of the compounds are important parts of the phytochemical diversity that should be included in measures of it. Using the CHEMODIV package, these components can easily be quantified and incorporated in diversity calculations with the novel use of functional Hill diversity, providing more comprehensive measures of chemodiversity.

Evaluating examples on floral volatiles

Analyses of the *A. millefolium* and *C. arvense* dataset indicate that phytochemical diversity of the floral scent bouquet was higher in the latter species (Fig. 4b,c; ANOVA, $F_{1,16} = 16.9$, $P = 0.001$), mainly due to a higher average number of compounds (*A. millefolium* = 36.3, *C. arvense* = 48.4). The floral scent composition was also clearly different between species (Fig. 4d). Illustrations

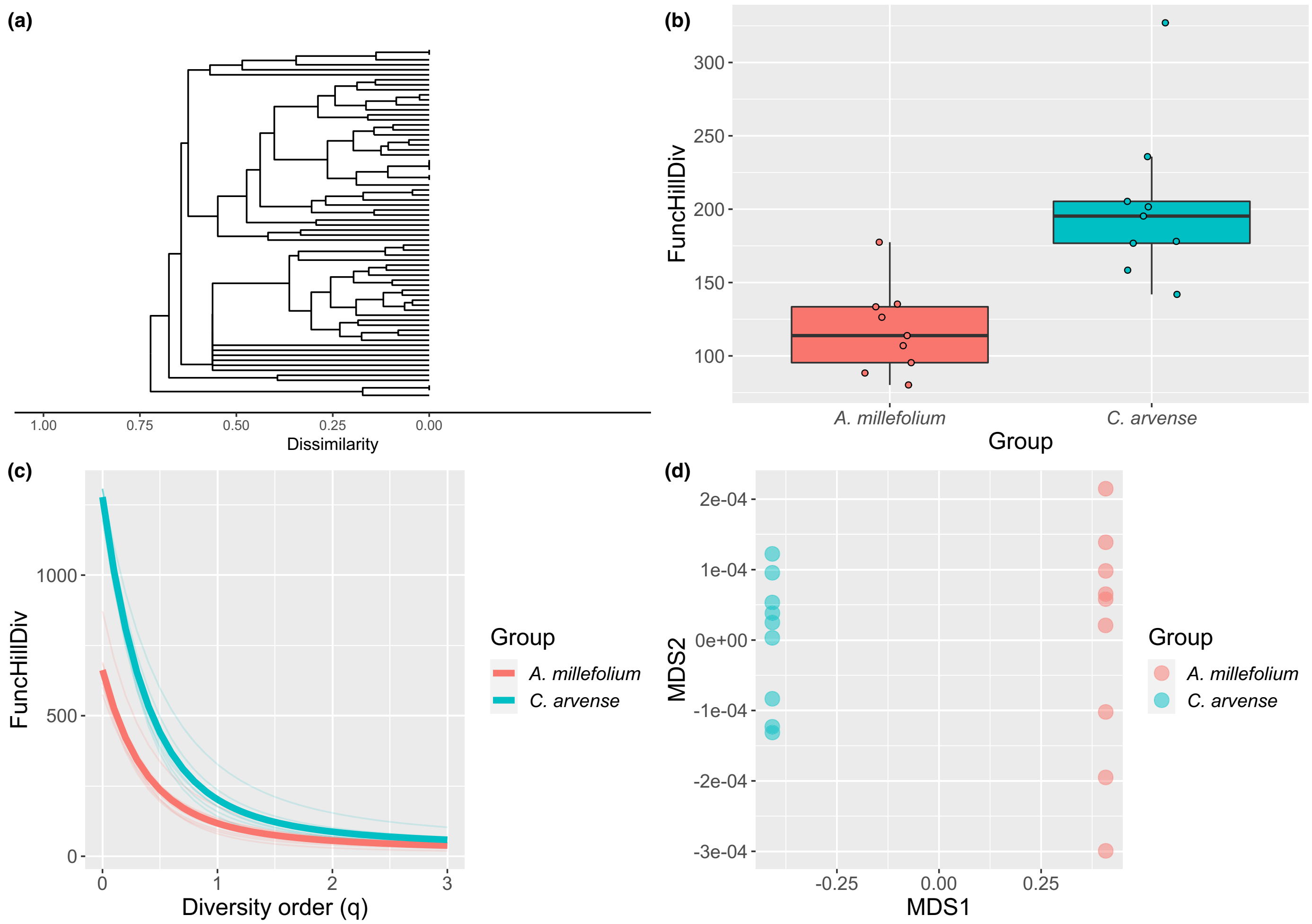

Fig. 4 Phytochemical diversity and dissimilarity for *Achillea millefolium* and *Cirsium arvense* ($n = 9$ for both species) floral scent, visualized by the *chemoDivPlot* function in the CHEMODIV package. (a) Dendrogram of compound dissimilarities based on *PubChem Fingerprints* (compound names have been excluded for clarity; a version with these included is presented in Supporting Information Fig. S5). (b) Functional Hill diversity ($q = 1$) for the two species. Boxes display median values and upper and lower quartiles, with whiskers extending up to 1.5 times the interquartile range. (c) Diversity profile showing the functional Hill diversity for $q = 0$–3. Thick lines represent species means while thin lines represent individual samples. (d) Non-metric multidimensional scaling (NMDS) plot visualizing sample dissimilarities (Generalized UniFracs) between the species.

