University of Nevada, Reno

A ^{1}H NMR Network Approach for the Study of Specialized Metabolites: Development and Applications

A thesis submitted in partial fulfillment of the

requirements for the degree of Doctor of Philosophy in

Chemistry

By

Celso Ricardo de Oliveira Júnior

Dr. Christopher S. Jeffrey/Dissertation Advisor

August, 2021

THE GRADUATE SCHOOL

We recommend that the dissertation
prepared under our supervision by

Celso Ricardo de Oliveira Jr

entitled

A 1H NMR Network Approach for the Study of Specialized Metabolites: Development and Applications

be accepted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

Christopher S. Jeffrey
Advisor

Matthew J. Tucker
Committee Member

Wesley A. Chalifoux
Committee Member

Lee A. Dyer
Committee Member

Ian S. Wallace
Graduate School Representative

David W. Zeh, Ph.D., Dean
Graduate School

August, 2021

Abstract

The history of the World could not be told without mentioning specialized metabolites. Long before we had the notion of organic molecules, we were already sailing across the oceans in the pursue of their valuable sources, and raising whole empires based on the economy of spices, tobacco and coffee. Our own survival and adaptability as a species can be credited to discoveries of nature-sourced drugs through the centuries, such as penicillin, paclitaxel and artemisinin. As our awareness towards the environment evolved, so did our perspective on these molecules, which are now recognized as important products and intermediates of the interaction between organisms and their ecosystems. It is not an overstatement thus, to affirm that in the light of an imminent environmental crisis, the study of specialized metabolites will be of fundamental importance for the preservation and recovery of biodiversity.

The development of untargeted analytical techniques has greatly advanced our ability to investigate these compounds in their natural contexts. Among those, the application of Proton Nuclear Magnetic Resonance (^{1}H NMR) spectroscopy for chemical profiling enabled the recognition of compound structural features and facilitated the identification of unknown compounds even prior to the laborious work of isolation. However, because multiple resonance signals arise from a single compound, a considerable amount of overlap is observed in biological samples which can limit our ability to detect those key structural markers. There are certain NMR pulse experiments that can aid in deconvoluting these signals, but a more practical approach resides in the statistical treatment of the ^{1}H NMR spectrum, in which regular variations across the spectrum are directly mapped to variables pertinent to the system in study, such as biological activity, biogeographical data, or phylogenetic classification.

This document presents an innovative strategy in which gene co-expression network analysis is adapted for the statistical treatment of ^{1}H NMR spectral data, resulting in the deconvolution of metabolite signals and simplification of the spectrum into a few variables. These variables represent statistically measurable chemical patterns that in conjunction with other measurements can support addressing a multitude of topics in Chemical Ecology. In Chapter 1, we describe the method development and validation in a controlled experiment with prepared mixtures of known compounds and demonstrate how it recognizes metabolite identity at different structural levels. We then demonstrate its applicability in the study of natural mixtures with the investigation of ontogenetic changes of metabolism in *Piper kelleyi.* In Chapter 2, we applied the method to the identification of biologically active compounds from a chemically heterogeneous set of 29 *Piper* extracts. By quantifying the association between chemical patterns and measurements of antifungal activity, we accurately identified specific targets for the isolation of antifungal compounds, while also establishing a framework to evaluate the effect of specific structural features in modulating the toxicity of different plant species. Finally, in Chapter 3, we adopted an untargeted approach to investigate the phylogenetic signal of specialized metabolites in a broader collection of *Piper* species. Based upon measurements of chemical similarity, we identified the chemical traits most strongly associated with the clade *Schilleria,* ultimately leading us to the characterization of novel lignans. Altogether, this ^{1}H NMR network approach represents a powerful tool for the study of specialized metabolites in a myriad of contexts relevant to the field of Chemical Ecology.

Dedication

In memory of Dr. Luiz Fernando da Silva Júnior and Dr. Paulo Teng An Sumodjo, who showed me how painfully hard learning can be, but also how priceless is the knowledge that emerges from it.

Acknowledgements

I write standing on the shoulders of so many people, famous and anonymous, who have contributed directly the development of science, in particular the field of Chemistry and Biology, by raising their determination to disseminate knowledge above the personal and societal challenges of their own times. This work is a tribute to those who dared to think differently and enter the uncharted territories of nature.

