Chemoecology
https://doi.org/10.1007/s00049-020-00322-4

SHORT COMMUNICATION

Correlation in plant volatile metabolites: physiochemical properties as a proxy for enzymatic pathways and an alternative metric of biosynthetic constraint

Jordan A. Dowell[1] · **Chase M. Mason**[1]

Received: 26 February 2020 / Accepted: 7 August 2020

Abstract
From intra-individual regulation of metabolism to entire ecosystem functioning, the thousands of biogenic compounds produced by organisms serve as a major component of ecological and evolutionary diversity mediating interactions across scales. Earlier work considers canonical reactions, defined as reactions specified along accepted (experimentally validated or theoretically postulated) biosynthetic pathways, as the primary form of constraint on chemical diversity. An emerging understanding of non-canonical reactions (reactions which occur independently of canonical reactions) suggests that the physical chemistry of compounds may play a larger role in constraining chemo-diversity than previously thought. We selected 24 studies of plant volatile profiles, satisfying a defined set of criteria, to assess the extent of correlation among profiles attributable to either shared biosynthetic enzymes or physiochemical properties. Across studies, regardless of treatment, 0.17 (±0.16 SD) adjusted R^2 was attributed to both shared biosynthetic enzymes and physiochemical properties; however, there were no significant differences between the amount of unique variance attributed to shared enzymes (0.05 ± 0.08 SD) or physiochemical properties (0.03 ± 0.06 SD). The amount of unique variance explained by physiochemical properties, independent of their canonical relationships, provides a metric for evaluating the role of non-enzymatic and non-canonical reactions in constraining molecular diversity.

Keywords Biosynthetic constraint · Enzyme promiscuity · Physiochemical · Specialized metabolism · Plant volatiles

Introduction

Chemistry is the primal language of life, given that across the diversity of taxonomic groups individuals have evolved unique ways to produce and interpret complex molecules from their surroundings. As chemistry provides a fundamental way for organisms to interact with one another, it is no surprise that compound diversity has been linked to the structuring of communities at multiple levels (Micallef et al. 2009; Moles et al. 2011; Rubin et al. 2015; Uesugi et al. 2016). For example, in plants at the community level, volatile chemical diversity can influence herbivory rates more strongly than taxonomic diversity, such that high chemo-diversity reduces herbivory rates (Salazar et al. 2016; Karban 2017; Dahlin et al. 2018). Plant volatile compounds (VOCs) are predominantly synthesized by a small number of biochemical pathways, although simple modifications of hydroxyl, acetyl and methyl groups provide much of the diversity of compounds emitted, and in turn the diverse responses in organisms receiving them (Dudareva 2004; Gang 2005; Dudareva et al. 2006). Within individuals, the diversity of compounds produced can vary across organs and may change with ontogeny (Borges et al. 2013; Hoffmeister et al. 2016; Vieira et al. 2016; Killiny and Jones 2017). The chemical diversity recorded among individuals is even more dramatic than comparisons across organs (Jaeger et al. 2016; Prieto-Benítez et al. 2016; Schrader et al. 2017). As there are many biotic and abiotic factors that play a role in altering chemical profiles, little is known about the genetic

Communicated by Günther Raspotnig.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00049-020-00322-4) contains supplementary material, which is available to authorized users.

✉ Jordan A. Dowell
Jordan.dowell@ucf.edu

[1] Department of Biology, University of Central Florida, Orlando, FL 32816, USA

Published online: 14 August 2020

control of VOCs, or inheritance patterns (Gross et al. 2009; Laothawornkitkul et al. 2009; Rowan et al. 2009; Kumar et al. 2015). The current major hypothesis for the control of volatile metabolism is that enzyme abundance and protein regulation are the primary drivers of chemo-diversity and compound abundance (Gang 2005; Khersonsky and Tawfik 2010). However, metabolic regulation is subject to enzymatic affinity, reaction rate, and substrate/product concentration, all of which are affected by the physical chemistry of the enzymes and substrates/products (Guzmán et al. 2015).

Junker et al. (2018) was one of the first studies that took into consideration the role of enzymes as constraining chemo-diversity. These constraints posit that the maximal variation in abundance between two compounds is dependent on the number of shared enzymes necessary to produce them. This concept is analogous to current assumptions made in phylogenetic comparative methods such that maximal trait variation achievable between two taxa is dependent on their shared evolutionary history (de Bello et al. 2015). In the absence of a phylogeny, an enzymatic dissimilarity matrix formed of enzymes necessary for metabolite production is compared to the variation observed among compounds within individual volatile metabolite profiles. In their study, the correlation (Mantel's *R*) between compounds produced by plant replicates and their enzymes varied from − 0.01 to 0.73 (Junker et al. 2018). The lack of stability in correlation suggests that there may be missing pieces to this theory.

A recent emerging understanding of 'underground metabolism' and the influence of non-canonical reactions across many domains of biology supplies a possible missing piece to the theory of biosynthetic constraint (D'Ari and Casadesu 1998; Liu et al. 2011; Keller et al. 2015; Noctor et al. 2015; Guzmán et al. 2018). A non-canonical reaction is an umbrella term used to describe reactions that are not a part of an 'accepted' and/or experimentally validated biochemical pathway. Types of non-canonical reactions include enzyme substrate/product promiscuity, single-element catalysis, and spontaneous reactions (D'Ari and Casadesu 1998; Guzmán et al. 2015; Piedrafita et al. 2015). Enzyme substrate/product promiscuity, or the acceptance of more than one substrate and/or production of more than one product, is common, as 37% of enzymes studied in the *E. coli* K12 genome have promiscuity at the product or substrate level under stable conditions (Nam et al. 2012). Variation in enzyme promiscuity and specificity can both be induced via environmental shifts, where primary driving forces include, but are not limited to: electric potential, substrate/product concentration, and temperature (Khersonsky and Tawfik 2010; Fried and Boxer 2017).

All chemical reaction types have one thing in common: the physical chemistry of the reactants, catalysts, and products govern their reactivity. Recently the field of computational retrosynthesis has grown significantly and begun to address key questions concerning compound and pathway prediction. Watson and colleagues demonstrated the utility of predictors of physical chemical dynamics in the recreation of many complex organic synthetic pathways, both with and without the help of enzymes (Watson et al. 2019). These predictors are created in silico where a molecule is described and traits such as ionization potential, total polar surface area, and eccentric connectivity are derived based on our current knowledge of physical chemistry. Earlier studies showing the utility of physical chemistry in predicting metabolic pathways—both canonical and non-canonical—have been highly successful (Lanzeni et al. 2008; Barupal and Fiehn 2017; Kunimoto et al. 2017; Delépine et al. 2018; Segler et al. 2018; Klamt et al. 2018; Watson et al. 2019). Between the fields of retrosynthetic chemistry and chemical ecology, our goals are still the same: 'how can we attribute variance in compound abundance based on the interdependence of observed compounds?'. Our coarse understanding of intracellular environmental heterogeneity paired with the postulation of major enzymes, in light of rampant convergent evolution of compounds and their precursors, leaves a large gap in general theories describing biosynthetic constraints of chemo-diversity. To further investigate these constraints, we must first gain an in-depth understanding of the relative role of non-canonical reactions.