of compound similarities by the molecular networks (Fig. 5) indicate the presence of two main clusters of structurally similar compounds mainly consisting of the pathways 'Terpenoids' and 'Shikimates and Phenylpropanoids', respectively. The scent bouquet of *A. millefolium* plants was dominated by compounds from the first group (Fig. 5a), while the scent bouquet of *C. arvense* plants was dominated by compounds from the second group (Fig. 5b).

Most examples on the effect of phytochemical diversity on ecological interactions regard herbivores, where the diversity represents a complex phenotype important for herbivore defence through toxic effects of compounds during consumption (Marion *et al.*, 2015; Kessler & Kalske, 2018). By contrast, for a pollinator in search of nectar, or a herbivore searching for host plants, phytochemical diversity, in the form of volatile organic compounds (VOCs), represents information in a complex environment (Kessler, 2015; O'Connor *et al.*, 2019). With potential correlations between compound properties and neural/behavioural response (Khan *et al.*, 2007; Haddad *et al.*, 2008; but see Knaden *et al.*, 2012), a structurally diverse set of VOCs may be functionally diverse, and may therefore enable a generalist plant to efficiently attract different pollinators and/or simultaneously repel antagonistic insects (Junker & Blüthgen, 2010; Schiestl, 2010; Gershenzon *et al.*, 2012; Junker, 2016). In other cases, diverse mixtures of leaf VOCs can make it difficult for herbivores to locate suitable host plants (Zu *et al.*, 2020). Using comprehensive measures of diversity may enable a better understanding of its effects on both antagonistic and mutualistic interactions between plants and other organisms.

Applicability and caveats

The CHEMODIV package allows users to comprehensively analyse phytochemical diversity. To do so, it utilizes other packages for retrieving and processing chemical data, including WEBCHEM (Szöcs *et al.*, 2020), CHEMMINER (Cao *et al.*, 2008a) and FMCSR

