the plant journal

SEB SOCIETY FOR EXPERIMENTAL BIOLOGY

The Plant Journal (2024) **120,** 406–419 doi: 10.1111/tpj.16906

TECHNICAL ADVANCE

Mapping of specialized metabolite terms onto a plant phylogeny using text mining and large language models

Lucas Busta[1,*], Drew Hall[2,†], Braidon Johnson[1], Madelyn Schaut[2], Caroline M. Hanson[2,‡], Anika Gupta[2,§], Megan Gundrum[2], Yuer Wang[2] and Hiroshi A. Maeda[2,*]

[1] *Department of Chemistry and Biochemistry, University of Minnesota Duluth, Duluth, Minnesota, USA,*

[2] *Department of Botany, University of Wisconsin-Madison, Madison, Wisconsin, USA*

Received 16 February 2024; revised 21 May 2024; accepted 19 June 2024; published online 8 July 2024.

*For correspondence (e-mail bust0037@d.umn.edu; maeda2@wisc.edu)

[†] Present address: Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada

[‡] Present address: Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA

[§] Present address: Molecular Engineering and Sciences Institute, University of Washington, Seattle, Washington, USA

SUMMARY

Plants produce a staggering array of chemicals that are the basis for organismal function and important human nutrients and medicines. However, it is poorly defined how these compounds evolved and are distributed across the plant kingdom, hindering a systematic view and understanding of plant chemical diversity. Recent advances in plant genome/transcriptome sequencing have provided a well-defined molecular phylogeny of plants, on which the presence of diverse natural products can be mapped to systematically determine their phylogenetic distribution. Here, we built a proof-of-concept workflow where previously reported diverse tyrosine-derived plant natural products were mapped onto the plant tree of life. Plant chemical-species associations were mined from literature, filtered, evaluated through manual inspection of over 2500 scientific articles, and mapped onto the plant phylogeny. The resulting "phylochemical" map confirmed several highly lineage-specific compound class distributions, such as betalain pigments and Amaryllidaceae alkaloids. The map also highlighted several lineages enriched in dopamine-derived compounds, including the orders Caryophyllales, Liliales, and Fabales. Additionally, the application of large language models, using our manually curated data as a ground truth set, showed that post-mining processing can largely be automated with a low false-positive rate, critical for generating a reliable phylochemical map. Although a high false-negative rate remains a challenge, our study demonstrates that combining text mining with language model-based processing can generate broader phylochemical maps, which will serve as a valuable community resource to uncover key evolutionary events that underlie plant chemical diversity and enable system-level views of nature's millions of years of chemical experimentation.

Keywords: technical advance, plant natural products, biochemical evolution, tyrosine, phylochemistry, artificial intelligence.

INTRODUCTION

Plants synthesize a vast array of chemical compounds that are essential for their own physiological processes and adaptation to the ecological niches they inhabit. These plant natural products also serve as vital nutrients, flavors, antioxidants, and neurotransmitters for other organisms. To date, researchers have identified over 200 000 such compounds, with estimates suggesting the existence of more than a million distinct chemicals in the more than 300 000 species that make up the plant kingdom (Afendi et al., 2012; Hartmann, 2007). Importantly, about one third of all pharmaceuticals approved in the last 35 years are obtained from plant-derived substances, either directly or as synthetic derivatives inspired by them (Leonard et al., 2009; Newman & Cragg, 2016). Furthermore, roughly two-thirds of the global population depends on plants as primary medicinal resources (McChesney et al., 2007; Wu & Chappell, 2008). Thus, while synthetic combinatorial chemistry has produced extensive chemical libraries for drug discovery, the intricate and specialized nature of plant-based compounds, honed by evolutionary processes, render them a treasured source for new pharmaceutical developments (Awan et al., 2016; Mitchell, 2011; Walsh & Fischbach, 2010) and underscore the need for expanded exploration of natural products across the plant tree of life.

Plant natural products are largely produced by complex networks of metabolic pathways that evolve via various mechanisms including gene duplication, enzyme neofunctionalization, and altered gene expression (Maeda & Fernie, 2021; Moghe & Last, 2015; Schenck & Busta, 2022; Weng et al., 2012). Many of these plant pathways and their products occur in lineage-specific patterns and thus are referred to as lineage-specific or specialized metabolism, as compared to primary metabolism which is conserved across most organisms. The taxonomic distributions of plant chemicals have been an active area of study for decades, with contributions coming from diverse research areas including chemotaxonomy, natural products research, botany, drug discovery, and many others. Thus, though the distribution of plant chemicals is complex, a large amount of information is available about the detection of various chemicals in different plant species. However, this wealth of information was collected by many different investigators using diverse approaches (including chromatography, mass spectrometry, nuclear magnetic resonance, infrared spectroscopy, etc.), and is distributed across a large number of publications.

Recent studies integrating molecular phylogenies with metabolite analyses have showcased the value of exploring chemical distributions across the plant tree of life (Defossez et al., 2021; Ernst et al., 2019; Zust et al., 2020). These analyses have focused on either a broad range of species with a targeted set of metabolites, or a broad range of metabolites across a specific group of species. Such targeted analyses have, for example, helped predict chemical diversity hotspots (Fahmi et al., 2023), helped us understand some of the genetic and evolutionary basis for specialized metabolism (Fan et al., 2017; Lichman et al., 2020), and uncovered variant metabolisms with industrial or agricultural applications (Wang et al., 2017). These outcomes alone are exciting but also hint at potentially transformative outcomes that could result from broader analyses of plant chemical distributions across the tree of life, i.e., by incorporating wide ranges of both metabolites and species. The ingredients for such untargeted analyses are already available in the form of (i) massive molecular phylogenies enabled by recent advances in sequencing technology and (ii) decades of literature documenting the presence of lineage-specific plant chemicals. Unfortunately, a bottleneck exists in the highly time consuming and laborious tasks of systematically extracting compound–species associations from the literature and documenting them in a way that enables them to be mapped onto a phylogeny. By overcoming this bottleneck, we can gain a much more holistic perspective of plant chemical diversity that will help us (i) understand the evolutionary history of diverse natural products and underlying metabolic innovations, (ii) conduct targeted searches for new bioactivities, and (iii) derive novel hypotheses and design experiments aimed at understanding the physiological and ecological functions of plant specialized metabolites.

Several other groups have worked toward systematically extracting compound–species associations from literature. These efforts have led to several natural product databases, including Duke's Plant Natural Product, KNAPSAcK, the Dictionary of Natural Products, and NAPRALERT (Afendi et al., 2012; Duke, 2023; Loub et al., 1895). These databases are excellent tools for targeted searches and provide detailed information about bioactivities, mass spectral signatures, physical properties of chemicals, and in some cases compound–species associations. Nonetheless, these resources are designed for targeted searches for a specific compound(s) or a specie(s) of interest and not for untargeted analyses that require comprehensive phylogenetic mapping of plant compounds. Some databases do make compound–species association data available in an untargeted fashion, but are either not open access (Elsevier's Reaxys) or are metadatabases that lack a systematic mechanism for expansion and updating, like LOTUS (Rutz et al., 2022). To bridge this gap, we have developed a novel workflow (Figure 1) for creating a "phylochemical" database in which chemical diversity is organized in the context of a plant phylogeny and all data are simultaneously accessible. This workflow consists of (i) text mining to identify potential associations between chemical compounds and specific plant species, (ii) evaluation of those associations, and (iii) mapping of positive hits on molecular phylogeny to populate a phylochemical database (Figure 1). Here, we first used text mining and manual curation to identify and evaluate over 3000 putative associations between specific plant species and specific tyrosine-derived compounds. We then analyzed the distribution of the tyrosine-derived compounds across various plant lineages to gain insights into their specialized metabolic pathways. Finally, we tested the potential of large language models to complement our text mining methods, revealing their effectiveness in processing and accurately categorizing candidate compound–species associations. Overall, our study demonstrates that a workflow combining text mining with language model-based processing can open the bottleneck preventing the construction of broader phylochemical maps that span large portions of both chemical and plant diversity.

RESULTS

Text mining and manual curation to identify compound–species associations

Our first aim was to identify previously reported associations between tyrosine-derived compounds and specific

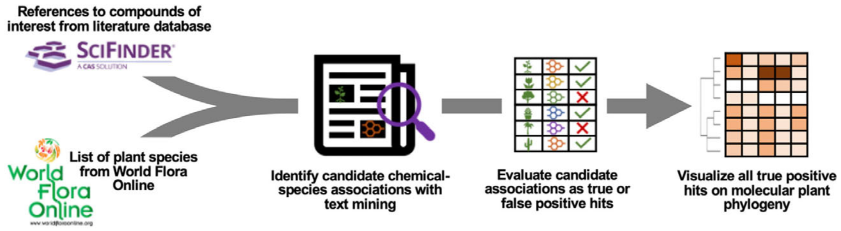
References to compounds of interest from literature database
SciFinder
A CAS SOLUTION
List of plant species from World Flora Online
World Flora Online
Identify candidate chemical-species associations with text mining
Evaluate candidate associations as true or false positive hits
Visualize all true positive hits on molecular plant phylogeny

Figure 1. Overview of the workflow for identifying compound–species associations.
From left to right, the workflow consists of gathering relevant articles from a literature database together with a list of accepted plant species names, then applying text mining to identify co-occurrences of the names of chemical compounds and chemical species in the titles or abstracts. Next, the co-occurrences (candidate compound–species associations) are evaluated to isolate only associations that are supported with experimental evidence. Finally, the supported compound–species associations are mapped on to a phylogeny.