The objectives of this study are: (1) evaluate the role of canonical and non-canonical biosynthetic constraints on plant volatile metabolic profiles, where such constraints are defined as the amount of variation among compounds produced which is attributable to either shared enzymes or physiochemical properties, (2) identify trends of potential shifts in constraints based on treatments, and (3) evaluate the utility of in silico physiochemical predictors as a suitable proxy for experimentally confirmed enzymatic pathways. To assess the role of non-canonical reactions in production of VOCs, we compared the correlation among VOC profiles across a diverse assemblage of species to two datasets: a presence/absence table of canonical biosynthetic enzymes and a table of physiochemical properties based on the three-dimensional structure of each compound. We hypothesized that, among studies, variation in the correlation of VOCs between individuals is explained by either shared biosynthetic enzymes or similarity in physiochemical properties; however, we expected there to be a higher degree of shared explainable variance, defined as the variance explained by both shared biosynthetic enzymes and physiochemical properties that is not unique to either sets of properties. A finding that shared variance explains a higher degree of VOC profile variation compared to the uniquely explained variance of either biosynthetic enzymes or physiochemical properties would support the utility of physiochemical predictors as valid proxies for experimentally validated enzymatic

pathways in the assessment of biosynthetic constraint. Furthermore, within studies we categorized individuals based on treatments (control, chemical-induction, and biotic-induction) and assessed variance partitions relevant to the degree of biosynthetic constraint. Between chemical-induction and biotic-induction treatments, we expected greater similarity among variance components in biotic-induction treatments over chemical-induction; as plant-interactions are a delicate balance of chemical reactions, the effect of a single compound may not be complex enough to elicit a biologically similar response.

Materials and methods

Data set description

To build a dataset with which to examine constraints on chemo-diversity, studies were identified from the Dryad Digital Repository using a search with combinations of terms (plant* volatile*), including only records published before 2019. The initial dataset included 74 studies; however, only 24 studies (containing 27 individual experiments) were included based on the following criteria (Online Resource 1):

- At minimum, 5 VOCs were quantified.
- 80% of the VOCs produced match compounds for which there is an annotated enzymatic pathway.
- Profiles collected had a minimum number of seven individuals per study, and at least five biological replicates per treatment.
- Profiles were collected via headspace or solvent extraction and analyzed on a gas-chromatograph coupled to a mass spectrometer.

Data were partitioned from 20 of the 24 studies into categories based on treatment type, either chemical-induction or biotic-induction. Chemical-induction was defined as the usage of an accepted plant hormone or other plant-derived compounds to elicit a response. From the data collected, these hormones included primarily methyl jasmonate and methyl salicylate. Biotic-induction included the use of an herbivore, pollinator, microbe, and/or viral agent as the response elicitor. Control samples from all studies present in these two categories were considered independently as well.

Biosynthetic enzymes (E-table and E-matrix)

Enzyme biosynthetic pathways were extracted for 137 volatile organic compounds from Junker et al. (2018). We reconstructed the analyses of Junker et al. by assessing the correlation of canonical reactions and compound profiles. The E-table begins with selected compounds as rows and biosynthetic enzymes as columns. A cell in the table contains a '1' if the enzyme of the corresponding column is involved in the production of the compound of the corresponding row. If the enzyme is not involved in compound production, the cell contains a '0'. To quantify a dissimilarity measure for the enzymatic dataset, a Sørensen's Index was used.

Metabolomic profiles (P-table and P-matrix)

To construct the P-table, we extracted volatile profiles produced from various organs and experimental conditions across studies. The P-table has compounds as columns, samples as rows, and each cell corresponds to the relative ratios of compounds observed in each sample. The P-table is made square and comparable to the E-matrix, forming the P-matrix, via a Bray–Curtis dissimilarity index of all compounds present.

Physiochemical properties (C-table and C-matrix)

The C-table initially includes compounds as rows and physiochemical properties as columns, with each cell filled with the value a physiochemical property for a given compound. For all molecules represented in the E-table, we obtained 207 continuous and binary traits that exemplify the three-dimensional structure and physics of each compound from PubChem, an open-source chemical information repository, for use as predictors of physical chemistry (Willighagen et al. 2017; Kim et al. 2019). These traits include descriptions of surface polarity, carbon backbone structure, hydrogen bond donors/acceptors, as well as other in silico predictors (Online Resource 2). We formed the C-matrix from the C-table and made it comparable to other matrices via the same processes as the P-matrix.

C- and E-matrix similarity

We compared the C- and E-matrices including all 137 VOCs for correlation with a Mantel test. A randomization approach with 10,000 iterations assessed the significance of the Mantel correlation. Tree visualization of E- and C-matrices were conduct via hierarchal clustering based on Sørensen's or Bray–Curtis dissimilarity values in a nearest neighbor joining framework (Galili 2015). We compared the hierarchal clustering between both trees for identification of novel groupings.

Biosynthetic constraint evaluation

We initially assessed constraint with the same methods of Junker via Mantel tests (2018). To supply an estimate of biosynthetic constraint that occurred in studies regardless

of treatment, we compared the P-matrix of each study independently to its corresponding and pruned E- and C-matrix. We pruned each E- and C-matrix to include only compounds observed in each study to ensure all three matrices were comparable. We correlated the E-matrix and C-matrix in each study with Mantel tests and assessed significance via the same randomization approach.

A redundancy analysis (RDA) partitioned the variance of each P-matrix to corresponding E- and C-matrices (Peres-Neto et al. 2006; Everitt and Dunn 2013). Variance partitioning in this framework is attempt to resolve the explanatory power of the E- and C-matrices in relation to the P-matrix (Oksanen 2008). A redundancy analysis (RDA) of two predictor matrices against a third response matrix yields five variance partitions: E|C + E:C + C|E, E|C, E|C + E:C, E:C, C|E + E:C, and C|E (Online Resource 3). In this framework, the combinatory partition is E|C + E:C + C|E, which includes all variance explained by the E- and C-matrices. E|C and C|E are variance components uniquely explained by the C-matrix or the E-matrix, while indistinguishable variance explained by both the E- and C-matrix is the component E:C, a non-testable fraction of shared explained variance. The shared variance component (E:C) explains variance that it is representative of both shared enzymes and physiochemical properties. Biologically, this fraction could potentially represent correlation attributable to enzyme abundance, variable reaction affinity, and other complex enzyme/substrate/product interactions. In explicitly examining the E|C component of variance explained, we remove the variance attributed to more complex enzyme/substrate/product interactions, such that the remaining variance explained by the E|C component represents a simple 'lock-and-key' model of metabolism, where simple presence/absence of enzymes contributes to correlation. Conversely, the C|E component of variance explained rejects the presence/absence nature of the 'lock-and-key' model and only explains non-enzymatically canonical reactions. Observation of variance explained by either the E- or C-matrix includes unique variance explained (C|E or E|C) and shared variance explained (E:C) yielding the observed variance explained as E|C + E:C or C|E + E:C for the E-matrix or C-matrix, respectively.

We used a one-way randomization approach (10,000 iterations) to assess the significance of individual component partitions compared to a reduced model. Each reduced model is simply missing the variance component of interest. For example, to test the E|C fraction the E|C fraction is compared to a model which explains zero variance. As there were differences in sample sizes among studies and Mantel correlations may be positively or negatively correlated, we further defined the adjusted R^2 of the RDA as an alternative measure of constraint. Although a traditional R^2 value would equal 1, we chose an adjusted R^2 value which will approximate 1, while adjusting for number of samples and predictors allowing for comparisons across studies with unequal experimental dimensions (Peres-Neto et al. 2006). In this instance, nonsignificant adjusted R^2values indicate that observed compounds are independently distributed, such that there is no observed correlation in metabolite profiles explainable by the variance component of interest.