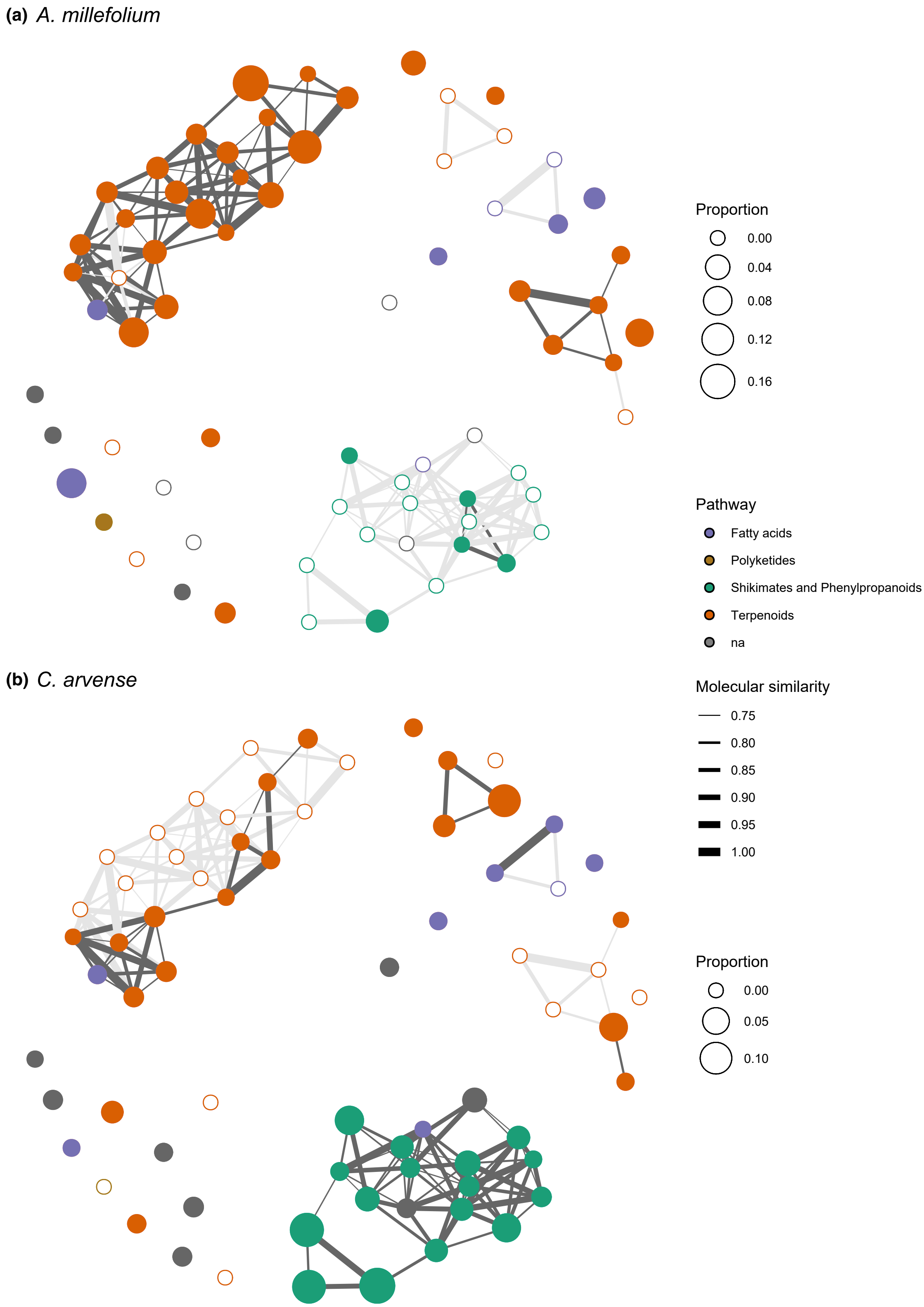

Fig. 5 Molecular networks of the compounds found in *Achillea millefolium* (a) and *Cirsium arvense* (b), visualized by the *molNetPlot* function in the CHEMODIV package. Edge width represents similarities between compounds. Only edges with a similarity ≥ 0.75 are plotted. This cut-off value influences the structure of the network, and was manually specified to separate compounds from the main different pathways. 'na' indicates that the compound could not be classified. Node colour represents the pathway classification from *NPClassifier*, indicating to which major biosynthetic group compounds belong. These are identical in (a, b). Node size represents proportional concentration (mean values for each species), and differs between (a) and (b). Note that nodes with white fill represent zero values, that is, compounds not present in that species, and that edges connecting to such nodes are a lighter shade of grey.

(Wang *et al.*, 2013); and for diversity and dissimilarity calculations, including VEGAN (Oksanen *et al.*, 2022), HILLR (Li, 2018) and GUNIFRAC (Chen *et al.*, 2022). These and similar packages include some of the functionalities in CHEMODIV, but no other package contains functions both for processing chemical data and analysing diversity. Including this functionality in a single package enables chemical ecologists to analyse their data in a comprehensive way and to straightforwardly examine patterns of chemodiversity in a wide range of datasets.

A few aspects should be considered when using the CHEMODIV package and quantifying diversity. The package requires that compounds have been identified to quantify compound dissimilarity. The functions can handle unknown compounds, and will inform the user in cases of missing data and similar issues. Still, CHEMODIV is of most use for datasets where most compounds are identified. If that is not the case, diversity can instead be calculated with indices not considering compound dissimilarities. Alternatively, other methods to quantify the dissimilarity of unknown compounds can be used. For example, in metabolomic analyses where individual compounds are not identified, the GNPS ecosystem provides methods for calculating cosine similarities between MS/MS spectra (Wang *et al.*, 2016), which could be used in diversity calculations. There are no strict limitations on the size of datasets that can be analysed, and computation times are overall reasonable (Fig. S6).

Even if chemodiversity is an ecologically relevant measure, other aspects of the phytochemical phenotype are undoubtedly also important. First, functions such as pollinator attraction, herbivore defence or plant–bacteria interactions often depend on individual compounds rather than mixtures (Junker, 2016; Zhou *et al.*, 2017; Burdon *et al.*, 2018). However, such examples often regard the effect of a compound on a single interacting species, while in nature, plants simultaneously interact with numerous mutualists and antagonists. Each of these might select for the occurrence of different compounds, resulting in evolution of increased chemical diversity (Berenbaum & Zangerl, 1996; Iason *et al.*, 2011; Whitehead *et al.*, 2021a). Second, an important aspect of the phytochemical phenotype is total abundance, which of course may be important for function. Diversity indices consider relative values, such that two samples with the same compounds in identical proportions will have identical diversity, even if they differ in absolute concentrations. Related to this, the number of compounds detected in a sample may partly depend on total abundance (Wetzel & Whitehead, 2020). Therefore, direct comparisons of phytochemical diversity should ideally be done for samples collected with identical methods. If this is not the case and total abundances vary substantially among samples, diversity can be quantified using Hill numbers at higher diversity orders (e.g. $q = 2$). Doing so, less weight is put on low concentration compounds, which may be less functionally important, decreasing any potential influence of differences in total abundance on measures of diversity.