plant species. Tyrosine-derived metabolites were useful in developing a phylochemical mapping workflow because some tyrosine-derived compounds are broadly distributed while others are highly lineage-specific (Schenck & Maeda, 2018). The latter include betalains and Amaryllidaceae alkaloids, which serve as internal positive controls. We began by compiling a list of 71 tyrosine-derived plant natural products, based on prior literature and the KEGG plant chemical compound list (Kanehisa et al., 2006; Schenck & Maeda, 2018), and grouped them based on their upstream precursor molecules or commonly used compound class names (e.g., "dopamine-derived compounds", "Amaryllidaceae alkaloids", etc., Table S1). We then used the Chemical Abstracts Service (CAS) SciFinder database (Gabrielson, 2018) to retrieve references, including titles and abstracts, for articles that mentioned the names of the 71 tyrosine-derived natural products. Subsequently, we filtered the SciFinder search results to include only the articles that were associated with the keyword "plant" or "taxonomy" in the SciFinder database, thereby eliminating a large number of clinical reports, which were likely irrelevant to our plant-derived chemical focus. Even after filtering, eight compounds, including epinephrine (adrenaline), noradrenaline, dopamine, and colchicine had over 10 000 hits (Figure S1), most of which appeared to be due to human clinical studies of these widely used pharmaceuticals. It seemed likely that many of the co-occurrences of these compounds and plant species names in the text of these articles would not reflect endogenous production in plants, so we excluded these compounds from further analysis. The references for the remaining 63 compounds were then filtered for articles that were written in English and had titles or abstracts mentioning at least one plant species name included in The World Flora Online (WFO) Plant List—an open access resource for the most comprehensive plant list maintained by the global community of taxonomy experts (https://wfoplantlist.org/plant-list). Overall, this process resulted in 2571 pertinent references, for which titles and abstracts were downloaded for further analysis.

In the next phase of our study, we assessed the accessibility of full texts for the 2571 references using PubChem, hosted by the National Center of Biotechnology Information (NCBI). Unfortunately, full texts were available via NCBI for only approximately 300 articles, with many of the other articles' full texts being located behind paywalls. Consequently, we first turned our attention to examining the titles and abstracts of the articles for compound–species associations using automated, pattern matching-based text mining. We recorded any occurrence of a taxon name (genus and species names together) within articles that mentioned one of the 63 tyrosine-derived compounds as an indicator of potential presence of the compound in that taxon, yielding 3631 candidate associations (Table S2). To integrate these findings with phylogenetic data, we mapped the associations onto a phylogeny of orders within angiosperms (Qian & Jin, 2016). Branches without reported compounds were pruned for clearer visualization (Figure 2A). To account for the varying species counts per order and the disparate number of articles for each compound, we introduced a normalized metric that highlights a compound's lineage specificity. This metric reflects the total number of reports for a compound per order, divided both by the total number of reports of that compound and by the species count within that order. The resulting phylochemical map showed the expected lineage-specific occurrences of the positive controls (betalains and Amaryllidaceae alkaloids in Caryophyllales and Asparagales orders, respectively), among other patterns (Figure 2B). This result demonstrates the ability of text mining to extract compound–species associations and capture their overall phylogenetic distributions.

To assess the reliability of our text mining approach, each of the 3631 candidate compound–species associations identified were manually reviewed. The review involved manually downloading and searching the full texts, figures, and tables of the 2571 articles to determine whether each association was a true positive (indicating the compound was indeed present in the species) or a false positive (where

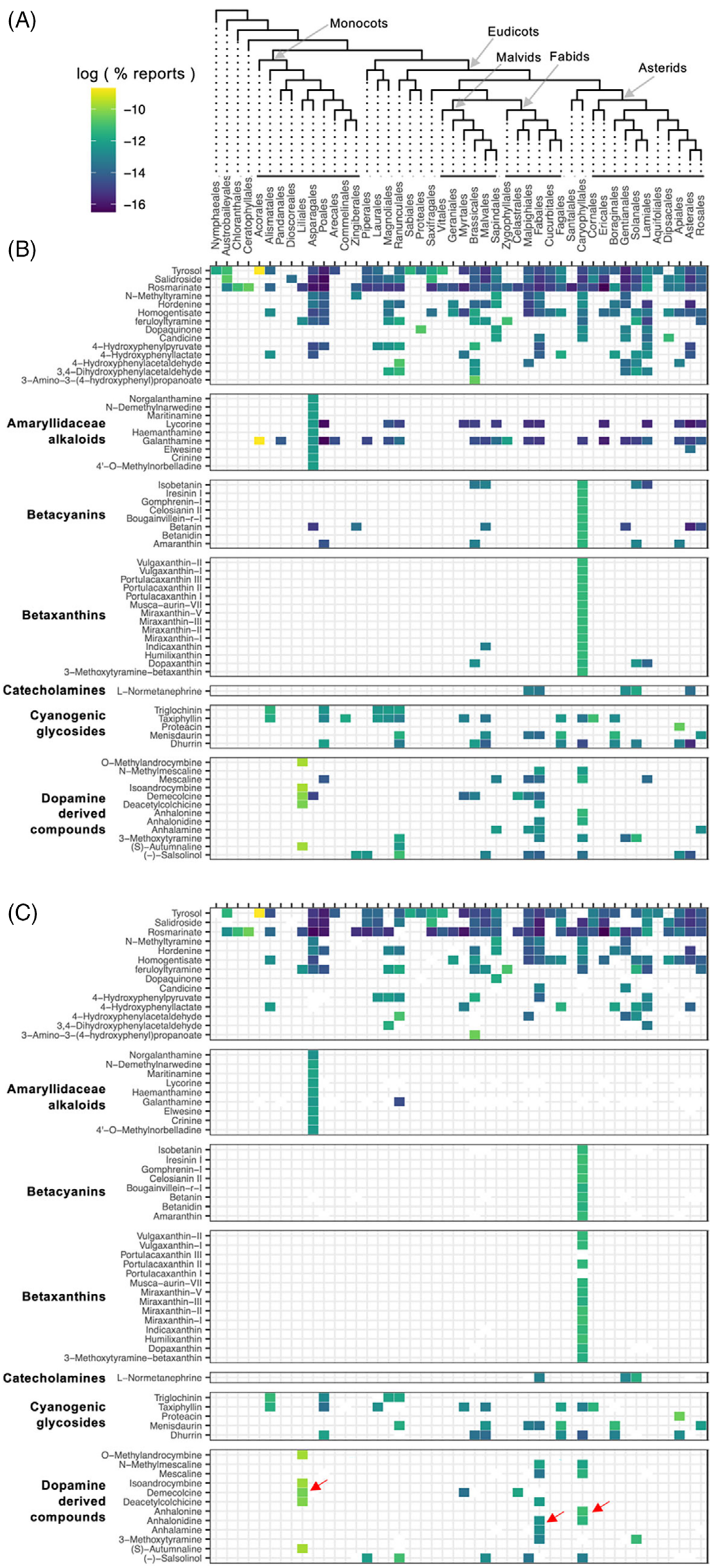
(A)
log (% reports)
−10
−12
−14
−16
Monocots
Eudicots
Malvids
Fabids
Asterids
Nymphaeales
Austrobaileyales
Chloranthales
Ceratophyllales
Acorales
Alismatales
Pandanales
Dioscoreales
Liliales
Asparagales
Poales
Arecales
Commelinales
Zingiberales
Piperales
Laurales
Magnoliales
Ranunculales
Sabiales
Proteales
Saxifragales
Vitales
Geraniales
Myrtales
Brassicales
Malvales
Sapindales
Zygophyllales
Celastrales
Malpighiales
Fabales
Cucurbitales
Fagales
Santalales
Caryophyllales
Cornales
Ericales
Boraginales
Gentianales
Solanales
Lamiales
Aquifoliales
Dipsacales
Apiales
Asterales
Rosales
(B)
Tyrosol
Salidroside
Rosmarinate
N−Methyltyramine
Hordenine
Homogentisate
feruloyltyramine
Dopaquinone
Candicine
4−Hydroxyphenylpyruvate
4−Hydroxyphenyllactate
4−Hydroxyphenylacetaldehyde
3,4−Dihydroxyphenylacetaldehyde
3−Amino−3−(4−hydroxyphenyl)propanoate
Amaryllidaceae alkaloids
Norgalanthamine
N−Demethylnarwedine
Maritinamine
Lycorine
Haemanthamine
Galanthamine
Elwesine
Crinine
4'−O−Methylnorbelladine
Betacyanins
Isobetanin
Iresinin I
Gomphrenin−I
Celosianin II
Bougainvillein−r−I
Betanin
Betanidin
Amaranthin
Betaxanthins
Vulgaxanthin−II
Vulgaxanthin−I
Portulacaxanthin III
Portulacaxanthin II
Portulacaxanthin I
Musca−aurin−VII
Miraxanthin−V
Miraxanthin−III
Miraxanthin−II
Miraxanthin−I
Indicaxanthin
Humilixanthin
Dopaxanthin
3−Methoxytyramine−betaxanthin
Catecholamines
L−Normetanephrine
Cyanogenic glycosides
Triglochinin
Taxiphyllin
Proteacin
Menisdaurin
Dhurrin
Dopamine derived compounds
O−Methylandrocymbine
N−Methylmescaline
Mescaline
Isoandrocymbine
Demecolcine
Deacetylcolchicine
Anhalonine
Anhalonidine
Anhalamine
3−Methoxytyramine
(S)−Autumnaline
(−)−Salsolinol
(C)
Tyrosol
Salidroside
Rosmarinate
N−Methyltyramine
Hordenine
Homogentisate
feruloyltyramine
Dopaquinone
Candicine
4−Hydroxyphenylpyruvate
4−Hydroxyphenyllactate
4−Hydroxyphenylacetaldehyde
3,4−Dihydroxyphenylacetaldehyde
3−Amino−3−(4−hydroxyphenyl)propanoate
Amaryllidaceae alkaloids
Norgalanthamine
N−Demethylnarwedine
Maritinamine
Lycorine
Haemanthamine
Galanthamine
Elwesine
Crinine
4'−O−Methylnorbelladine
Betacyanins
Isobetanin
Iresinin I
Gomphrenin−I
Celosianin II
Bougainvillein−r−I
Betanin
Betanidin
Amaranthin
Betaxanthins
Vulgaxanthin−II
Vulgaxanthin−I
Portulacaxanthin III
Portulacaxanthin II
Portulacaxanthin I
Musca−aurin−VII
Miraxanthin−V
Miraxanthin−III
Miraxanthin−II
Miraxanthin−I
Indicaxanthin
Humilixanthin
Dopaxanthin
3−Methoxytyramine−betaxanthin
Catecholamines
L−Normetanephrine
Cyanogenic glycosides
Triglochinin
Taxiphyllin
Proteacin
Menisdaurin
Dhurrin
Dopamine derived compounds
O−Methylandrocymbine
N−Methylmescaline
Mescaline
Isoandrocymbine
Demecolcine
Deacetylcolchicine
Anhalonine
Anhalonidine
Anhalamine
3−Methoxytyramine
(S)−Autumnaline
(−)−Salsolinol