All analyses were conducted within the R environment version 3.6.0, with the following packages: ChemmineR, dendextend, rcdk, and vegan (Cao et al. 2008; Oksanen 2008; Guha and Rojas-Chertó 2010; R Core Team 2019; Galili 2015).

Results

Enzyme and physiochemical similarity

The reference datasets of biosynthetic enzymes for 137 VOCs and their physiochemical properties were positively correlated ($r = 0.65$, $p < 0.001$) (Fig. 1). The average enzymatic dissimilarity among VOCs was 0.79 (±0.36 SD), in comparison with physiochemical properties with a dissimilarity of 0.25 (±0.15 SD). However, within studies the average observed correlations were less than half at 0.21 (±0.04 SD), with mean enzymatic dissimilarities of 0.68 (±0.10 SD) and physiochemical dissimilarities of 0.21 (±0.04 SD).

Full study comparisons

Among studies disregarding treatment, 17% had a significant correlation between physiochemical properties and individual metabolite profiles, while 30% of studies had a significant correlation between shared enzymes and metabolite profiles ($p < 0.05$). In 91% of studies, the physiochemical properties of compounds present were significantly correlated with shared biosynthetic enzymes ($p < 0.05$).

After variance partitioning, only four studies had variance not significantly attributed to either shared biosynthetic enzymes or physiochemical properties. The average variance component attributed to the joint use of chemical and enzymatic matrices (E|C + E:C + C|E) was significantly different from all other components, while the shared variance alone (E:C) and unique variance partitions (C|E and E|C) were not significantly different from each other (Fig. 2, Table 1). Each unique variance component (E|C or C|E) was significantly different from the combination of each unique variance component plus the inclusion of shared variance (E|C + E:C or C|E + E:C).

Treatment comparisons

Among observed compounds in control groups ($N = 23$), physiochemical properties were significantly correlated

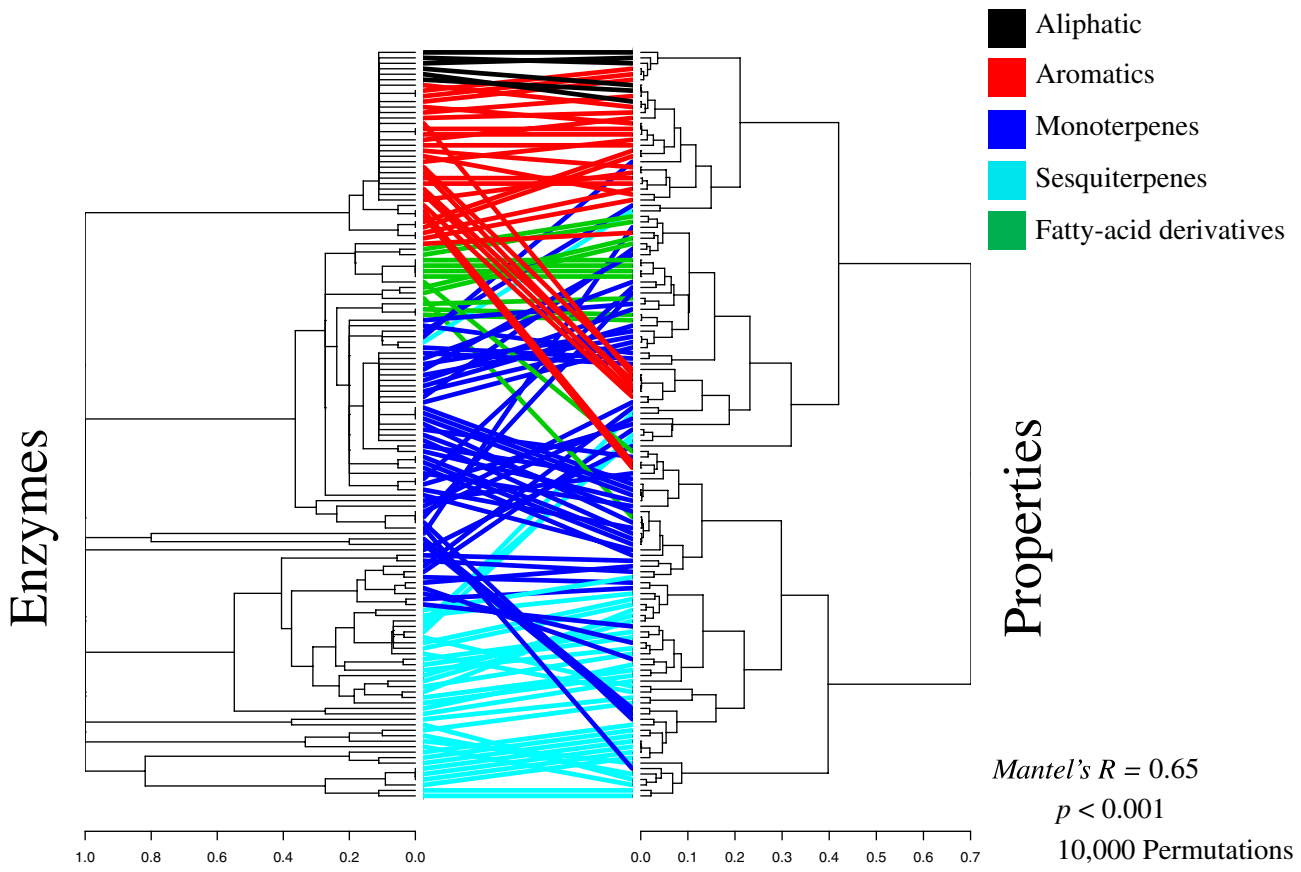

Fig. 1 Tangle gram of 137 VOCs, comprised of datasets concerning shared enzymes (left) and shared physiochemical properties (right). Metabolites are classed as aliphatics, aromatics, monoterpenes, sesquiterpenes, or fatty-acid derivatives (FAD)

with shared biosynthetic enzymes in 87% of groups. 71% of chemical-induction ($N=5$) groups and 87% of biotic-induction ($N=15$) groups had significant correlations between physiochemical properties and shared biosynthetic enzymes, respectively. Average correlations between shared biosynthetic enzymes and physiochemical properties of compounds observed were comparable among treatment groups (Table 1).

Surprisingly, among studies with a biotic-induction treatment, 27% of replicated treatment groups within a study had a significant correlation between physiochemical properties and VOC profiles, while 27% of treatments had a significant correlation between shared biosynthetic enzymes and VOC profiles. In comparison, approximately 15% of chemical-induction groups had a significant correlation between shared biosynthetic enzymes or physiochemical properties and VOC profiles. Control and biotic-induction treatments were comparable as 30% of replicate groups had a significant correlation between shared enzymes and VOC profiles, and 26% of Chemical-induction replicate groups had correlations between chemical properties and variation in compound abundance (Table 1).