Notably, quantification of molecular structural dissimilarity is not limited to phytochemicals. Diversity of chemical mixtures has been quantified for, for example, fungi VOCs (Guo *et al.*, 2021), snake venom (Holding *et al.*, 2021), coralline algae metabolites (Jorissen, 2021) and fish fatty acids (Feiner *et al.*, 2018). Instead of using traditional measures, such studies can also include the structural dissimilarities of compounds for more comprehensive measures of chemodiversity. It is also worth mentioning that we have focused on quantifying diversity on the level of individual samples, often likely to represent individual plants. Others have instead utilized phytochemical compounds as functional traits to measure diversity on a community level (Salazar *et al.*, 2016), representing a complementary approach which can answer questions related to trait-based ecology and niche processes (Müller & Junker, 2022; Walker *et al.*, 2022).

Conclusions

Plants produce a remarkable number of phytochemicals. By now, it is widely accepted that, rather than being metabolic waste products, they are functionally important, and their diversity is the result of adaptive processes (Hartmann, 2007). However, much is still unknown about how the complex phenotype that is the composition of phytochemical compounds affects interactions between plants and other organisms (Wetzel & Whitehead, 2020). We believe the diversity of the compounds, including their structural and biosynthetic properties, to be a potentially important dimension of this variation that deserves further attention. The CHEMODIV package provides an easy yet comprehensive way to quantify this diversity for many types of data collected by chemical ecologists. By providing this tool, we hope to give researchers the opportunity to more efficiently test in what ways phytochemical variation influences ecological interactions and evolutionary processes, which should increase our understanding of the vast diversity of phytochemical compounds found in plants.

Acknowledgements

We thank the members of the Research Unit FOR3000 ‘Ecology and Evolution of Intraspecific Chemodiversity of Plants’ for fruitful discussions on the topic of chemodiversity, and members of the Junker lab for helpful comments on the manuscript. This project was conducted within the Research Unit FOR3000 funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, JU 2856/5-1). Open Access funding enabled and organized by ProjektDEAL.

Competing interests

None declared.

Author contributions

All authors conceived the study and contributed to study design. HP evaluated statistical methods and created the R package with input from RRJ. HP wrote the manuscript with contributions from RRJ and TGK. All authors approved of the final version of the manuscript.

ORCID

Robert R. Junker https://orcid.org/0000-0002-7919-9678
Tobias G. Köllner https://orcid.org/0000-0002-7037-904X
Hampus Petrén https://orcid.org/0000-0001-6490-4517

Data availability

The R package is available on CRAN (https://CRAN.R-project.org/package=chemodiv) and developed on GitHub (https://github.com/hpetren/chemodiv). Scripts and data of the examples in the paper are available in Dataset S1 and Notes S1.

References

Bakhtiari M, Glauser G, Defossez E, Rasmann S. 2021. Ecological convergence of secondary phytochemicals along elevational gradients. *New Phytologist* **229**: 1755–1767.

Becerra JX, Noge K, Venable DL. 2009. Macroevolutionary chemical escalation in an ancient plant–herbivore arms race. *Proceedings of the National Academy of Sciences, USA* **106**: 18062–18066.

Berenbaum MR, Zangerl AR. 1996. Phytochemical diversity. In: Romeo JT, Saunders JA, Barbosa P, eds. *Phytochemical diversity and redundancy in ecological interactions*. Boston, MA, USA: Springer, 1–24.

Bolton EE, Wang Y, Thiessen PA, Bryant SH. 2008. PubChem: integrated platform of small molecules and biological activities. In: *Annual reports in computational chemistry, vol. 4*. Amsterdam, Netherlands: Elsevier, 217–241.

Bruce TJA, Wadhams LJ, Woodcock CM. 2005. Insect host location: a volatile situation. *Trends in Plant Science* **10**: 269–274.

Brückner A, Heethoff M. 2017. A chemo-ecologists' practical guide to compositional data analysis. *Chemoecology* **27**: 33–46.

Burdon RCF, Junker RR, Scofield DG, Parachnowitsch AL. 2018. Bacteria colonising *Penstemon digitalis* show volatile and tissue-specific responses to a natural concentration range of the floral volatile linalool. *Chemoecology* **28**: 11–19.

Cacho NI, Kliebenstein DJ, Strauss SY. 2015. Macroevolutionary patterns of glucosinolate defense and tests of defense-escalation and resource availability hypotheses. *New Phytologist* **208**: 915–927.