Figure 2. Map of tyrosine-derived compounds across the angiosperm phylogeny.
(A) Cladogram of angiosperm orders. We mapped compound–species associations onto a previously published angiosperm order phylogeny (Qian & Jin, 2016), removing branches where no compounds were detected for better clarity.
(B) All putative compound–species associations.
(C) All true positives. In B and C, to adjust for differences in species count per order and the number of articles for each compound, we used a normalized metric to color the heat map. This metric is the log of the number of reports for a compound in each order, divided by the total reports of that compound and by the species count in the order, thus emphasizing the compound's specificity to certain lineages. White boxes in C represent associations that were present in B but were all false positives and thus do not appear in the true positive map.

the compound and species co-occurred in the text without evidence of the species producing the compound). We considered instances in which the source article reported the pipeline-identified compound–species association either in the abstract, the main text, or in a figure/table as true positives. Our manual evaluation revealed that, of the 3631 associations detected via text mining, 2409 were true positives (66.3%, Figure 2C), 863 were false positives (23.7%, Figure S2), and 359 cases (10%) remained ambiguous due to unclear language or indeterminate figures. Overall, the revised map showing only these true positives (Figure 2C) contained very similar patterns to the original map (Figure 2B), though a number of associations for betalains and Amaryllidaceae alkaloids were eliminated outside of Caryophyllales and Asparagales.

During our manual evaluation of the literature, several patterns emerged that enabled us to distinguish several types of false positive associations. False positives often included: (i) non-original mentions of compounds without supporting data, (ii) discussions of biosynthetic pathways, (iii) mentions of enzyme assays related to a compound without reporting the compound itself, (iv) reports of a compound's absence in a species, and (v) references to compounds in a context of artificial presence or absence (Table S2). For example, instances of "non-original mention" were flagged when papers cited a compound without providing new evidence of its presence in the species. Similarly, "biosynthetic precursor/derivative" hits occurred when articles discussed related metabolic pathways or derivatives without confirming the compound's accumulation. For example, a study mentioned "homogentisic acid" while investigating its metabolic derivatives from *Miliusa balansae* but did not provide evidence of homogentisic acid accumulation (Huong et al., 2004). Some false positives also derived from sources reporting enzymes related to a given compound, such as a large number of papers referencing galantamine, an Amaryllidaceae alkaloid used to treat early-stage Alzheimer's disease (Ago et al., 2011), as a positive control in acetylcholinesterase inhibition assays using the extracts from various other plants (Mocan et al., 2018; Siatka et al., 2017). Interestingly, the text mining pipeline also produced false positive hits from papers describing the absence of a compound (for example, papers using targeted metabolomics approaches that reported "not detected"). Similarly, broad surveys of multiple species that reported the detection of multiple compounds resulted in false positives. For example, one study surveyed 10 Amaryllidaceae species for various compounds including galantamine in *Clivia miniata* (Cortes et al., 2018); however, galantamine was not detected in *C. miniata* but found in different species presented in the same table. Finally, some false positives resulted from studies using transgenic plants or artificial introduction of compounds to a heterologous species for the purposes of screening enzyme activity or inhibition. These examples and observations will help reduce false positive hits in future iterations of extracting true compound–species associations.

Exploration of tyrosine-derived compounds across the plant phylogeny

Having compared maps of all putative associations against maps of only true positive hits, we next examined patterns in the true positive map in detail and observed several surprising patterns. Unlike compounds with lineage-specific occurrence such as betalains and Amaryllidaceae alkaloids, we observed that other tyrosine-derived compounds were unexpectedly distributed across multiple lineages, despite them being generally considered as lineage-specific specialized metabolites. Some examples are the tyramine-derived compounds *N*-methyl-4-tyramine and hordenine, which were detected in 12 and 6 different plant orders, respectively, many of which were only distantly related to each other (e.g., Asparagales from monocots and Gentianales and Sapindales from eudicots; Figure 2).

To further experimentally validate the somewhat unexpected result described above, we conducted targeted chemical investigations of hordenine and related compounds (Figure 3A). We chose plant species that were available in the Botany Greenhouse at UW-Madison and belong to the same lineages reported to have hordenine, harvested live tissues, and analyzed metabolites using liquid chromatography-high-resolution mass spectrometry (LC-HRMS). We found that hordenine and its precursors (i.e., tyramine, *N*-methyl-4-tyramine) were indeed present in diverse lineages of plants; four orders (Poales, Sapindales, Gentianales, and Caryophyllales) with at least one species in each order testing positive for hordenine

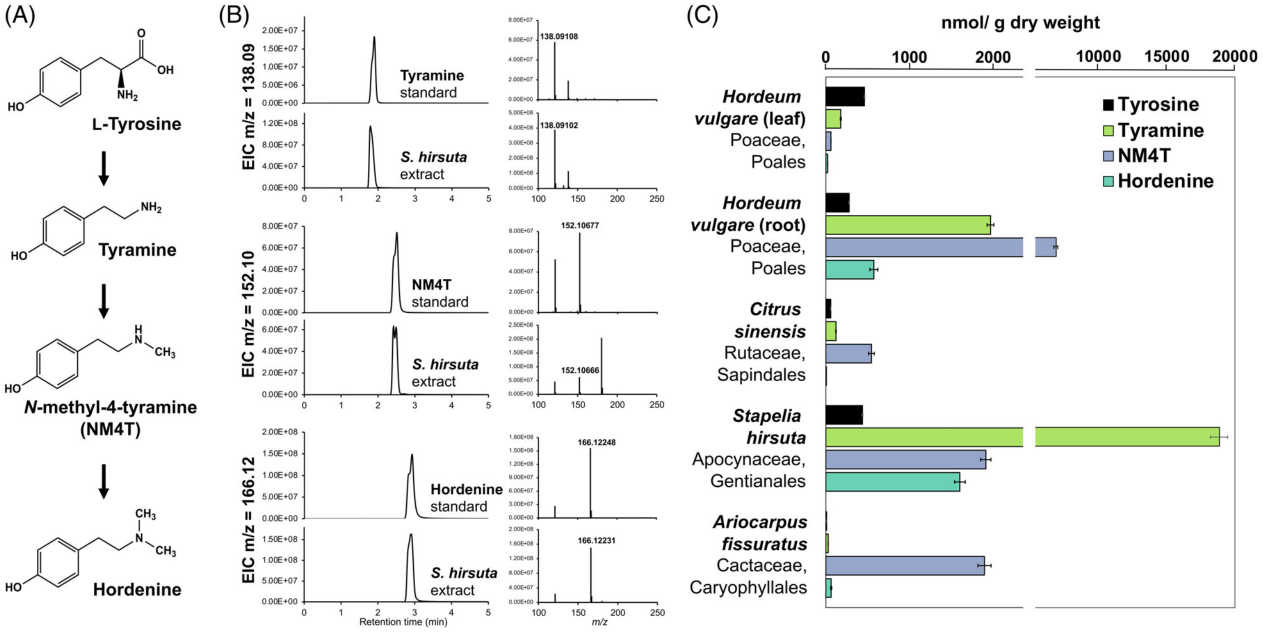

Figure 3. Experimental validation of hordenine and its precursors.
(A) Hordenine is synthesized from the aromatic amino acid L-tyrosine via tyramine and *N*-methyl-4-tyramine (NM4T) as intermediates.
(B) Metabolites were extracted from the leaf tissue of *Stapelia hirsuta* (Appocynaceae) and analyzed for the presence of hordenine and its upstream precursors using LC-HRMS by comparison with corresponding authentic standards (100 μM concentration). Mass searches were performed using the exact mass of each compound with an allowable error of 10 ppm at 5 decimal places, generating extracted ion current (EIC) traces shown.
(C) Further metabolite analysis of leaf and root tissues of *Hordeum vulgare* and the leaf tissues of other plant species showed that some non-Poaceae plants also accumulate comparable or even greater levels (nmol/g dry weight of plant tissues, $n = 3$ technical replicates) of hordenine and its precursors than *H. vulgare* (Poaceae).

production (Figure 3B). Some non-Poaceae species accumulated comparable or even higher levels of hordenine and its precursors than *Hordeum vulgare* (Poaceae, Figure 3C). These measurements showed that hordenine is not restricted to one plant family and illustrated the ability of the pipeline to provide unexpected, yet accurate phylogenetic distributions of certain chemicals.