With respect to variance partitioning, among control, biotic-induction, and chemical-induction groups, 92%,71%, and 100% had variance significantly attributed to either shared biosynthetic enzymes or physiochemical properties, respectively ($p<0.05$). Like full study comparisons, there were significant differences between adjusted R^2 values among variance components (Fig. 3). The joint effect of physiochemical and enzymatic matrices (E|C + E:C + C|E) in control and biotic-, and chemical-induction groups were significantly higher than unique variance components (E|C or C|E); however, the joint effect was not significantly different compared to the combination of each unique component and the inclusion of shared variance (E|C + E:C or C|E + E:C).

Discussion

Shared enzymes and chemical properties

The interdependent nature of metabolite biosynthesis precludes the use of conventional dissimilarity measures for comparisons of metabolic profiles. Previous work presents

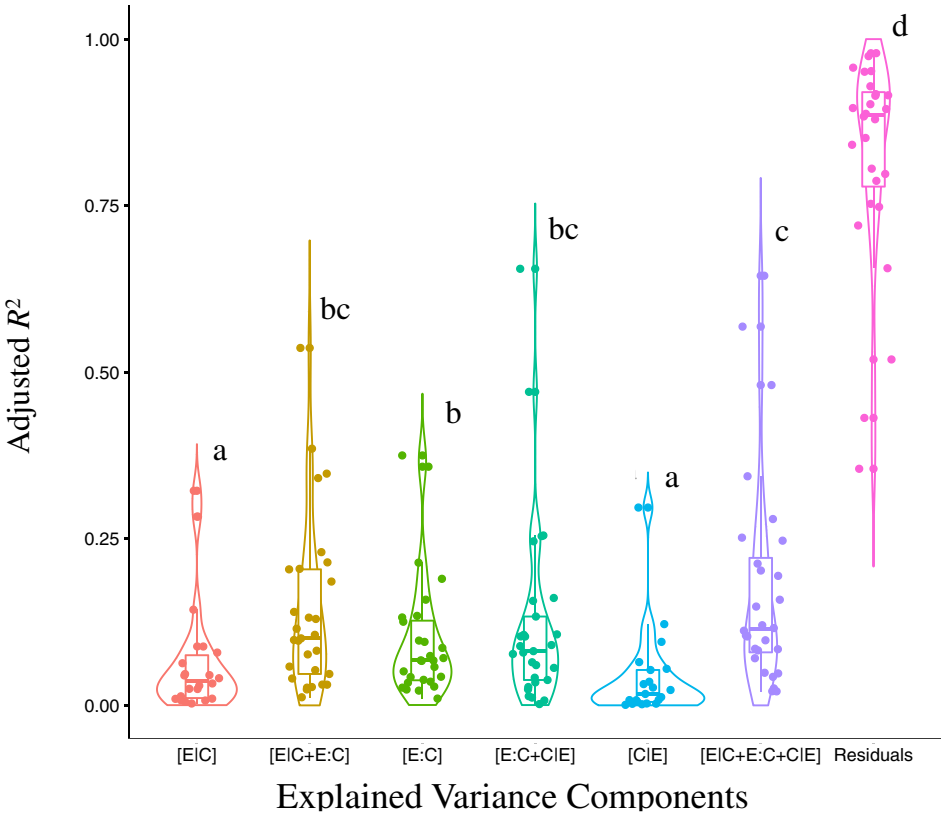

Fig. 2 Violin plots of adjusted R^2 of variance components for the full dataset. Letters indicate significant differences among groups assessed via a pairwise permutation test ($p < 0.01$). The combinatory partition is E|C + E:C + C|E, where E|C and C|E are variance components uniquely explained by the C-matrix or the E-matrix. Indistinguishable variance explained by both the E- and C-matrix is E:C, a non-testable fraction of shared explained variance. Observation of variance explained by the E- or C-matrix includes unique variance explained (C|E or E|C) and shared variance explained (E:C) yielding the observed variance explained as E|C + E:C or C|E + E:C for the E-matrix or C-matrix, respectively. The residual component is the residual variance that is not explained by either the C- or E-matrix

Table 1 Mantel's *R* correlations between physiochemical properties (C-matrix), biosynthetic enzymes (E-matrix), and individual plant volatile metabolite profiles within studies (P-matrix) (±standard deviation)

Datasets	C and E	C and P	E and P
Full studies	0.59 (±0.26)	0.10 (±0.22)	0.14 (±0.22)
Controls	0.55 (±0.25)	0.12 (±0.20)	0.11 (±0.22)
Chemical-induction	0.53 (±0.40)	0.05 (±0.13)	0.13 (±027)
Biotic-induction	0.57 (±0.27)	0.09 (±0.16)	0.11 (±0.22)

a conceptual framework for addressing this issue and poses the use of shared enzymes within biosynthetic pathways as a dependency structure which can be assessed and accounted for (Junker 2018). However, a major caveat in using shared enzymes to establish biosynthetic constraints is that the metabolic pathways in question must be well characterized (Junker 2018). As a potential alternative, it has been suggested that using chemical classes to derive biosynthetic constraints is likely to have comparable success to canonical pathways (Junker 2018). However, chemical classes are not always indicative of potential compound interdependence, and using chemical classes precludes the examination of interdependence among metabolites within a single class.

Observed metabolite properties are the result of previous chemical reactions, in which enzymes play a part (Patti et al. 2012; Singh et al. 2015). Chemical classes typically denote shared structure or properties; thus, we hypothesized that physiochemical properties should serve as a suitable, more descriptive proxy. In support of this hypothesis, physiochemical properties and shared enzymes were positively correlated. However, the average dissimilarity observed among metabolites derived from shared enzymes was higher than the average dissimilarity observed due to physiochemical properties. This greater separation of individual metabolites reflects deep separations based on major enzymatic pathways, in comparison with potentially more nuanced description of observed physiochemical properties. Among shared enzymes there were deep nodes separating major classes; however, at the individual metabolite level there were several shallow separations, as well as many polytomies (Fig. 1). In the tree derived from shared enzymes, there was a strong recovery of chemical classes (aliphatics, aromatics, monoterpenes, sesquiterpenes, and fatty-acid derivatives (Fig. 1). The tree derived from physiochemical properties had reduced recovery of clusters successfully parsing chemical class, in comparison with the tree derived from shared enzymes (Fig. 1). We expected the lack of full resolution of chemical classes in the physiochemical property-based tree, as multiple pathways can give rise to compounds with comparable properties regardless of class. Even with reduced efficacy, our results further support the potential usage of