Cao Y, Charisi A, Cheng L-C, Jiang T, Girke T. 2008a. ChemmineR: a compound mining framework for R. *Bioinformatics* **24**: 1733–1734.

Cao Y, Jiang T, Girke T. 2008b. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* **24**: i366–i374.

Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. 2015. Molecular fingerprint similarity search in virtual screening. *Methods* **71**: 58–63.

Chao A, Chiu C, Villéger S, Sun I, Thorn S, Lin Y, Chiang J, Sherwin WB. 2019. An attribute-diversity approach to functional diversity, functional beta diversity, and related (dis)similarity measures. *Ecological Monographs* **89**: e01343.

Chao A, Chiu C-H, Jost L. 2014. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through hill numbers. *Annual Review of Ecology, Evolution, and Systematics* **45**: 297–324.

Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. 2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. *Bioinformatics* **28**: 2106–2113.

Chen J, Zhang X, Zhou H. 2022. *GUniFrac: generalized UniFrac distances, distance-based multivariate methods and feature-based univariate methods for microbiome data analysis.* R package v.1.6. [WWW document] URL https://CRAN.R-project.org/package=GUniFrac. [accessed 15 December 2022].

Chiu C-H, Chao A. 2014. Distance-based functional diversity measures and their decomposition: a framework based on hill numbers. *PLoS ONE* **9**: e100014.

Cna'ani A, Dener E, Ben-Zeev E, Günther J, Köllner TG, Tzin V, Seifan M. 2021. Phylogeny and abiotic conditions shape the diel floral emission patterns of desert Brassicaceae species. *Plant, Cell & Environment* **44**: 2656–2671.

Conchou L, Lucas P, Meslin C, Proffit M, Staudt M, Renou M. 2019. Insect odorscapes: from plant volatiles to natural olfactory scenes. *Frontiers in Physiology* **10**: 972.

Cosmo LG, Yamaguchi LF, Felix GMF, Kato MJ, Cogni R, Pareja M. 2021. From the leaf to the community: distinct dimensions of phytochemical diversity shape insect–plant interactions within and among individual plants. *Journal of Ecology* **109**: 2475–2487.

Daly A, Baetens J, De Baets B. 2018. Ecological diversity: measuring the unmeasurable. *Mathematics* **6**: 119.

Dowell JA, Mason CM. 2020. Correlation in plant volatile metabolites: physiochemical properties as a proxy for enzymatic pathways and an alternative metric of biosynthetic constraint. *Chemoecology* **30**: 327–338.

Doyle L. 2009. Quantification of information in a one-way plant-to-animal communication system. *Entropy* **11**: 431–442.

Dyer LA, Philbin CS, Ochsenrider KM, Richards LA, Massad TJ, Smilanich AM, Forister ML, Parchman TL, Galland LM, Hurtado PJ *et al.* 2018. Modern approaches to study plant–insect interactions in chemical ecology. *Nature Reviews Chemistry* **2**: 50–64.

Ehrlich PR, Raven PH. 1964. Butterflies and plants: a study in coevolution. *Evolution* **18**: 586–608.

Ellison AM. 2010. Partitioning diversity. *Ecology* **91**: 1962–1963.

Feiner ZS, Swihart RK, Coulter DP, Höök TO. 2018. Fatty acids in an iteroparous fish: variable complexity, identity, and phenotypic correlates. *Canadian Journal of Zoology* **96**: 859–868.

Fraenkel GS. 1959. The Raison d'Être of secondary plant substances. *Science* **129**: 1466–1470.

Gershenzon J, Fontana A, Burow M, Wittstock U, Degenhardt J. 2012. Mixtures of plant secondary metabolites: metabolic origins and ecological benefits. In: Iason GR, Dicke M, Hartley SE, eds. *The ecology of plant secondary metabolites.* Cambridge, UK: Cambridge University Press, 56–77.

Glassmire AE, Jeffrey CS, Forister ML, Parchman TL, Nice CC, Jahner JP, Wilson JS, Walla TR, Richards LA, Smilanich AM *et al.* 2016. Intraspecific phytochemical variation shapes community and population structure for specialist caterpillars. *New Phytologist* **212**: 208–219.

Glassmire AE, Philbin C, Richards LA, Jeffrey CS, Snook JS, Dyer LA. 2019. Proximity to canopy mediates changes in the defensive chemistry and herbivore loads of an understory tropical shrub, *Piper kelleyi*. *Ecology Letters* **22**: 332–341.

Glassmire AE, Zehr LN, Wetzel WC. 2020. Disentangling dimensions of phytochemical diversity: alpha and beta have contrasting effects on an insect herbivore. *Ecology* **101**: e03158.