Finally, the phylochemical map of tyrosine-derived compounds presented here also revealed high lineage specificity in the dopamine-derived compound class, with clusters in Caryophyllales, Liliales, and Fabales (Figure 2C, red arrows). For example, Liliales was reported to accumulate several dopamine-derived compounds, including the intermediates of colchicine biosynthesis, such as (*S*)-Autumnaline and Demecolcine. In addition, 3,4,5-Trimethoxyphenethylamine (mescaline), which is well known to be produced in peyote cactus (*Lophophora spp.*) in Cactaceae (Caryophyllales order) (Aragane et al., 2011; El-Seedi et al., 2005), was also detected in Fabales, specifically in the Fabaceae family, the *Acasia* genus within the mimosoid clade (Clement et al., 1997, 1998). Finally, both Caryophyllales and Fabales were found to produce other dopamine-derived compounds such as anhalonidine and anhalamine (Figure 2C, red arrows). These examples illustrate that the pipeline can highlight lineages with concentrated occurrences of a group of related chemicals, a potentially valuable tool in detecting examples of convergent metabolic evolution beyond those already reported (Pichersky & Lewinsohn, 2011).

The potential of large language models to complement text mining approaches

So far, our work indicated that using text mining to find co-occurrences of the names of plant chemicals and species could generate a large number of candidate compound–species associations (3631 in a set of tyrosine-derived queries). However, by manually inspecting each full text yielding these candidates, we found that 23.7% of hits were false positive associations, that is, the name of a given compound and species appeared close together in an article's text, but the article did not provide empirical data for the compound actually being detected in an extract from that species. Still, even with a high false positive rate, the text mining approach yielded useful and sometimes unexpected, yet accurate, information as described above. These outcomes supported the rationale for constructing larger phylochemical maps with a text mining approach but suggested that methods to reduce the false positive rate (other than laborious manual inspection) should be explored. Recent studies in natural

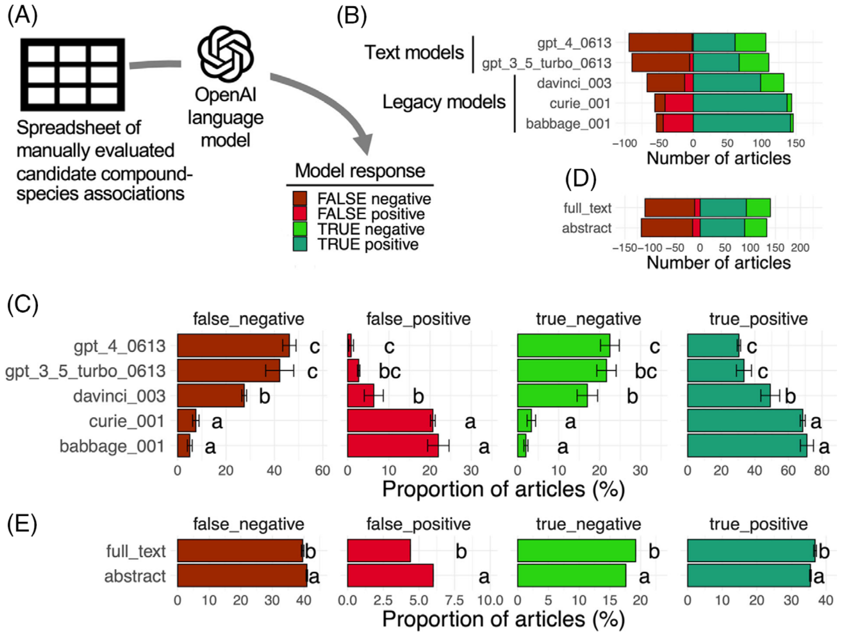

Figure 4. Language model-based evaluation of compound–species associations.
(A) Schematic of language model evaluation process.
(B, C) Comparison of OpenAI language models' ability to assess compound–species associations based on title and abstract text in terms of the number of articles (B) and proportion of articles (C).
(D, E) Comparison of compound–species association assessment abilities of gpt-3.5-turbo-0613 using only title and abstract text versus when full texts are also supplied in terms of number of articles (D) and proportion of articles (E). Note that the proportion of false positives and true negatives was identical between the independent samples, so the error bar sizes are zero. Bar heights and error bars represent the means and standard deviations of *n* = 3 independent tests ("replicates") conducted using random sampling, respectively. Letters ("a", "b", "c") above bars indicate significant differences as determined by ANOVA and post-hoc Tukey tests, *P* < 0.05. Some error bars in E are not visible due to constant proportions of false positive and true negatives observed among these replicate analyses.

language processing have highlighted the abilities of large language models to answer questions about the meaning of text strings, including text describing chemical and biological concepts (Dunn et al., 2022; Jablonka et al., 2023; Xiao et al., 2023), and such models are now readily accessible. Accordingly, we next evaluated the abilities of large language models to categorize candidate compound–species associations by using the manually curated dataset as a ground truth set (Figure 4A). If effective, such models could greatly enhance our ability to generate accurate, large-scale compound–species association maps directly from existing literature.

Of the large language models that are publicly available, we tested three of OpenAI's legacy models, babbage, currie, and davinci, as well as two text models, gpt-3.5-turbo-0613 and gpt-4-0613, because these were the most accessible and advanced models at the time of study. To test the models' abilities, we began by filtering our manually evaluated compounds-species associations for only those 3275 associations that could be distinguished as true positives (2409) or false positives (863) from manual inspection, thus eliminating associations whose validity was ambiguous. To test a language model's abilities, we created a query containing the text of an article's abstract and title, as well as a question: "Here is an article's title and its abstract <abstract>. Answer yes or no: is the compound <compound name> found in <taxon name>?". In a first round of queries, we randomly selected 200 putative associations from among the 3275 candidates and passed each association, as a separate query, to each model. We conducted a total of three rounds of queries using a different random selection each time. These random selections served as independent experiments (or "replicates") in this context. Then, we compared the models' responses with the results of our manual evaluation to categorize the models' responses as correct or incorrect (Figure 4A). Thus, our approach consisted of "question-answering" tasks performed by a language model, that is, we asked the language model to answer a presence/absence question based on the abstract of the peer-reviewed scientific article we provided. This type of question-answering task is one of the many tasks that language models are built to do at a fundamental level (for a larger list of language model tasks, please see huggingface.co/tasks). Our approach did not rely on the content of the original training data used by OpenAI (the model's "built in knowledge"), nor did we train (or "fine-tune") an existing model using our manually curated data on tyrosine-derived compounds.

Using the question-answering task approach, our first observation was that, compared with the legacy models, the text models (e.g., gpt-4-0613) yielded a substantially lower number of false positives (a negative association erroneously identified by the model as positive, <5% versus >5%). However, the lower false positive rate was accompanied by lower proportion of true positives (correctly identified positive associations; ~30% versus >30%; Tukey HSD test, $P < 0.05$, Bonferroni correction applied; Figure 4B,C). Relative to the legacy models, the text models also produced a higher number of negatives—both true negatives (correctly identified instances in which the co-occurrence of a compound name and a species name did *not* indicate the presence of that compound in that species) and false negatives (erroneous claims that the co-occurrence of a compound name and a species name did *not* indicate the presence of that compound in that species). At the risk of anthropomorphizing, one way to think about the increase in negatives is an increase in "skepticism" as model sophistication increases, or that the more sophisticated models are able to distinguish between textual co-occurrence and biological co-occurrence. Overall, our results indicate that language models, particularly the text models, can greatly reduce the proportion of false positives, which is critical for enhancing the overall quality of the resulting phylochemical mapping dataset, in an automated fashion.

In addition to comparing different language models, we examined whether providing the full text of an article improved the ability of a language model to classify candidate compound–species associations. For this test we chose to work with gpt-3.5-turbo-0613 for three reasons: (i) there were no significant differences in the performance of gpt-3.5-turbo-0613 and the most advanced model, gpt-4-0613, in our initial tests (Figure 4C), (ii) at the time of our experiments there were rate limits imposed on gpt-4-0613 usage (only 25 queries per hour), and (iii) the availability of gpt-3.5-turbo-16 k-0613, which was identical to gpt-3.5-turbo-0613 except that it had a 16 000 token context window that would enable working with full text articles. Using gpt-3.5-turbo-0613/gpt-3.5-turbo-16 k-0613, we again performed three rounds of queries. In each round, we randomly selected a set of 250 compound–species associations and constructed two types of queries for each – one that asked whether a given compound was found in the species of interest based on the title and abstract alone, and a second that asked the same question but also provided the full text in the query. When provided with full texts, the model was able to produce a slightly higher proportion of true negatives and true positives, as well as a slightly lower proportion of false negatives and false positives (Tukey HSD test, $P < 0.05$, Bonferroni correction applied; Figure 4D,E). In other words, providing full texts seems to result in a small but detectable improvement in language model-based classification of candidate compound–species associations. It could be that the improvements due to full texts were only modest because of the "lost in the middle" phenomenon in which language models (at least, at the time of our study) were prone to overlooking a specific fact provided to them if it was buried in the middle of a large corpus of text, as described at github.com/gkamradt/LLMTest_NeedleInAHaystack.