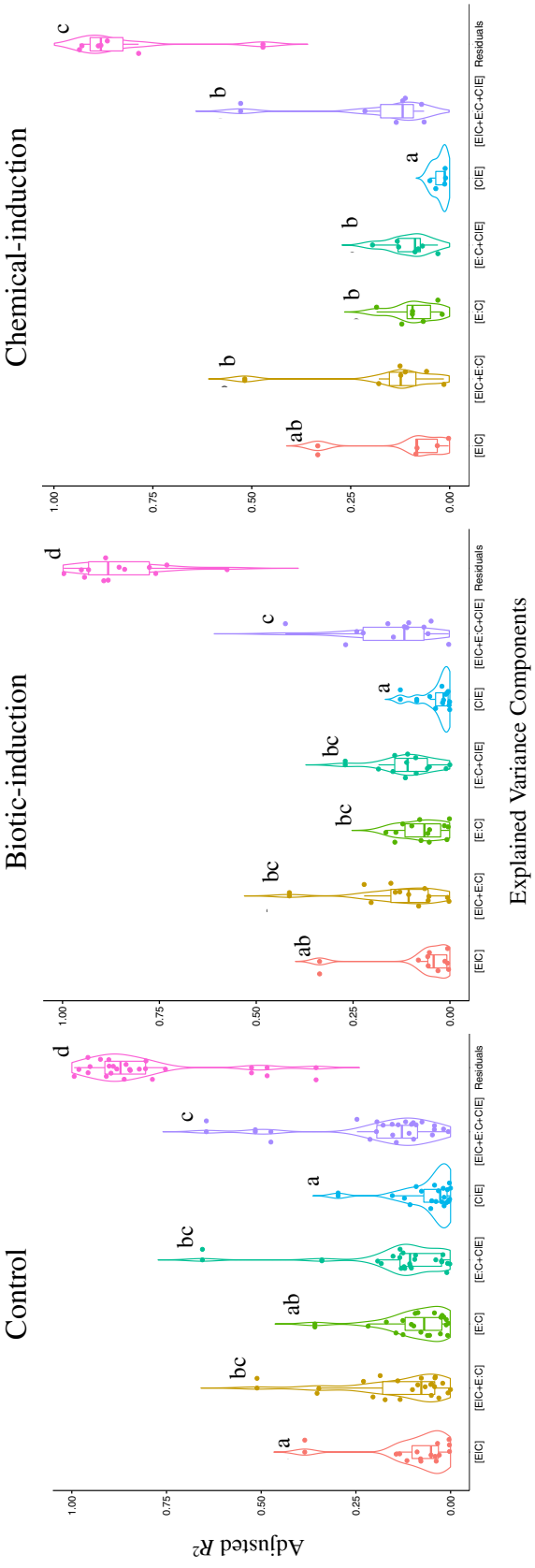

Fig. 3 Violin plots of adjusted R^2 of variance components for each treatment dataset. Letters indicate significant differences among groups assessed via a pairwise permutation test ($p < 0.01$). The combinatory partition is E|C + E:C + C|E, where. E|C and C|E are variance components uniquely explained by the C-matrix or the E-matrix. Indistinguishable variance explained by both the E- and C-Matrix is E:C, a non-testable fraction of shared explained variance. Observation of variance explained by the E- or C-matrix includes unique variance explained (C|E or E|C) and shared variance explained (E:C) yielding the observed variance explained as E|C + E:C or C|E + E:C for the E-matrix or C-matrix, respectively. The residual component is the residual variance that is not explained by either the C- or E-matrix

these physiochemical properties as predictors of biosynthetic constraint given the significant positive correlation (Table 1; Fig. 1).

Biosynthetic constraints in plant VOC experiments

From an analytical perspective, biosynthetic constraint is defined by Junker (2018) as the correlation (*Pearson's r*) between observed compounds and their corresponding shared biosynthetic enzymes. However, we suggest the use of explainable variance among metabolites that is attributable to shared biosynthetic enzymes or other relevant predictors as a more appropriate parameter for quantifying 'biosynthetic constraint'. The bounding nature of an R^2 value between 0 and 1 for explained variation supplies a more intuitive assessment of constraint and allows for incorporation and comparison of both positive and negative correlations. Additionally, in metabolomic dataset with hundreds of metabolites, subsets of metabolites may be positively or negatively correlated within the full profile which would obscure a traditional correlation measure (Borges et al. 2013; Yip et al. 2017). The large percentage of studies demonstrating significant variation attributed to physiochemical properties or shared enzymes support our initial hypothesis that there is rampant correlation among observed plant volatile profiles, regardless of experimental treatment, establishing the need to develop novel methods to address the non-independence of metabolite data (Table 1). Although the absolute variance explained was low, metabolite pathways do not conform to simple tree-like models and are better represented by networks; thus this approach may be an underestimation of interdependence, as one cannot account for potential feedback loops associated with biosynthesis with current tree-based models (Kruger and Ratcliffe 2012; Notebaart et al. 2018) (Table 1). However, there was an equivalent amount of unique variance explained by either shared enzymes or physiochemical properties, supporting the use of physiochemical properties and canonical enzymatic pathways as complementary datasets for assessment of biosynthetic constraint under current assumptions/methods (Kruger and Ratcliffe 2012; Costello and Martin 2018).

Across treatment types there were no significant differences observed between unique variance components, in opposition to our expectation that biotic-induction treatments would exhibit greater similarity among variance components over chemical-induction treatments. However, as chemical-induction treatments had a smaller sample size in comparison with biotic-induction and control treatment groups, sample size differences may have affected the result by reducing statistical power. The results obtained do, however, support the validity of experiments using plant hormones or other plant-derived compounds to elicit metabolic responses in lieu of more difficult biotic-induction treatments, as effects were comparable (Tables 1, 2).

Unique variance explained by physiochemical properties presents a potential quantification metric for physiochemical impacts on constraint of VOC profiles. As well, considering the similarity in values, this provides support for the use of physiochemical parameters as a suitable proxy to canonical biosynthetic pathways in this context, with minor information loss (Table 2). Although disentangling the effects of non-canonical non-enzymatic reactions from canonical non-enzymatic reactions is not possible with the use of physiochemical parameters alone, the similarity of the unique variance explained by the C-matrix and E-matrix does however add support to the potential of 'underground metabolism' in contribution to variation among observed compounds (Gutensohn et al. 2014; Piedrafita et al. 2015; Notebaart et al. 2018).

Potential utility

As mentioned in Junker (2018), assessment and use of biosynthetic constraints is easily transferable to studies concerning metabolite variation across evolutionary and ecological scales. Methods which assess biosynthetic constraint based on an enzymatic measure can assess metabolic pathways and enzymes shared by individuals and their function in a meaningful way, as opposed to dissimilarity measures based on metabolite identity and abundance (Junker 2018). Assessment of variation based on physiochemical properties provides a directly applicable quantification of

Table 2 Adjusted R^2 of variance components across dataset partitions (±standard deviation)

Datasets	E|C + E:C + C|E	E|C	E|C + E:C	E:C	E:C + C|E	C|E
Full studies	0.17 (±0.16)	0.05 (±0.08)	0.14 (±0.13)	0.09 (±0.09)	0.12 (±0.14)	0.03 (±0.06)
Controls	0.16 (±0.16)	0.05 (±0.09)	0.12 (±0.13)	0.07 (±0.09)	0.12 (±0.14)	0.05 (±0.07)
Chemical-induction	0.18 (±0.16)	0.07 (±0.12)	0.16 (±0.17)	0.09 (±0.06)	0.10 (±0.05)	0.02 (±0.02)
Biotic-induction	0.13 (±0.12)	0.04 (±0.09)	0.11 (±0.11)	0.07 (±0.06)	0.09 (±0.08)	0.02 (±0.04)

The combinatory partition is E|C + E:C + C|E, where E|C and C|E are variance components uniquely explained by the C-matrix or the E-matrix. Indistinguishable variance explained by both the E- and C-matrix is E:C, a non-testable fraction of shared explained variance. Observation of variance explained by the E- or C-matrix includes unique variance explained (C|E or E|C) and shared variance explained (E:C) yielding the observed variance explained as E|C + E:C or E:C + C|E for the E-matrix or C-matrix, respectively

biosynthetic constraint, in light of unknown biosynthetic factors with minimal information loss. One can potentially infer the degree of this information loss, based on the similarity between unique variances attributed to either enzymatic or physiochemical properties observed in this study.