I would like to also acknowledge the fantastic scientists who directly helped me through my doctoral studies. First and foremost, my advisor Dr. Christopher Jeffrey, for being so incredibly supportive and for nurturing a mentorship that encouraged me to be independent. Chris has been an invaluable source of knowledge for NMR spectroscopy, and if today I am so passionate about this topic, it is largely because of what I learned from him, and because of the enthusiasm he showed at every spectrum that we discussed. He has also been a model lecturer and an example of interdisciplinary collaboration to me, and I am grateful that he challenged me with projects that I enjoyed, but that demanded me to learn more about ecology and bioinformatics. And Dr. Lora Richards who me into her office and shared her expertise in these two subjects with me. Lora has been far more than a mentor in R coding and statistical analysis, she has been a colleague to chat about papers, projects and the academic life over a coffee during my lunch breaks. These exchanges often helped me to regain stability in moments of uncertainty. I am also thankful to Dr. Lee Dyer, who first welcomed me into the UNR chemical ecology group in 2014, when I was still an undergraduate student. Lee is a cornerstone in our studies with *Piper* and along with Chris and Lora, a proponent for the development of this NMR network approach. Through my graduate studies, Lee has contributed to strengthen my knowledge in statistical analysis and experimental design by means of directly ministered classes or during discussions in collaborative projects. He also helped collecting and transporting from Costa Rica some of the *Piper* species I used in the isolation projects. To Dr. Ian Wallace I express my gratitude for contributing with his expertise in biochemistry and microbiology to the study of bioactive compounds from *Piper*. Ian has also entertained me with insightful discussions, either as a collaborator or as a committee member, that helped me to develop a more critical outlook on the biological extension of my projects. I also thank Dr. Matthew Tucker and Dr. Wesley Chalifoux for serving as members of my academic committee, and for bringing their experience in spectroscopy and organic synthesis, respectively, in discussions that impacted my development as a scientist and which are ultimately implicated in this research work.

To the staff at the Chemistry Department, my gratitude for ensuring the administrative and technical support necessary for the development of my research, in special to Dr. Steven Spain for the upkeep of the NMR spectrometers and for coming to my aid during the times when the instruments did not behave as I expected. Also, Dr. Sarah Cummings, Dr. Lyndsay Munro, and Dr. Janell Mahoney, who supervised me during my teaching duties, and who granted me the position of head TA during many of the semesters that I served in the teaching labs. Dr. Sean Casey, Dr. Benjamin King and Dr. Vincent Catalano, who served as department chair during my doctorate studies. Beyond our department walls, I thank Adilia Ross for always being so kindly helpful with the bureaucracies of my student visa. I was also fortunate to receive the generous endowments from Dr. Scott Chadwick and Dr. Mick Hitchcock, who recognized the importance of fomenting the advancement of science and deemed the merit my research work. Mick was also a person of utmost importance in the establishment of the Hitchcock Center for Chemical Ecology, which has nurtured the collaborative effort of the projects here highlighted and many others in our research group.

To the countless friends and colleagues who morally supported me through my years of graduate school, especially Dr. Jorge Monteiro, Dr. Paulo Regis, Dr. Dylan Jones, Samuel Silva, Dr. Farzaneh Chalyavi, Dr. Natalie Fetto, Dr. Andrea Glassmire and Dr. Tara Massad. To the members of the Jeffrey group, former and present who have contributed to my research through direct collaboration or through the informal exchange of knowledge, and who have made our work environment a place of friendly cooperation, particularly Dr. Kaitlin Ochsenrider, Dr. Casey Philbin, Dr. Mary Fennimore, Alex English and Jennifer McCracken. Also, to the undergraduate students Arran Rumbaugh and Megan Warner for their direct contribution to some of my projects, and to Dr. Flavia Nogueira, for challenging me with the task of mentoring a biologist into NMR spectroscopy, and for reminding me that learning is an exercise of humility.