Overall, our work with the language models revealed them to be an excellent complement to text mining in building phylochemical maps. For the scientific community to utilize and further explore this approach, we provide the code we used to perform our language model assessments at https://github.com/thebustalab/phylochemical_mapping. We also provide raw data in the form of candidate compound–species associations with annotations, and true compound–species associations missed by the text mining pipeline (Table S2), as well as raw results from the language model evaluations and the code used to create the figures presented in this article at the link provided above. These open resources will allow other scientists to further expand the map and associate with other traits as discussed below.

DISCUSSION

In this study, we developed a proof-of-concept workflow for building maps of chemical occurrence across a plant phylogeny using text mining and downstream processing. By using tyrosine-derived compounds as test cases, we were able to validate that this approach can indeed identify the known presence of compounds in well-studied lineages and highlight previously unrecognized patterns of chemical distribution, illustrating the efficacy and power of our workflow. In addition, the application of large language models, combined with our manually curated dataset, showed that downstream evaluation after text mining is critical but can largely be automated in future iterations. Here, we discuss the limitations of our workflow and potential solutions, as well as future applications and how we can make more comprehensive phylochemical maps that capture larger bodies of literature and plant chemical diversity.

Limitations of the text mining workflow and potential solutions

Our manual curations (i) showed that the text mining identifies a high rate of false positive hits for compound–species associations (Figure S2) and (ii) identified several types of false positive associations. Notably, further application of language models to evaluate the potential associations, derived from the text-mining, produced a low rate of false positives, which is advantageous in generating reliable and accurate phylochemical maps. However, our results indicates that these models are quite conservative in their assessments and tend to accept very few incorrect compound–species associations, leading to the dismissal a considerable number of correct associations and hence a high rate of false negatives. Such filtering resulted in sacrificing the size of the output map for the sake of accuracy, leaving room for improvement in the sophistication of the evaluation portion of our workflow. At the time of the work reported here, OpenAI was the only sophisticated, API-accessible language model, and the pace of language model development was quite rapid. Accordingly, we opted not to benchmark the false positive rates of a broad array of language models with regard to the application described here, in part since such benchmarking would likely be obsolete within a short period of time. Nevertheless, one route for improving false positive rates could be to use techniques like chain of thought prompting (Wei et al., 2023), or dedicated techniques for the extraction of structured data from unstructured texts (Li et al., 2023). Alternatively, prior works have shown that performance gains in classification tasks can be achieved by fine tuning a language model using annotated examples (Bakker et al., 2022; Chae & Davidson, 2023; Howard & Ruder, 2018). Therefore, the evaluation portion of a future workflow could be enhanced with a similar approach using the several types of false negatives and false positives that we manually annotated in this study to create a fine-tuned classification model. Other approaches include a pipeline that combines the natural language processing abilities of language models with the classification abilities of machine learning models. Regardless of the approach used, it seems that it will be highly important for investigators to explore a variety of cutting-edge methods for optimizing compound–species extraction/mapping, since the newest models and prompting techniques developed by the natural language processing community are likely to have higher performance. Particular attention should be paid to open-source models, since they may be less expensive than closed source models when used at scale.

We also sought to evaluate the comprehensiveness of the text mining process itself by looking for relevant studies that our text mining method might have missed. Through manual searches using Google Scholar for each of the 63 tyrosine-derived compounds (excluding 8 compounds with very high number of medical articles, Figure S1), we discovered 291 overlooked articles (Table S3). These articles reported a total of 344 valid compound–species associations, which we again confirmed through manual inspection. However, these articles and associations were not identified through our text mining process (Table S2), likely because they were not included in the literature database (i.e., SciFinder) that we initially used to search for information on tyrosine-derived compounds. When we mapped these 344 newly found associations, we found similar overall trends compared to our existing text-mined dataset (Figure S3). Nevertheless, these results suggest that when we scale to cover more species and a broader range of chemical compounds, we need to expand the range of databases that we query to build more comprehensive phylochemical maps. Also, we need to always keep in mind that a lack of reports does not necessarily indicate the absence of a compound in a certain species.

Another limitation of the method described here was that certain compounds had a very high number of papers describing them (e.g., epinephrine, morphine, Figure S1). In our workflow, we elected not to map these compounds described in more than 10 000 papers, because many of the reports detailed medicinal uses rather than the presence of these compounds in specific plant species. In other words, the signal-to-noise ratio for widely studied medicinal compounds is low. Our decision to omit such compounds led to notable absences in our maps: for example, the absence of positive hits for isoquinoline alkaloids (such as codeine and morphine) in Ranunculales, an order that includes poppy plants and is well-known for producing these compounds (Bisset, 1985). This limitation could be addressed in future works by developing procedures to effectively map compounds that are the subject of a large volume of research, such as by eliminating only certain articles (e.g., those from medical journals) for these selected compounds or by training language models to identify and dismiss medical articles while preserving others. Alternative strategies could involve only mining certain portions of an article (its results but not introduction section, for example) when generating candidate compound–species associations.

Finally, a lack of sufficient reported data for some compounds may create gaps within the map and their underlying patterns of chemical distribution may be invisible (even if they exist). This situation can be compared to the concept of sequencing depth: without sufficient data, it is challenging to accurately identify genetic features. Furthermore, since phylochemical maps are two-dimensional, there is a parallel issue regarding phylogenetic resolution. In cases where the phylogenetic data is limited (e.g., certain plant families), even abundant chemical data might not reveal patterns of interest. While notable patterns of certain chemicals at the level of plant orders emerged in this study (Figure 2C), distributions at the levels of plant families or genera may require further species sampling and refinement in the molecular phylogeny. Therefore, our study also highlights the importance of continued data expansion of both axes—chemical and phylogenetic—within the scientific community to achieve adequate depth for generating useful maps.

Applications of the phylochemical database

One interesting finding from our mapping efforts was the presence of dopamine-derived compounds in Caryophyllales, Liliales, and Fabales (Figure 2C). These results raise the possibility that these distant lineages somehow independently acquired highly active upstream pathways to produce elevated amounts or availability of the dopamine precursor. Prior studies showed that Caryophyllales and Fabaceae within Fabales have highly active biosynthetic pathways of tyrosine, an upstream precursor of dopamine (Coley et al., 2019; Lopez-Nieves et al., 2018, 2022). The current finding suggests that Liliales may also have altered regulation of the upstream precursor supply pathway, leading to enriched occurrence of dopamine-derived compounds. This demonstrates one example where phylochemical mapping can provide a novel hypothesis in an unbiased or untargeted manner, which could be further evaluated experimentally in future works. Similarly, a larger, more comprehensive phylochemical database developed in the future, could be used to isolate data from certain plant orders or families and assess the characteristics of certain regions of metabolism within that lineage, which will offer a system-level view of the phylochemical map and underlying metabolic evolution. Such data could also be used with phylogenetic comparative methods, ancestral state reconstruction, hot node identification, and other approaches to potentially predict the presence of compounds in unexplored lineages.

Another interesting pattern that emerged from our mapping efforts was repeated reports of tyramine, *N*-methyl-4-tyramine, and hordenine in multiple lineages (e.g., Sapindales, Gentianales, Poales, Figures 2C and 3C). While we know that these compounds are the intermediates and product (respectively) of the hordenine biosynthetic pathway (Figure 3A), this result illustrates that phylochemical mapping has the potential to illuminate lineage-specific metabolic innovations, such as lineage-specific modification or evolution of a specific branch of a metabolic network. Increased resolution of the mapping, as discussed above, could also predict *when* such evolutionary innovation might have occurred. These results suggest that, for chemicals generated by yet unknown biosynthetic networks, phylochemical maps could identify potential pathway intermediates. Together with systems that apply logical rules to chemical structures or biochemical pathways (Hafner et al., 2021; Rodriguez-Lopez et al., 2022), it may be possible to enhance predictions about unknown or novel biochemical pathways and reconstruct their lineage-specific metabolic evolution. A phylochemical database could be also linked to metabolic network databases, such as Plant Metabolic Network (Hawkins et al., 2021) to facilitate our understanding of phylogenetic distribution and thereby evolutionary history of plant metabolic networks. Finally, we note that, in addition to metabolic networks, metabolomic data could be integrated with a phylochemical database which could potentially help assign unknown peaks in metabolomic datasets by linking the phylogenetic occurrence of those peaks to previously reported areas of metabolism within a given lineage.

A phylochemical map could be further merged with maps of other traits such as genetic, biochemical, morphological, and physiological traits, as well as climatic and geographic distribution patterns. Comparing the phylogenetic distributions of multiple traits with compound occurrence could potentially illuminate novel and important links. For example, a recent study applied a similar idea to identify a link between certain geographic patterns and enzyme biochemical traits (Fahmi et al., 2023). Also, since metabolites are synthesized through enzymes encoded in genome of each individual, the phylogenetic distribution of certain chemicals with genomic and biochemical information could help identify the genetic basis of chemical synthesis. Finally, the comparison of chemical distributions and ecological and geographic patterns could infer potential biological functions and ecological roles of plant specialized metabolites.