Guo Y, Jud W, Weikl F, Ghirardo A, Junker RR, Polle A, Benz JP, Pritsch K, Schnitzler J-P, Rosenkranz M. 2021. Volatile organic compound patterns predict fungal trophic mode and lifestyle. *Communications Biology* **4**: 673.

Haddad R, Khan R, Takahashi YK, Mori K, Harel D, Sobel N. 2008. A metric for odorant comparison. *Nature Methods* **5**: 425–429.

Hartmann T. 2007. From waste products to ecochemicals: fifty years research of plant secondary metabolism. *Phytochemistry* **68**: 2831–2846.

He J, Fandino RA, Halitschke R, Luck K, Köllner TG, Murdock MH, Ray R, Gase K, Knaden M, Baldwin IT *et al.* 2019. An unbiased approach elucidates variation in (*S*)-(+)-linalool, a context-specific mediator of a tri-trophic interaction in wild tobacco. *Proceedings of the National Academy of Sciences, USA* **116**: 14651–14660.

Hilker M. 2014. New synthesis: parallels between biodiversity and chemodiversity. *Journal of Chemical Ecology* **40**: 225–226.

Hill MO. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54**: 427–432.

Holding ML, Strickland JL, Rautsaw RM, Hofmann EP, Mason AJ, Hogan MP, Nystrom GS, Ellsworth SA, Colston TJ, Borja M *et al.* 2021. Phylogenetically diverse diets favor more complex venoms in North American pitvipers. *Proceedings of the National Academy of Sciences, USA* **118**: e2015579118.

Hopkins RJ, van Dam NM, van Loon JJA. 2009. Role of glucosinolates in insect–plant relationships and multitrophic interactions. *Annual Review of Entomology* **54**: 57–83.

Iason GR, Lennon JJ, Pakeman RJ, Thoss V, Beaton JK, Sim DA, Elston DA. 2005. Does chemical composition of individual Scots pine trees determine the biodiversity of their associated ground vegetation? *Ecology Letters* **8**: 364–369.

Iason GR, O'Reilly-Wapstra JM, Brewer MJ, Summers RW, Moore BD. 2011. Do multiple herbivores maintain chemical diversity of Scots pine monoterpenes? *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **366**: 1337–1345.

Jorissen H. 2021. Coral larval settlement preferences linked to crustose coralline algae with distinct chemical and microbial signatures. *Scientific Reports* **11**: 14610.

Jost L. 2006. Entropy and diversity. *Oikos* **113**: 363–375.

Jost L. 2007. Partitioning diversity into independent alpha and beta components. *Ecology* **88**: 2427–2439.

Junker RR. 2016. Multifunctional and diverse floral scents mediate biotic interactions embedded in communities. In: Blande JD, Glinwood R, eds. *Deciphering chemical language of plant communication.* Cham, Switzerland: Springer International, 257–282.

Junker RR. 2018. A biosynthetically informed distance measure to compare secondary metabolite profiles. *Chemoecology* **28**: 29–37.

Junker RR, Blüthgen N. 2010. Floral scents repel facultative flower visitors, but attract obligate ones. *Annals of Botany* **105**: 777–782.

Junker RR, Kuppler J, Amo L, Blande JD, Borges RM, van Dam NM, Dicke M, Dötterl S, Ehlers BK, Etl F *et al.* 2018. Covariation and phenotypic integration in chemical communication displays: biosynthetic constraints and eco-evolutionary implications. *New Phytologist* **220**: 739–749.

Junker RR, Tholl D. 2013. Volatile organic compound mediated interactions at the plant–microbe interface. *Journal of Chemical Ecology* **39**: 810–825.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Researach* **28**: 27–30.

Kessler A. 2015. The information landscape of plant constitutive and induced secondary metabolite production. *Current Opinion in Insect Science* **8**: 47–53.

Kessler A, Kalske A. 2018. Plant secondary metabolite diversity and species interactions. *Annual Review of Ecology, Evolution, and Systematics* **49**: 115–138.

Khan RM, Luk C-H, Flinker A, Aggarwal A, Lapid H, Haddad R, Sobel N. 2007. Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. *Journal of Neuroscience* **27**: 10015–10023.

Kim HW, Wang M, Leber CA, Nothias L-F, Reher R, Kang KB, van der Hooft JJJ, Dorrestein PC, Gerwick WH, Cottrell GW. 2021. NPClassifier: a deep neural network-based structural classification tool for natural products. *Journal of Natural Products* **84**: 2795–2807.

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B *et al.* 2021. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research* **49**: D1388–D1395.

Knaden M, Strutz A, Ahsan J, Sachse S, Hansson BS. 2012. Spatial representation of odorant valence in an insect brain. *Cell Reports* **1**: 392–399.

Larue A-AC, Raguso RA, Junker RR. 2016. Experimental manipulation of floral scent bouquets restructures flower–visitor interactions in the field. *Journal of Animal Ecology* **85**: 396–408.