I am eternally grateful to the people who unknowingly raised me to be a scientist, my parents. Having been denied the privileged of education to work at an early age, they dedicated their lives to ensure that my sister and I had a better opportunity to show our potential. Their rigor and humility taught me to be a better student, and their unconditional love and support provided me with the encouragement to pursue a doctoral degree and develop my career in another country. Also, to the Torres and Hettrick families, especially Greg, Connie and Holly, for supporting me as one of their own kids. And finally, to Tyler Torres, who shared virtually every day of this experience with me, believing in me even when I doubted myself. Thank you, Tyler, for your dedicated friendship.

Table of Contents

List of Tables

List of Figures

Chapter 1: Development of a Network Analysis Framework for ^{1}H NMR Data

The work presented in this chapter was published[1] and this author contributed in the execution of research and elaboration of the manuscript.

1.1 Introduction

1.1.1 The study of specialized metabolites

The notion that the planet holds an unmeasurable natural diversity, much of which remains unexplored,[2] makes a formidable claim for the bioprospection of biologically relevant compounds. With that motivation, the avid researcher that arrives to an untouched segment of forest may contemplate the possibility of chemically dissecting all of the plant species he can collect, and his ambition will be well justified if one of them yields a *new* molecule of pharmaceutical importance, or one that holds unique properties and structure. However feasible, this serendipitous search is lengthy, costly and may lead to duplicated or individually inactive compounds, which is not surprising given growing evidence for the combinatorial effect of some metabolites.[3] Additionally, the search for biologically active compounds is typically biased towards a few targets of clinical relevance, often within the narrow scope of expertise of each research group, thus underestimating the functionality of some compounds. The efforts to maintain public molecular libraries may help to close gaps in the functional space, but still, a focus on clinical application underscores an array of specialized metabolites that serve important functions in their natural context, albeit being irresponsive to targeted assays.

An alternative ecological-guided approach prioritizes the *all* over the *part*, in which entire chemical phenotypes are perceived as the outcome of the dynamic interaction between organisms and their ecossystems.[4] This perspective is rooted in Erlich and Raven's coevolutionary hypothesis[5] and is supported by modern phylogenetic analysis of plants and their herbivores.[6] As a consequence of this evolutionary selection, the contextualized study of specialized metabolites has the potential to offer direct leads towards compounds with enhanced biological properties, newly described biological functions, and unique structural features. For instance, the quantification of chemical phenotypes and its patterns of variation is desirable not only for chemical ecology, but also for natural products research.

1.1.2 ^{1}H NMR-based metabolomics

Unlike the traditional approaches to compound discovery, a holistic investigation of chemical phenotypes demands the untargeted collection and analysis of large volumes of multivariate data (metabolomics). In order to extract relevant *chemical signals* from the metabolic baseline, the analytical method must be capable to detect compounds at widely different concentrations across chemical space, while also providing distinguishable signals for each metabolite. These challenges have been partially overcome by important technological advancements in Mass Spectrometry (MS) and ^{1}H NMR spectroscopy, markedly towards the improvement of detection limits, structural resolution, and dynamic ranges.[7, 8]

NMR spectroscopy and MS provide complementary types of information about the metabolome.[9, 10] The latter is arguably the most commonly applied technique due to instrument accessibility, high sensitivity and deconvolution of compound signals through coupling with chromatography. However, ^{1}H NMR spectroscopy presents several advantages over MS that are suitable for comparing organisms at varying taxonomic levels.[11] Namely, it can virtually detect any type of compound independently of its physicochemical properties (volatility, polarity, molecular weight, etc.), which also translates into facile sample preparation methods.[12] In addition, ^{1}H NMR spectroscopy data is reproducible, quantitative and, more importantly, it provides a high degree of molecular information that can expedite compound identification.

Structural resolution comes at the expense of a convoluted spectrum, so the application of ^{1}H NMR spectroscopy in metabolomics depends on statistical analyses for data mining. Typically, ordination techniques such as PCA (Principal Component Analysis) and PLS-DA (Partial Least Squares Discriminant Analysis) are utilized to detect a few peaks in the spectrum that amount to the largest variation across samples. While these techniques are well suited for discriminating chemical markers from moderately homogeneous samples (e.g. specimens under differentiated treatment or different cultivars),[13-16] they do not perform well for the identification of subtle similarities across taxa, where the number of distinguishing variables is far more expressive. Meanwhile, recent advancements on MS data analysis demonstrated the potential of molecular networks to deconvolute complex chemical datasets and identify shared structural features across samples,[17, 18] thus providing an attractive precedent for the development of a similar strategy towards ^{1}H NMR data.