In addition to providing research-oriented benefits, maps of multiple trait types are also excellent catalysts for interdisciplinary systems thinking, which have the potential to help us strengthen the STEM workforce (Busta & Russo, 2020). The manual curation process of this study was carried out as a part of virtual undergraduate research during the COVID pandemic, where five students worked closely to critically evaluate scientific literature, contributed to the interpretation of the findings, and also learned various functions and distribution of diverse plant chemicals at a larger phylogenetic context. To further facilitate such research and educational efforts, software tools that allow interaction with maps of multiple data types could be highly useful. For example, a previously published fatty acid database, PlantFADB, illustrates one potential platform by which researchers and students could interact with phylochemical maps (Ohlrogge et al., 2018). One excellent feature of such a website is the ability to view the underlying data on user selected phylogenetic level, such as the abundance or presence of a given compound across different plant orders, different plant families, or different genera. Depending on when a biochemical pathway of interest evolved, viewing the data at one phylogenetic level might be more useful than other views. For example, higher-level phylogenetic views might not provide any resolution for more lineage-specific compounds, while a view at the genus or even species-level might only highlight patchiness in the data and such a map might not be a good tool to test a given hypothesis. In short, larger phylochemical databases developed in the future could benefit from the option to be viewed in a dynamic fashion. Finally, we also note that the open-source project LOTUS (Rutz et al., 2022), hosted at Natural Products Online (lotus.naturalproducts.net), accepts community additions of data via entries on Wikidata, and the pipeline described here or its future versions could greatly expand that repository.

Generating more comprehensive phylochemical maps

While this study built a phylochemical map for a targeted class of compounds (tyrosine-derived compound) as a case study, the workflow we developed is capable of generating plant chemical occurrence maps in a much more comprehensive and untargeted manner. This untargeted approach would involve processing scientific articles with a language model to extract compound–species associations without focusing on specific chemical compounds or plant species. As the capabilities of AI systems advance, an approach that will probably be worth exploring is the use of so-called "vision" AI models, models that understand information embedded in images (jpeg, png, etc.), and whether such can reliably extract compound–species relationships from figures and tables, such as those often included in review articles. Another important consideration is the source of material used to develop more exhaustive databases. Currently, many research institutions only allow data in the public domain to be passed to closed source language models. For this reason, scaled up workflows will likely need to be centered around public domain resources like PubMed, or will need to rely on open-source language models run on local hardware. Either way, a scaled-up, untargeted approach to compound–species relationship extraction would yield untargeted maps, providing a broader view of plant chemical distributions to help us identify novel molecular mechanisms that underpin chemical evolution, predict pathway plasticity, unearth novel pathways (Rodriguez-Lopez et al., 2022), and identify variant enzymes with potential industrial or agricultural applications.

The creation of large, untargeted plant chemical maps from extensive literature must consider the scalability of the workflow, which comprises three main steps: accessing or compiling a literature collection, extracting candidate compound–species associations, and evaluating these associations. Scaling the literature database(s) that can be used for input into a mapping workflow seems essentially complete with the availability of near-comprehensive and programmatically accessible public databases like PubMed, even if only abstracts and not full texts are available. Scaling up the extraction process of candidate compound–species associations is also practical, since the list of plant chemicals and species can be found in large public repositories, such as PubChem and botanical databases like World Flora Online. Finally, the evaluation of these associations can be done through a computer program capable of understanding natural language, though small-scale manual curations could be helpful for validation and further training of language models. Fortunately, unlike tasks with exponentially increasing solution spaces, such as phylogeny building, the evaluation of compound–species associations based on text excerpts is linear in complexity. Therefore, the required time and effort can be scaled linearly with the number of evaluations, which makes the extensive expansion of phylochemical mapping feasible. In fact, we could theoretically expand the workflow to encompass all known chemical space as well as species outside of the plant kingdom. This scalability highlights the relevance of our work and its potential value to researchers across various fields of natural product research that span the tree of life.

MATERIALS AND METHODS

Text mining for compound–species associations

To prepare a list of tyrosine derived compounds for subsequent literature mining, we inspected prior published literature and the KEGG plant chemical compound list to identify 71 tyrosine derived compounds that spanned a range from very common or ubiquitous compounds produced in nearly all lineages to highly lineage specific compounds known to be produced in only a select few genera. SciFinder searches were performed manually by querying the database for the CAS number associated with each of the 71 tyrosine derived compounds. The results page from each search was filtered using the SciFinder filtering tool to obtain articles associated with the keywords "plant" or "taxonomy". Compounds with more than 10 000 hits were excluded manually using SciFinder. The records for the remaining articles were temporarily downloaded and compared against the World Flora Online Plant List (https://wfoplantlist.org/plant-list) using regular expression-based pattern matching in R (provided by the R Core Team). Synonyms and unofficial names were not considered in our pattern matching. The 2571 articles that mentioned a plant genus and species name present in the World Flora Online Plant List were retained and organized into a spreadsheet (Table S2), and the other records were deleted from the local storage space.

Manual curation of compound–species associations

Full texts for each of the 2571 articles were obtained through our university libraries. Each was manually inspected to determine whether the putative compound–species associations found by the pattern matching approach were true, false, or unclear. Additional metadata was also collected from each full text, including data on the tissue types reported to accumulate the compound of interest, other plant species reported to accumulate the target compound within each article, and whether the report of the compound included analytical quantification. To visualize the resulting data, we pruned a published phylogenetic tree to include only the orders that were represented in our association data. Phylochemical maps were visualized by using the R package "ggtree" (Yu et al., 2016) to visualize a pruned phylogeny (Qian & Jin, 2016) plotted alongside a heatmap representing the association data generated with ggplot2 (Wickham, 2016).

Analysis of metabolites by ultra-high-performance liquid chromatography (UHPLC)-MS

Approximately 100 mg of healthy fresh plant tissue was harvested from plants grown in the Botany Greenhouses of University of Wisconsin-Madison and flash frozen in liquid nitrogen. The tissues were then freeze dried using the Benchtop Freeze Dryer

(Labconco) for no less than 48 h and pulverized with a bead beater. Ten to twenty milligrams of frozen powders were aliquoted to new frozen tubes, weighed, and resuspended in 800 μl of LC–MS grade methanol: chloroform, (v/v 2:1) with 10 μM ^{13}C(6)-tyrosine (Cambridge Isotope, CLM-1542-0.25) as an internal recovery standard. After rigorous vortexting for 10 min, the extracts were centrifuged for 10 min at 21 000 ***g*** at room temperature, and 800 μl of the supernatant was transferred to a new tube, mixed with 600 μl water and 250 μl chloroform, vortexed for 30 sec, and then centrifuged again for 5 min at 21 000 ***g*** to generate phase separation. The upper aqueous phase was transferred into a new tube and dried down in a CentriVap SpeedVac (Labconco) overnight at room temperature. The dried samples were resuspended in 70 μl of 80% methanol by vortexing for 30 sec and sonicating for 5 min, followed by centrifugation for 5 min at 21 000 ***g***. Then, 30 μl of the supernatant was transferred to glass vials for injection onto a Vanquish Horizon UHPLC Binary Pump H coupled to a Q-Exactive MS (Thermo Scientific). All reagents were of LC–MS quality. One microliter of the sample was injected onto a HSS T3 C18 reversed phase column (100 × 2.1 mm i.d., 1.8-μm particle size; Waters) and eluted using a 26-mi gradient comprising 0.1% formic acid in water (solvent A) and 0.1% formic acid in acetonitrile (solvent B) at a flow rate of 0.40 ml/min and column temperature of 40°C, with the following linear gradient of solvent B: 0–1 min, 1%; 1–10 min, 1–10%; 10–13 min, 10–25%; 13–18 min, 25–99%; 18–22 min, 99%; 22–22 min, 99–1%; 22–26 min, 1%. Full MS spectra were recorded between 0.55 and 18 min using the full scan positive mode, with the following parameters: sheath gas flow rate, 45; auxiliary gas flow rate, 13; sweep gas flow rate, 1; spray voltage, 3.75 kV; capillary temperature, 350°C; S-lens RF level, 50; resolution, 140 000; AGC target 1 × 10e6, maximum scan time 200 ms; scan range 100–1000 *m*/*z*. Metabolites were identified through a mass search of the compounds' mass to charge ratios (*m*/*z*), generating an extracted ion chromatogram (EIC) from which retention time and mass spectra were matched to high purity authentic standards: tyrosine (A11141, Alfa Aesar), tyramine (AC140610050, Acros Organics), and *N*-methyl-4-tyramine (OR-4655, Combi-Blocks), hordenine (04476, Sigma Aldrich). Quantification of each compound was done by manual integration of peak area using Xcalibur 3.0 (Qual Browser, Thermo Scientific) and abundance was calculated by reference to the authentic standards. The final compound contents were determined by normalizing based on the internal standard (% recovery) and tissue weight for each sample.

Language model evaluation of candidate compound–species associations

Candidate compound–species associations were evaluated using language models created by OpenAI (2023). Compound species associations were evaluated inside a loop in which each candidate association was passed to the OpenAI API using an R script like:

```
query
 "Here is the title: ", data$title[i],
 "Here is the abstract: ", data$abstract[i],
 " classify your response as \"yes\" or \"no\": was ",
 data$compound_name[i], " found in ", data$species[i],
 "? Do not include a period in your response. Respond only
with yes or no and no other explanation or text."
)
```

The system prompt used was "You are an expert in scientific literature and can easily and accurately parse text for scientific literature tasks." The output from the model was used directly in downstream analysis. A full example of the code used to evaluate the language models is included at the project GitHub site. Once the responses from the language model had been tabulated, ANOVA and post-hoc Tukey tests, implemented in R via the "statix" (rstatix, 2020) were used to test for any significant differences.