Li D. 2018. HILLR: taxonomic, functional, and phylogenetic diversity and similarity through Hill Numbers. *Journal of Open Source Software* **3**: 1041.

Marion ZH, Fordyce JA, Fitzpatrick BM. 2015. Extending the concept of diversity partitioning to characterize phenotypic complexity. *The American Naturalist* **186**: 348–361.

Martin YC, Kofron JL, Traphagen LM. 2002. Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry* **45**: 4350–4358.

Mirzaei M, Züst T, Younkin GC, Hastings AP, Alani ML, Agrawal AA, Jander G. 2020. Less is more: a mutation in the chemical defense pathway of *Erysimum cheiranthoides* (Brassicaceae) reduces total cardenolide abundance but increases resistance to insect herbivores. *Journal of Chemical Ecology* **46**: 1131–1143.

Müller C, Bräutigam A, Eilers E, Junker R, Schnitzler J-P, Steppuhn A, Unsicker S, van Dam N, Weisser W, Wittmann M. 2020. Ecology and evolution of intraspecific chemodiversity of plants. *Research Ideas and Outcomes* **6**: e49810.

Müller C, Junker RR. 2022. Chemical phenotype as important and dynamic niche dimension of plants. *New Phytologist* **234**: 1168–1174.

O'Connor MI, Pennell MW, Altermatt F, Matthews B, Melián CJ, Gonzalez A. 2019. Principles of ecology revisited: integrating information and ecological theories for a more unified science. *Frontiers in Ecology and Evolution* **7**: 219.

Oksanen J, Blanchet F, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin P, O'Hara R, Simpson G, Solymos P *et al.* 2022. *VEGAN: community ecology package.* R package v.2.6-2. [WWW document] URL https://CRAN.R-project.org/package=vegan. [accessed 15 December 2022].

Petchey OL, Gaston KJ. 2006. Functional diversity: back to basics and looking forward. *Ecology Letters* **9**: 741–758.

Philbin CS, Dyer LA, Jeffrey CS, Glassmire AE, Richards LA. 2022. Structural and compositional dimensions of phytochemical diversity in the genus *Piper* reflect distinct ecological modes of action. *Journal of Ecology* **110**: 57–67.

R Core Team. 2022. *R: a language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. [WWW document] URL https://www.R-project.org/.

Raguso RA, Agrawal AA, Douglas AE, Jander G, Kessler A, Poveda K, Thaler JS. 2015. The raison d'être of chemical ecology. *Ecology* **96**: 617–630.

Rasmann S, Agrawal AA. 2011. Latitudinal patterns in plant defense: evolution of cardenolides, their toxicity and induction following herbivory. *Ecology Letters* **14**: 476–483.

Richards LA, Dyer LA, Forister ML, Smilanich AM, Dodson CD, Leonard MD, Jeffrey CS. 2015. Phytochemical diversity drives plant–insect community diversity. *Proceedings of the National Academy of Sciences, USA* **112**: 10973–10978.

Richards LA, Glassmire AE, Ochsenrider KM, Smilanich AM, Dodson CD, Jeffrey CS, Dyer LA. 2016. Phytochemical diversity and synergistic effects on herbivores. *Phytochemistry Reviews* **15**: 1153–1166.

Salazar D, Jaramillo A, Marquis RJ. 2016. The impact of plant chemical diversity on plant–herbivore interactions at the community level. *Oecologia* **181**: 1199–1208.

Schiestl FP. 2010. The evolution of floral scent and insect chemical communication. *Ecology Letters* **13**: 643–656.

Sedio BE. 2017. Recent breakthroughs in metabolomics promise to reveal the cryptic chemical traits that mediate plant community composition, character evolution and lineage diversification. *New Phytologist* **214**: 952–958.

Sedio BE, Devaney JL, Pullen J, Parker GG, Wright SJ, Parker JD. 2020. Chemical novelty facilitates herbivore resistance and biological invasions in some introduced plant species. *Ecology and Evolution* **10**: 8770–8792.

Sedio BE, Rojas Echeverri JC, Boya PCA, Wright SJ. 2017. Sources of variation in foliar secondary chemistry in a tropical forest tree community. *Ecology* **98**: 616–623.

Shimatani K. 2001. On the measurement of species diversity incorporating species differences. *Oikos* **93**: 135–147.

Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C. 2021. COCONUT online: collection of open natural products database. *Journal of Cheminformatics* **13**: 2.

Szöcs E, Stirling T, Scott ER, Scharmüller A, Schäfer RB. 2020. WEBCHEM: an R package to retrieve chemical information from the web. *Journal of Statistical Software* **93**: 1–17.