The incorporation of publicly available physiochemical information through resources like PubChem can reduce the burden on researchers to track down canonical pathway information or wait for pathways of focal metabolites to be experimentally validated before estimating biosynthetic constraint in a given study system. Previous work made use of metrics of biosynthetic constraint for comparisons across species and experimental conditions to describe relative increases and decreases in constraint based on conditions or as independent trait characteristics (Junker et al. 2018). However, we propose several advantages of using biosynthetic constraints in manipulative studies.

The identification and quantification of biosynthetic constraints on metabolite profiles from a study with multiple treatments supplies a unique set of parameters. Traditionally, treatments tasked with identifying biotic resistance focus on stimulating hormone pathways (e.g. jasmonic or salicylic acid) (Thaler et al. 2012; Hoffmeister and Junker 2017). Hormone pathways may interact with biosynthetic pathways by regulating hypothesized tree-like biosynthetic pathways at their base or closer to their terminal ends. Given this assumption, how metabolite profiles and individual metabolites change in response to a given treatment can supply insight into how treatments affect the underlying biosynthetic pathway or focal group of compounds. For example, if a treatment primarily stimulates the base of a biosynthetic pathway, one would expect high biosynthetic constraint on metabolites observed if they are derived from the stimulated biosynthetic pathway, with deviations present due to potential noncanonical reactions (Fig. 4). Under high biosynthetic constraint, we expect profiles to shift uniformly in the same direction, supplying a visual diagnostic of treatment effects on biosynthesis (Fig. 4a**).** In a more quantitative sense, biosynthetic constraint can be used as a parameter to identify individual metabolites that are potentially being regulated independently of other observed metabolites. In other words, if one estimates the proportion of variance in a dataset explainable by similarities in metabolite physiochemical properties, this value is an informative effect-size threshold. In an example of high biosynthetic constraint, treatment effects on metabolite profile variation remain below what is expected due to shared biosynthetic enzymes or physiochemical properties indicating that the treatment maybe be stimulating the base of the metabolic pathway (Fig. 4).

Regulation of biosynthesis may occur on multiple parts of a biosynthetic pathway leading to potentially lower biosynthetic constraints. Under a lower biosynthetic constraint, we postulate two extreme examples, which have arbitrarily been given the terms γ and δ (Fig. 4). In low constraint γ,

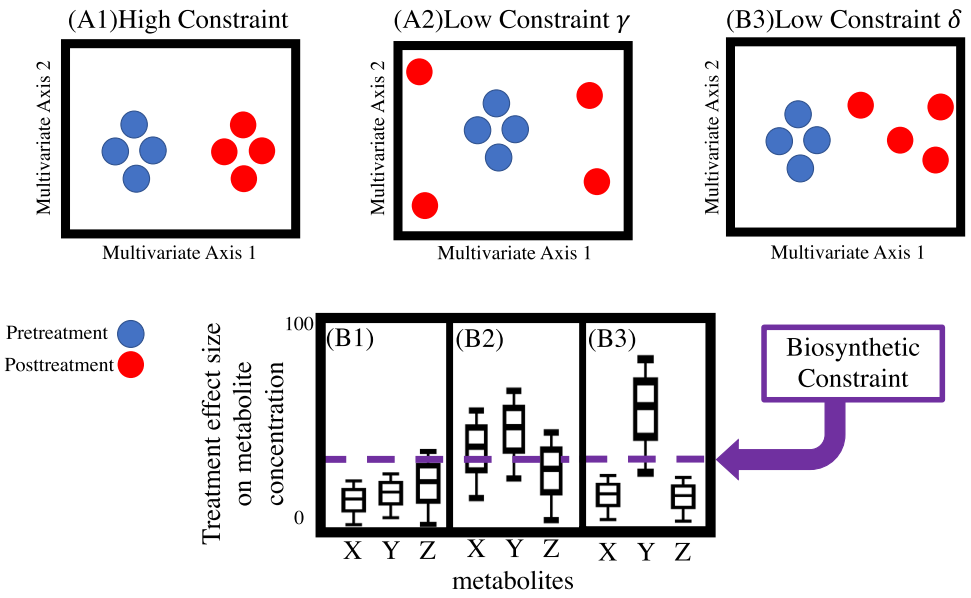

Fig. 4 Conceptual biosynthetic constraint scenarios. Axes in A represent 2 axes of multivariate trait space. Under high biosynthetic constraint (A1 and B1), we expect profiles to shift uniformly, while treatment effects on variation observed among compounds remain below what is expected due to shared biosynthetic enzymes or physiochemical properties. In low constraint (A2 and B2), we expect regulation occurring closer to the terminal ends of all metabolites, corresponding with treatment effects than expected due to shared properties. A second scenario of low constraint (A3 and B3) represents independent regulation of a single compound, near the terminal end of its biosynthesis combined with the stimulation of the base of a biosynthetic pathway. We expect a uniform directional shift of varying magnitudes, and the treatment effect on a single metabolite would be above that expected due to shared properties

we demonstrate regulation occurring closer to the terminal ends of all metabolites; thus, the treatment effect on variance is higher than what would be expected due to shared properties. In a visual sense, one would expect a general 'spreading' of data points in multivariate trait space, being individual metabolite profiles, in comparison with more concentrated pretreatment observations (Fig. 4). In the alternative scenario of low constraint δ, independent regulation of a single compound near the terminal end of its biosynthesis is combined with the stimulation of the base of a biosynthetic pathway (Fig. 4). Visually we expect a uniform directional shift of variable magnitudes, and the treatment effect on a single metabolite would be above that expected due to biosynthetic constraint (Fig. 4).

Apart from assessing treatment effects, constraint can potentially assess the simplicity of response to a given treatment. The primary goals of many induction studies are to characterize the induced state, such that it is identifiable and replicable across conditions. Biosynthetic constraints supply a metric of metabolite interdependence. For example, if the goal of an experiment comparing two induction treatments is to elicit a stable post-treatment profile that consists of the treatment affecting all observed compounds. Then the treatment which exhibits the highest biosynthetic constraint could be described as eliciting a more simplistic response, such that the base of the biosynthetic pathway is being stimulated. Theoretically this is comparable to exposing a plant to jasmonic or salicylic acid to stimulate either pathway, with the goal of producing a 'stable' induced response phenotype. This increased constraint-state would result in reduction of overall variability among observed among compounds. In contrast, within a given treatment if variance attributed to shared properties is low or non-existent, then all compounds are acting statistically independently of their shared biosynthetic enzymes and physiochemical properties, which may or may not indicate complex or context dependent regulation.

Conclusions

As we begin to understand the role of biosynthetic constraint and the interdependent nature of metabolite biosynthesis, a necessary goal of further studies is the development of methods which allow for proper statistical treatment of metabolites in analytical frameworks (Barkman 2001; Van Dam and Poppy 2008; Barupal and Fiehn 2017). The implications of enzymatic promiscuity and non-enzymatic reactions remain an understudied topic, notably absent from many plant chemical ecology and comparative biochemistry theories (Weng and Noel 2012; Weng 2014; Keller et al. 2015; Boachon et al. 2018). While the method demonstrated here has by no means solved the problem of non-independence within metabolite datasets, we here provide and validate the use of freely available and easily accessible data on physiochemical properties of metabolites which can quantify a measure of biosynthetic constraint. In the future, incorporating metabolomic networks as dependency structures and incorporating machine learning algorithms to identify important physiochemical properties may help to improve and refine the quantification of biosynthetic constraints. In addition, providing inclusion–exclusion criteria for individual metabolites based on identification of biosynthetic constraint for hypothesis testing with non-independent metabolite-based datasets would benefit the field. Given a small window into the implications of non-canonical reactions and their effects on constraining chemo-diversity, we can begin to develop better theory that incorporates physiochemical dynamics into diversity assessments and comparative analyses.