AUTHOR CONTRIBUTIONS

LB and HAM conceived the research. LB conducted the text mining and phylogenetic mapping with the help of HAM. The curation of candidate compound–species associations was performed by DH, CMH, AG, YW, MG, and MS. Language model evaluation was conducted by BJ and LB. LC-MS validation of hordenine and its precursors was done by MS with an initial assistant from MG. The manuscript was drafted by LB, HAM, DH, and MS with input from all authors. All authors contributed to the analyses and interpretation of the data and read the final version of the manuscript.

ACKNOWLEDGMENTS

We thank Drs. Ricardo Kriebel and Ken Sytsma for helpful discussions for mapping chemical traits across the plant phylogeny and Dr. Bethany Moore for discussing and testing a possibility of a machine learning approach. We are also grateful to Dr. Alan Oyler for assistance with preparing this manuscript. LB gratefully acknowledges support from the Swenson College of Science and Engineering in the form of start-up funds and a SciFinder Future Leader Award from CAS SciFinder. BJ acknowledges the support of a UROP award from the University of Minnesota. This work was supported by grants from the US National Science Foundation (IOS-PGRP-1836824 and DEB-1938597) to HAM.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The datasets generated and analyzed during this study as well as the relevant code are available on GitHub at https://github.com/thebustalab/phylochemical_mapping. This includes the annotated 3000+ compound–species associations, code used to evaluate language models, and code used to create phylochemical maps presented in this work.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Text mining to identify of compound–species associations.

Figure S2. A map of false positive hits from text mining.

Figure S3. Maps of false negatives from text mining and true compound–species associations missed by the text mining pipeline.

Table S1. Compounds considered for phylochemical mapping.

Table S2. Compound–species associations examined in this study.

Table S3. True compound–species associations missed by text mining that were identified by manual searching.

REFERENCES

Afendi, F.M., **Okada, T.**, **Yamazaki, M.**, **Hirai-Morita, A.**, **Nakamura, Y.**, **Nakamura, K.** ***et al.*** (2012) KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant & Cell Physiology*, **53**, e1. Available from: https://doi.org/10.1093/pcp/pcr165

Ago, Y., **Koda, K.**, **Takuma, K.** & **Matsuda, T.** (2011) Pharmacological aspects of the acetylcholinesterase inhibitor galantamine. *Journal of Pharmacological Sciences*, **116**, 6–17. Available from: https://doi.org/10.1254/jphs.11R01CR

Aragane, M., **Sasaki, Y.**, **Nakajima, J.**, **Fukumori, N.**, **Yoshizawa, M.**, **Suzuki, Y.** ***et al.*** (2011) Peyote identification on the basis of differences in morphology, mescaline content, and trnL/trnF sequence between *Lophophora williamsii* and *L. diffusa*. *Journal of Natural Medicines*, **65**, 103–110. Available from: https://doi.org/10.1007/s11418-010-0469-7

Awan, A.R., **Shaw, W.M.** & **Ellis, T.** (2016) Biosynthesis of therapeutic natural products using synthetic biology. *Advanced Drug Delivery Reviews*, **105**, 96–106. Available from: https://doi.org/10.1016/j.addr.2016.04.010

Bakker, M.A., **Chadwick, M.J.**, **Sheahan, H.R.**, **Tessler, M.H.**, **Campbell-Gillingham, L.**, **Balaguer, J.** ***et al.*** (2022) Fine-tuning language models to find agreement among humans with diverse preferences. *arXiv*. 2211.15006. Available from: https://doi.org/10.48550/arXiv.2211.15006

Bisset, N.G. (1985) Plants as a Source of Isoquinoline Alkaloids. In: Phillipson, J.D., Roberts, M.F. & Zenk, M.H. (Eds.) *The Chemistry and Biology of Isoquinoline Alkaloids*. Berlin: Springer-Verlag. Available from: https://doi.org/10.1007/978-3-642-70128-3_1

Busta, L. & **Russo, S.E.** (2020) Enhancing interdisciplinary and systems thinking with an integrative plant chemistry module applied in diverse undergraduate course settings. *Journal of Chemical Education*, **97**, 4406–4413. Available from: https://doi.org/10.1021/acs.jchemed.0c00395

Chae, Y. & **Davidson, T.** (2023) Large language models for text classification: from zero-shot learning to fine-tuning. *Open Science Foundation*. SocArXiv. Available from: https://doi.org/10.31235/osf.io/sthwk

Clement, B.A., **Goff, C.M.** & **Forbes, T.D.A.** (1997) Toxic amines and alkaloids from *Acacia berlandieri*. *Phytochemistry*, **46**, 249–254. Available from: https://doi.org/10.1016/s0031-9422(97)00240-9

Clement, B.A., **Goff, C.M.** & **Forbes, T.D.A.** (1998) Toxic amines and alkaloids from acacia rigidula. *Phytochemistry*, **49**, 1377–1380. Available from: https://doi.org/10.1016/s0031-9422(97)01022-4

Coley, P.D., **Endara, M.J.**, **Ghabash, G.**, **Kidner, C.A.**, **Nicholls, J.A.**, **Pennington, R.T.** ***et al.*** (2019) Macroevolutionary patterns in overexpression of tyrosine: an anti-herbivore defence in a speciose tropical tree genus, Inga (Fabaceae). *Journal of Ecology*, **107**, 1620–1632. Available from: https://doi.org/10.1111/1365-2745.13208

Cortes, N., **Castaneda, C.**, **Osorio, E.H.**, **Cardona-Gomez, G.P.** & **Osorio, E.** (2018) Amaryllidaceae alkaloids as agents with protective effects against oxidative neural cell injury. *Life Sciences*, **203**, 54–65. Available from: https://doi.org/10.1016/j.lfs.2018.04.026

Defossez, E., **Pitteloud, C.**, **Descombes, P.**, **Glauser, G.**, **Allard, P.M.**, **Walker, T.W.N.** ***et al.*** (2021) Spatial and evolutionary predictability of phytochemical diversity. *Proceedings of the National Academy of Sciences of the United States of America*, **118**, e2013344118. Available from: https://doi.org/10.1073/pnas.2013344118

Duke, J.A. (2023) Dr. Duke's Phytochemical and Ethnobotanical Databases. https://data.nal.usda.gov/dataset/dr-dukes-phytochemical-and-ethnobotanical-databases https://doi.org/10.15482/USDA.ADC/1239279

Dunn, A., **Dagdelen, J.**, **Walker, N.**, **Lee, S.**, **Rosen, A.S.**, **Ceder, G.** ***et al.*** (2022) Structured information extraction from complex scientific text with fine-tuned large language models. *arXiv*, 2212.05238v1. Available from: https://doi.org/10.48550/arXiv.2212.05238

El-Seedi, H.R., **De Smet, P.A.**, **Beck, O.**, **Possnert, G.** & **Bruhn, J.G.** (2005) Prehistoric peyote use: alkaloid analysis and radiocarbon dating of archaeological specimens of Lophophora from Texas. *Journal of Ethnopharmacology*, **101**, 238–242. Available from: https://doi.org/10.1016/j.jep.2005.04.022

Ernst, M., **Nothias, L.F.**, **van der Hooft, J.J.J.**, **Silva, R.R.**, **Saslis-Lagoudakis, C.H.**, **Grace, O.M.** ***et al.*** (2019) Assessing specialized metabolite diversity in the cosmopolitan plant genus euphorbia L. *Frontiers in Plant Science*, **10**, 846. Available from: https://doi.org/10.3389/fpls.2019.00846

Fahmi, M., **Takanashi, K.**, **Kakei, Y.**, **Kubota, Y.** & **Kanemoto, K.** (2023) Mappig natural product biosynthetic hotspots: prioritizing conservation for medicinal resources. *bioRxiv*, 2023.08.12.553062. Available from: https://doi.org/10.1101/2023.08.12.553062

Fan, P., **Miller, A.M.**, **Liu, X.**, **Jones, A.D.** & **Last, R.L.** (2017) Evolution of a flipped pathway creates metabolic innovation in tomato trichomes through BAHD enzyme promiscuity. *Nature Communications*, **8**, 2080. Available from: https://doi.org/10.1038/s41467-017-02045-7

Gabrielson, S.W. (2018) SciFinder. *Journal of the Medical Library Association*, **106**, 2018.515. Available from: https://doi.org/10.5195/jmla.2018.515

Hafner, J., **Payne, J.**, **MohammadiPeyhani, H.**, **Hatzimanikatis, V.** & **Smolke, C.** (2021) A computational workflow for the expansion of heterologous biosynthetic pathways to natural product derivatives. *Nature Communications*, **12**, 1760. Available from: https://doi.org/10.1038/s41467-021-22022-5

Hartmann, T. (2007) From waste products to ecochemicals: fifty years research of plant secondary metabolism. *Phytochemistry*, **68**, 2831–2846. Available from: https://doi.org/10.1016/j.phytochem.2007.09.017

Hawkins, C., **Ginzburg, D.**, **Zhao, K.**, **Dwyer, W.**, **Xue, B.**, **Xu, A.** ***et al.*** (2021) Plant metabolic network 15: a resource of genome-wide metabolism databases for 126 plants and algae. *Journal of Integrative Plant Biology*, **63**, 1888–1905. Available from: https://doi.org/10.1111/jipb.13163

Howard, J. & **Ruder, S.** (2018) Universal language model fine-tuning for text classification. *arXiv*, 1801.06146. Available from: https://doi.org/10.48550/arXiv.1801.06146