Tewes LJ, Michling F, Koch MA, Müller C. 2018. Intracontinental plant invader shows matching genetic and chemical profiles and might benefit from high defence variation within populations. *Journal of Ecology* **106**: 714–726.

Tripathi A, Vázquez-Baeza Y, Gauglitz JM, Wang M, Dührkop K, Nothias-Esposito M, Acharya DD, Ernst M, van der Hooft JJJ, Zhu Q *et al.* 2021. Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. *Nature Chemical Biology* **17**: 146–151.

Tuomisto H. 2012. An updated consumer's guide to evenness and related indices. *Oikos* **121**: 1203–1218.

Volf M, Segar ST, Miller SE, Isua B, Sisol M, Aubona G, Šimek P, Moos M, Laitila J, Kim J *et al.* 2018. Community structure of insect herbivores is driven by conservatism, escalation and divergence of defensive traits in *Ficus*. *Ecology Letters* **21**: 83–92.

Walker B, Kinzig A, Langridge J. 1999. Plant attribute diversity, resilience, and ecosystem function: the nature and significance of dominant and minor species. *Ecosystems* **2**: 95–113.

Walker TWN, Alexander JM, Allard P, Baines O, Baldy V, Bardgett RD, Capdevila P, Coley PD, David B, Defossez E *et al.* 2022. Functional traits 2.0: the power of the metabolome for ecology. *Journal of Ecology* **110**: 4–20.

Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, Nguyen DD, Watrous J, Kapono CA, Luzzatto-Knaan T *et al.* 2016. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nature Biotechnology* **34**: 828–837.

Wang S, Alseekh S, Fernie AR, Luo J. 2019. The structure and function of major plant metabolite modifications. *Molecular Plant* **12**: 899–919.

Wang Y, Backman TWH, Horan K, Girke T. 2013. FMCSR: mismatch tolerant maximum common substructure searching in R. *Bioinformatics* **29**: 2792–2794.

Wetzel WC, Whitehead SR. 2020. The many dimensions of phytochemical diversity: linking theory to practice. *Ecology Letters* **23**: 16–32.

Whitehead SR, Bass E, Corrigan A, Kessler A, Poveda K. 2021a. Interaction diversity explains the maintenance of phytochemical diversity. *Ecology Letters* **24**: 1205–1214.

Whitehead SR, Schneider GF, Dybzinski R, Nelson AS, Gelambi M, Jos E, Beckman NG. 2021b. Fruits, frugivores, and the evolution of phytochemical diversity. *Oikos* **2022**: e08332.

Wink M. 2010. *Biochemistry of plant secondary metabolism. Annual plant reviews, vol. 40.* Oxford, UK: Wiley-Blackwell.

Zhou W, Kügler A, McGale E, Haverkamp A, Knaden M, Guo H, Beran F, Yon F, Li R, Lackus N *et al.* 2017. Tissue-specific emission of (*E*)-α-bergamotene helps resolve the dilemma when pollinators are also herbivores. *Current Biology* **27**: 1336–1341.

Zu P, Boege K, del Val E, Schuman MC, Stevenson PC, Zaldivar-Riverón A, Saavedra S. 2020. Information arms race explains plant–herbivore chemical communication in ecological communities. *Science* **368**: 1377–1381.

Züst T, Mirzaei M, Jander G. 2018. *Erysimum cheiranthoides*, an ecological research system with potential as a genetic and genomic model for studying cardiac glycoside biosynthesis. *Phytochemistry Reviews* **17**: 1239–1251.

Züst T, Strickler SR, Powell AF, Mabry ME, An H, Mirzaei M, York T, Holland CK, Kumar P, Erb M *et al.* 2020. Independent evolution of ancestral and novel defenses in a genus of toxic plants (*Erysimum*, Brassicaceae). *eLife* **9**: e51712.

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Dataset S1 Dataset used for all analyses and figures in the manuscript.

Fig. S1 Comparisons of different measures of phytochemical diversity and dissimilarity for the semi-simulated dataset with cardenolides and glucosinolates.

Fig. S2 Comparison of glucosinolate diversity and dissimilarity, calculated with and without taking compound dissimilarities into account, for 48 *Erysimum* species.

Fig. S3 Comparisons of different measures of phytochemical diversity for eight groups of phytochemical samples simulated to have a high or low richness, evenness and compound dissimilarity.

Fig. S4 Comparisons of compound dissimilarities calculated using the three different methods in the *compDis* function.

Fig. S5 Dendrogram showing compound dissimilarities for floral scent compounds found in *Achillea millefolium* and *Cirsium arvense*.

Fig. S6 Computation times of the *quickChemoDiv* function for datasets with varying number of compounds, computed with and without calculation of compound dissimilarity data.

Notes S1 R-script used to perform statistical analyses and create figures in the manuscript.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.