Acknowledgements The authors wish to thank Pedro Quintana-Ascencio, Eric Goolsby, Robert R. Junker, and one anonymous reviewer for helpful comments on this manuscript.

Author contributions JAD designed the study, collected, and analyzed the data. JAD wrote the manuscript with input from CMM.

Funding J.A.D. was supported in part by the Bill and Melinda Gates Foundation Millennium Scholars predoctoral fellowship.

Data availability Executable code to access physiochemical property data from PubChem and assess biosynthetic constraint will be submitted to the Dryad digital repository and stored publicly on GitHub (https://github.com/jordandowell/PhysiochemicalBiosyntheticConstrait.git) upon acceptance.

Compliance with ethical standards

Conflict of interest The author(s) declare that they have no conflict of interest.

References

Barkman TJ (2001) Character coding of secondary chemical variation for use in phylogenetic analyses. Biochem Syst Ecol 29:1–20. https://doi.org/10.1016/S0305-1978(00)00031-4

Barupal DK, Fiehn O (2017) Chemical similarity enrichment analysis (ChemRICH) as alternative to biochemical pathway mapping for metabolomic datasets. Sci Rep 7:1–11. https://doi.org/10.1038/s41598-017-15231-w

Boachon B, Buell CR, Crisovan E et al (2018) Phylogenomic mining of the mints reveals multiple mechanisms contributing to the evolution of chemical diversity in Lamiaceae. Mol Plant 11:1084–1096. https://doi.org/10.1016/j.molp.2018.06.002

Borges RM, Bessière JM, Ranganathan Y (2013) Diel variation in fig volatiles across syconium development: making sense of scents. J Chem Ecol 39:630–642. https://doi.org/10.1007/s10886-013-0280-5

Cao Y, Charisi A, Cheng LC et al (2008) ChemmineR: a compound mining framework for R. Bioinformatics. https://doi.org/10.1093/bioinformatics/btn307

Costello Z, Martin HG (2018) A machine learning approach to predict metabolic pathway dynamics from time-series multiomics data. NPJ Syst Biol Appl 4:19. https://doi.org/10.1038/s41540-018-0054-3

D'Ari RD, Casadesu J (1998) Underground metabolism. Bio Essays 20:181–186

Dahlin I, Rubene D, Glinwood R, Ninkovic V (2018) Pest suppression in cultivar mixtures is influenced by neighbor-specific plant–plant communication. Ecol Appl 28:2187–2196. https://doi.org/10.1002/eap.1807

de Bello F, Berg MP, Dias ATC et al (2015) On the need for phylogenetic "corrections" in functional trait-based approaches. Folia Geobot 50:349–357. https://doi.org/10.1007/s12224-015-9228-6

Delépine B, Duigou T, Carbonell P, Faulon J-L (2018) RetroPath2.0: a retrosynthesis workflow for metabolic engineers. Metab Eng 45:158–170. https://doi.org/10.1016/j.ymben.2017.12.002

Dudareva N (2004) Biochemistry of plant volatiles. Plant Physiol 135:1893–1902. https://doi.org/10.1104/pp.104.049981

Dudareva N, Negre F, Nagegowda DA, Orlova I (2006) Plant volatiles: recent advances and future perspectives. CRC Crit Rev Plant Sci 25:417–440. https://doi.org/10.1080/07352680600899973

Everitt BS, Dunn G (2013) Multivariate data and multivariate statistics. Applied multivariate data analysis. Wiley, West Sussex, pp 1–8

Fried SD, Boxer SG (2017) Electric fields and enzyme catalysis. Annu Rev Biochem 86:387–415. https://doi.org/10.1146/annurev-biochem-061516-044432

Galili T (2015) dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. Bioinformatics. https://doi.org/10.1093/bioinformatics/btv428

Gang DR (2005) Evolution of flavors and scents. Annu Rev Plant Biol 56:301–325. https://doi.org/10.1146/annurev.arplant.56.032604.144128

Gross M, Lewinsohn E, Tadmor Y et al (2009) The inheritance of volatile phenylpropenes in bitter fennel (Foeniculum vulgare Mill. var. vulgare, Apiaceae) chemotypes and their distribution within the plant. Biochem Syst Ecol 37:308–316. https://doi.org/10.1016/j.bse.2009.05.007

Guha R, Rojas-Chertó M (2010) rcdk: integrating the CDK with R. Chem Inf Funct R, pp 1–17

Gutensohn M, Nguyen TTH, McMahon RD et al (2014) Metabolic engineering of monoterpene biosynthesis in tomato fruits via introduction of the non-canonical substrate neryl diphosphate. Metab Eng 24:107–116. https://doi.org/10.1016/J.YMBEN.2014.05.008

Guzmán GI, Utrilla J, Nurk S et al (2015) Model-driven discovery of underground metabolic functions in *Escherichia coli*. Proc Natl Acad Sci 112:929–934. https://doi.org/10.1073/pnas.1414218112

Guzmán GI, Sandberg TE, Lacroix RA et al (2018) Enzyme promiscuity shapes evolutionary innovation and optimization. bioRxiv. https://doi.org/10.1101/310946

Hoffmeister M, Junker RR (2017) Herbivory-induced changes in the olfactory and visual display of flowers and extrafloral nectaries affect pollinator behavior. Evol Ecol 31:269–284. https://doi.org/10.1007/s10682-016-9875-y

Hoffmeister M, Wittköpper N, Junker RR (2016) Herbivore-induced changes in flower scent and morphology affect the structure of flower–visitor networks but not plant reproduction. Oikos 125:1241–1249. https://doi.org/10.1111/oik.02988

Jaeger DM, Runyon JB, Richardson BA (2016) Signals of speciation: volatile organic compounds resolve closely related sagebrush taxa, suggesting their importance in evolution. New Phytol 211:1393–1401. https://doi.org/10.1111/nph.13982

Junker RR (2018) A biosynthetically informed distance measure to compare secondary metabolite profiles. Chemoecology 28:29–37. https://doi.org/10.1007/s00049-017-0250-4

Junker RR, Kuppler J, Amo L et al (2018) Covariation and phenotypic integration in chemical communication displays: biosynthetic constraints and eco-evolutionary implications. New Phytol 220:739–749. https://doi.org/10.1111/nph.14505

Karban R (2017) Plant communication increases heterogeneity in plant phenotypes and herbivore movement. Funct Ecol 31:990–991. https://doi.org/10.1111/1365-2435.12806

Keller MA, Piedrafita G, Ralser M (2015) The widespread role of non-enzymatic reactions in cellular metabolism. Curr Opin Biotechnol 34:153–161. https://doi.org/10.1016/j.copbio.2014.12.020

Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem 79:471–505. https://doi.org/10.1146/annurev-biochem-030409-143718

Killiny N, Jones SE (2017) Profiling of volatile organic compounds released from individual intact juvenile and mature citrus leaves. J Plant Physiol 208:47–51. https://doi.org/10.1016/j.jplph.2016.11.001