Huong, D.T., **Kamperdick, C.** & **Sung, T.V.** (2004) Homogentisic acid-derivatives from *miliusa balansae*. *Journal of Natural Products*, **67**, 445–447. Available from: https://doi.org/10.1021/np030195z

Jablonka, K.M., **Ai, Q.**, **al-Feghali, A.**, **Badhwar, S.**, **Bocarsly, J.D.**, **Bran, A.M.** ***et al.*** (2023) 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digit Discov*, **2**, 1233–1250. Available from: https://doi.org/10.1039/d3dd00113j

Kanehisa, M., **Goto, S.**, **Hattori, M.**, **Aoki-Kinoshita, K.F.**, **Itoh, M.**, **Kawashima, S.** ***et al.*** (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research*, **34**, D354–D357. Available from: https://doi.org/10.1093/nar/gkj102

Leonard, E., **Runguphan, W.**, **O'Connor, S.** & **Prather, K.J.** (2009) Opportunities in metabolic engineering to facilitate scalable alkaloid production. *Nature Chemical Biology*, **5**, 292–300. Available from: https://doi.org/10.1038/nchembio.160

Li, X., **Liu, Z.**, **Xiong, C.**, **Yu, S.**, **Gu, Y.**, **Liu, Z.** ***et al.*** (2023) Structure-aware language model pretraining improves dense retrieval on structured data. *arXiv*, 2305.19912v1. Available from: https://doi.org/10.48550/arXiv.2305.19912

Lichman, B.R., **Godden, G.T.**, **Hamilton, J.P.**, **Palmer, L.**, **Kamileen, M.O.**, **Zhao, D.** ***et al.*** (2020) The evolutionary origins of the cat attractant nepetalactone in catnip. *Science Advances*, **6**, eaba0721. Available from: https://doi.org/10.1126/sciadv.aba0721

Lopez-Nieves, S., **el-Azaz, J.**, **Men, Y.**, **Holland, C.K.**, **Feng, T.**, **Brockington, S.F.** ***et al.*** (2022) Two independently evolved natural mutations additively deregulate TyrA enzymes and boost tyrosine production in planta. *The Plant Journal*, **109**, 844–855. Available from: https://doi.org/10.1111/tpj.15597

Lopez-Nieves, S., **Yang, Y.**, **Timoneda, A.**, **Wang, M.**, **Feng, T.**, **Smith, S.A.** ***et al.*** (2018) Relaxation of tyrosine pathway regulation underlies the evolution of betalain pigmentation in Caryophyllales. *The New Phytologist*, **217**, 896–908. Available from: https://doi.org/10.1111/nph.14822

Loub, W.D., **Farnsworth, R.**, **Soejarto, D.D.** & **Quinn, M.L.** (1895) NAPRALERT: computer handling of natural product research data. *Journal of Chemical Information and Computer Sciences*, **25**, 99–103. Available from: https://doi.org/10.1021/ci00046a009

Maeda, H.A. & **Fernie, A.R.** (2021) Evolutionary history of plant metabolism. *Annual Review of Plant Biology*, **72**, 185–216. Available from: https://doi.org/10.1146/annurev-arplant-080620-031054

McChesney, J.D., **Venkataraman, S.K.** & **Henri, J.T.** (2007) Plant natural products: back to the future or into extinction? *Phytochemistry*, **68**, 2015–2022. Available from: https://doi.org/10.1016/j.phytochem.2007.04.032

Mitchell, W. (2011) Natural products from synthetic biology. *Current Opinion in Chemical Biology*, **15**, 505–515. Available from: https://doi.org:10.1016/j.cbpa.2011.05.017

Mocan, A., **Zengin, G.**, **Mollica, A.**, **Uysal, A.**, **Gunes, E.**, **Crişan, G.** ***et al.*** (2018) Biological effects and chemical characterization of Iris schachtii Markgr. Extracts: a new source of bioactive constituents. *Food and Chemical Toxicology*, **112**, 448–457. Available from: https://doi.org/10.1016/j.fct.2017.08.004

Moghe, G.D. & **Last, R.L.** (2015) Something old, something new: conserved enzymes and the evolution of novelty in plant specialized metabolism. *Plant Physiology*, **169**, 1512–1523. Available from: https://doi.org/10.1104/pp.15.00994

Newman, D.J. & **Cragg, G.M.** (2016) Natural products as sources of new drugs from 1981 to 2014. *Journal of Natural Products*, **79**, 629–661. Available from: https://doi.org/10.1021/acs.jnatprod.5b01055

Ohlrogge, J., **Thrower, N.**, **Mhaske, V.**, **Stymne, S.**, **Baxter, M.**, **Yang, W.** ***et al.*** (2018) PlantFAdb: a resource for exploring hundreds of plant fatty acid structures synthesized by thousands of plants and their phylogenetic relationships. *The Plant Journal*, **96**, 1299–1308. Available from: https://doi.org/10.1111/tpj.14102

OpenAI. (2023) *OpenAI API*. https://platform.openai.com/docs/api-reference

Pichersky, E. & **Lewinsohn, E.** (2011) Convergent evolution in plant specialized metabolism. *Annual Review of Plant Biology*, **62**, 549–566. Available from: https://doi.org/10.1146/annurev-arplant-042110-103814

Qian, H. & **Jin, Y.** (2016) An updated megaphylogeny of plants, a tool for generating plant phylogenies and an analysis of phylogenetic community structure. *Journal of Plant Ecology*, **9**, 233–239. Available from: https://doi.org/10.1093/jpe/rtv047

Rodriguez-Lopez, C.E. ***et al.*** (2022) Phylogeny-aware chemoinformatic analysis of chemical diversity in Lamiaceae enables iridoid pathway assembly and discovery of Aucubin synthase. *Molecular Biology and Evolution*, **39**, msac057. Available from: https://doi.org/10.1093/molbev/msac057

rstatix. (2020) rstatix: Pipe-friendly framework for basic statistical tests.

Rutz, A., **Sorokina, M.**, **Galgonek, J.**, **Mietchen, D.**, **Willighagen, E.**, **Gaudry, A.** ***et al.*** (2022) The LOTUS initiative for open knowledge management in natural products research. *eLife*, **11**, 70780. Available from: https://doi.org/10.7554/eLife.70780

Schenck, C.A. & **Busta, L.** (2022) Using interdisciplinary, phylogeny-guided approaches to understand the evolution of plant metabolism. *Plant Molecular Biology*, **109**, 355–367. Available from: https://doi.org/10.1007/s11103-021-01220-1

Schenck, C.A. & **Maeda, H.A.** (2018) Tyrosine biosynthesis, metabolism, and catabolism in plants. *Phytochemistry*, **149**, 82–102. Available from: https://doi.org/10.1016/j.phytochem.2018.02.003

Siatka, T., **Adamcová, M.**, **Opletal, L.**, **Cahlíková, L.**, **Jun, D.**, **Hrabinová, M.** ***et al.*** (2017) Cholinesterase and prolyl Oligopeptidase inhibitory activities of alkaloids from Argemone platyceras (Papaveraceae). *Molecules*, **22**, 22071181. Available from: https://doi.org/10.3390/molecules22071181

Walsh, C.T. & **Fischbach, M.A.** (2010) Natural products version 2.0: connecting genes to molecules. *Journal of the American Chemical Society*, **132**, 2469–2493. Available from: https://doi.org/10.1021/ja909118a

Wang, M., **Lopez-Nieves, S.**, **Goldman, I.L.** & **Maeda, H.A.** (2017) Limited tyrosine utilization explains lower Betalain contents in yellow than in red table beet genotypes. *Journal of Agricultural and Food Chemistry*, **65**, 4305–4313. Available from: https://doi.org/10.1021/acs.jafc.7b00810

Wei, J., **Wang, X.**, **Schuurmans, D.**, **Bosma, M.**, **Ichter, B.**, **Xia, F.** ***et al.*** (2023) Chain-of-thought prompting elicits reasoning in large language models. *arXiv*, 2201.11903v6. Available from: https://doi.org/10.48550/arXiv.2201.11903

Weng, J.-K., **Philippe, R.N.** & **Noel, J.P.** (2012) The rise of Chemodiversity in plants. *Science*, **336**, 1667–1670. Available from: https://doi.org/10.1126/science.1217411

Wickham, H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

Wu, S. & **Chappell, J.** (2008) Metabolic engineering of natural products in plants; tools of the trade and challenges for the future. *Current Opinion in Biotechnology*, **19**, 145–152. Available from: https://doi.org/10.1016/j.copbio.2008.02.007

Xiao, Z., **Li, W.**, **Moon, H.**, **Roell, G.W.**, **Chen, Y.** & **Tang, Y.J.** (2023) Generative artificial intelligence GPT-4 accelerates knowledge mining and machine learning for synthetic biology. *ACS Synthetic Biology*, **12**, 2973–2982. Available from: https://doi.org/10.1021/acssynbio.3c00310

Yu, G., **Smith, D.K.**, **Zhu, H.**, **Guan, Y.** & **Lam, T.T.Y.** (2016) Ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*, **8**, 28–36. Available from: https://doi.org/10.1111/2041-210x.12628

Zust, T., **Strickler, S.R.**, **Powell, A.F.**, **Mabry, M.E.**, **An, H.**, **Mirzaei, M.** ***et al.*** (2020) Independent evolution of ancestral and novel defenses in a genus of toxic plants (*Erysimum*, Brassicaceae). *eLife*, **9**, 5171. Available from: https://doi.org/10.7554/eLife.51712