Kim S, Chen J, Cheng T et al (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 47:D1102–D1109. https://doi.org/10.1093/nar/gky1033

Klamt S, Müller S, Regensburger G, Zanghellini J (2018) A mathematical framework for yield (vs. rate) optimization in constraint-based modeling and applications in metabolic engineering. Metab Eng 47:153–169. https://doi.org/10.1016/j.ymben.2018.02.001

Kruger NJ, Ratcliffe RG (2012) Pathways and fluxes: exploring the plant metabolic network. J Exp Bot 63:2243–2246. https://doi.org/10.1093/jxb/ers073

Kumar S, Rowan D, Hunt M et al (2015) Genome-wide scans reveal genetic architecture of apple flavour volatiles. Mol Breed. https://doi.org/10.1007/s11032-015-0312-7

Kunimoto R, Vogt M, Bajorath J (2017) Tracing compound pathways using chemical space networks. Medchemcomm 8:376–384. https://doi.org/10.1039/C6MD00628K

Lanzeni S, Messina E, Archetti F (2008) Graph models and mathematical programming in biochemical network analysis and metabolic engineering design. Comput Math with Appl 55:970–983. https://doi.org/10.1016/j.camwa.2006.12.101

Laothawornkitkul J, Taylor JE, Paul ND, Hewitt CN (2009) Biogenic volatile organic compounds in the Earth system. New Phytol 183:27

Liu J, Litt L, Segal MR et al (2011) Metabolomics of oxidative stress in recent studies of endogenous and exogenously administered intermediate metabolites. Int J Mol Sci 12:6469–6501. https://doi.org/10.3390/ijms12106469

Micallef SA, Shiaris MP, Colón-Carmona A (2009) Influence of arabidopsis thaliana accessions on rhizobacterial communities and natural variation in root exudates. J Exp Bot 60:1729–1742. https://doi.org/10.1093/jxb/erp053

Moles AT, Bonser SP, Poore AGB et al (2011) Assessing the evidence for latitudinal gradients in plant defence and herbivory. Funct Ecol 25:380–388. https://doi.org/10.1111/j.1365-2435.2010.01814.x

Nam H, Lewis NE, Lerman JA et al (2012) Network context and selection in the evolution to enzyme specificity. Science (80–) 337:1101–1104. https://doi.org/10.1126/science.1216861

Noctor G, Lelarge-Trouverie C, Mhamdi A (2015) The metabolomics of oxidative stress. Phytochemistry 112:33–53. https://doi.org/10.1016/j.phytochem.2014.09.002

Notebaart RA, Kintses B, Feist AM, Papp B (2018) Underground metabolism: network-level perspective and biotechnological potential. Curr Opin Biotechnol 49:108–114. https://doi.org/10.1016/J.COPBIO.2017.07.015

Oksanen J, Guillaume Blanchet F, Friendly M, Roeland Kindt F, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHM, Szoecs E, Wagner H (2019) Vegan: community ecology package. R package version 2.5–6. https://CRAN.R-project.org/package=vegan

Patti GJ, Yanes O, Siuzdak G (2012) Innovation: metabolomics: the apogee of the omics trilogy. Nat Rev Mol Cell Biol 13:263–269. https://doi.org/10.1038/nrm3314

Peres-Neto PR, Legendre P, Dray S, Borcard D (2006) Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87:2614–2625. https://doi.org/10.1890/0012-9658(2006)87[2614:VPOSDM]2.0.CO;2

Piedrafita G, Keller M, Ralser M (2015) The Impact of non-enzymatic reactions and enzyme promiscuity on cellular metabolism during (oxidative) stress conditions. Biomolecules 5:2101–2122. https://doi.org/10.3390/biom5032101

Prieto-Benítez S, Millanes AM, Dötterl S, Giménez-Benavides L (2016) Comparative analyses of flower scent in Sileneae reveal a contrasting phylogenetic signal between night and day emissions. Ecol Evol 6:7869–7881. https://doi.org/10.1002/ece3.2377

R Core Team (2019) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rowan DD, Hunt MB, Alspach PA et al (2009) Heritability and genetic and phenotypic correlations of apple (malus x domestica) fruit volatiles in a genetically diverse breeding population. J Agric Food Chem 57:7944–7952. https://doi.org/10.1021/jf901359r

Rubin IN, Ellner SP, Kessler A, Morrell KA (2015) Informed herbivore movement and interplant communication determine the effects of induced resistance in an individual-based model. J Anim Ecol 84:1273–1285. https://doi.org/10.1111/1365-2656.12369

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. https://doi.org/10.1007/s00442-016-3629-y

Schrader J, Unsicker SB, Bhattacharya S, Mummenhoff K (2017) Growth form rather than phylogenetic relationship predicts broad volatile emission patterns in the Brassicaceae. Plant Syst Evol 303:653–662. https://doi.org/10.1007/s00606-017-1397-8

Segler MHS, Preuss M, Waller MP (2018) Planning chemical syntheses with deep neural networks and symbolic AI. Nature 555:604–610. https://doi.org/10.1038/nature25978

Singh P, Kalunke RM, Giri AP (2015) Towards comprehension of complex chemical evolution and diversification of terpene and phenylpropanoid pathways in *Ocimum* species. RSC Adv 5:106886–106904. https://doi.org/10.1039/c5ra16637c

Thaler JS, Humphrey PT, Whiteman NK (2012) Evolution of jasmonate and salicylate signal crosstalk. Trends Plant Sci 17:260–270. https://doi.org/10.1016/j.tplants.2012.02.010

Uesugi A, Morrell K, Poelman EH et al (2016) Modification of plant-induced responses by an insect ecosystem engineer influences the colonization behaviour of subsequent shelter-users. J Ecol 104:1096–1105. https://doi.org/10.1111/1365-2745.12587

Van Dam NM, Poppy GM (2008) Why plant volatile analysis needs bioinformatics—detecting signal from noise in increasingly complex profiles. Plant Biol 10:29–37. https://doi.org/10.1055/s-2007-964961

Vieira DDSS, Emiliani G, Michelozzi M et al (2016) Polyploidization alters constitutive content of volatile organic compounds (VOC) and improves membrane stability under water deficit in Volkamer lemon (*Citrus limonia Osb.*) leaves. Environ Exp Bot 126:1–9. https://doi.org/10.1016/j.envexpbot.2016.02.010

Watson IA, Wang J, Nicolaou CA (2019) A retrosynthetic analysis algorithm implementation. J Cheminform 11:1. https://doi.org/10.1186/s13321-018-0323-6

Weng JK (2014) The evolutionary paths towards complexity: a metabolic perspective. New Phytol 201:1141–1149. https://doi.org/10.1111/nph.12416

Weng J-K, Noel JP (2012) The remarkable pliability and promiscuity of specialized metabolism. Cold Spring Harb Symp Quant Biol 77:309–320. https://doi.org/10.1101/sqb.2012.77.014787

Willighagen EL, Mayfield JW, Alvarsson J et al (2017) The chemistry development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminform 9:33. https://doi.org/10.1186/s13321-017-0220-4

Yip EC, De Moraes CM, Mescher MC, Tooker JF (2017) The volatile emission of a specialist herbivore alters patterns of plant defence, growth and flower production in a field population of goldenrod. Funct Ecol 31:1062–1070. https://doi.org/10.1111/1365-2435.12826