1.1.3 Weighted Gene Co-Expression Network Analysis (WGCNA)

Networks represent a mathematical reconstruction of the relationships between the elements within a system. Each element is represented as a *node*, the connection between a pair of nodes is defined by an *edge*, and the overall arrangement of these components define the *network topology*. This concept gained particular relevance in the field systems biology and became a powerful means to explain the intricacies of biochemical processes.[19] Zhang and Horvath consolidated a framework for the study of gene expression data that utilizes weighted network analysis to identify clusters of co-expressed genes (we define weighted and unweighted networks in section 1.2).[20] From this methodology, each gene is a node and the strength of an edge is defined by the degree of similarity in the expression profiles of its pair of genes. If a group of genes fits the same overall profile, they are grouped into *modules*, which can then be verified against phenotypical data to reveal biologically meaningful sets of genes.

We consider that the chemical profile obtained from ^{1}H NMR spectra is structurally very similar to gene expression data, and that the WGCNA methodology could be easily adapted to this type of data. For instance, a network constructed with hydrogen resonance data (chemical shifts and integration) will result in modules composed of co-occurring peaks, which represent compounds or molecular features present in the sample set. Therefore, we predicted that (a) these chemical modules should reveal shared chemotypes across samples, and (b) the correlation between module values and biological or ecological measurements should offer direct leads to functional compounds. We first tested prediction (a) by applying the WGCNA approach to a controlled experiment with prepared compound mixtures, and then validated both assumptions with the analysis of a set of crude extracts of the plant species *Piper kelleyi*, where we verified the association of chemical modules with qualitative ontogenetic data.

1.2 Methods

Sample preparation: We prepared 196 artificial mixtures containing three to four different compounds from a collection of 31 plant specialized metabolites, obtained commercially or extracted from natural sources (Table 1.2). Admittedly, these compounds comprise only a small fraction of the phytochemical landscape, but they represent the structural features observed in major groups of metabolites implied in plant defenses, including terpenoids, flavonoids, phenylpropanoids, furanocoumarins, amides and alkaloids.[21] The mixtures were prepared in deuterated methanol at a concentration of 10 mg/mL, and while this is not an optimal solvent for ^{1}H NMR spectroscopy analysis due to the presence of two wide residual peaks that can overlap with sample peaks, it was the one that best solubilized the tested compounds. Additionally, methanol is the preferable solvent for extracting a broad range of metabolites in untargeted experiments,[22] so this choice of analytical solvent is also determining for generating network modules that can be applied annotate phytochemical data obtained in future studies. Mixture compositions were designed to reflect varying degrees of metabolite complexity, as observed for phytochemical extracts, and they can be divided in three groups. "Intraclass" samples included 21 combinations of three compounds within the same metabolic group at a mass ratio of 3:1:1. "Interclass" samples included 97 combinations of three compounds from two different metabolic groups also at the ratio of 3:1:1. Lastly, 78 "4-component" mixtures were prepared with four compounds from three different groups at a ratio of 2:1:1:1, thus yielding samples with higher compound diversity and evenness. Both interclass and 4-component mixtures contained two compounds from the same metabolic group in order to reflect the observation from natural extracts that compounds of the same biosynthetic pathways tend to co-occur.[23] In addition to the specialized metabolites, eicosanol was also included as a component in some of the mixtures to simulate the effect of long chain fatty acids, which are typically present in plant extracts and contribute to increased peak overlap in the upfield region of the ^{1}H NMR spectrum (δ_H 0.5–2).

Leaf samples of *P. kelleyi* were collected from Yanayacu Biological Station, Napo Province, Ecuador (0°36′ S, 77°53′ W, 2080 m). The most recently expanded leaves were collected for each individual plant, and also young leaves when available. Plants were divided into three age categories according to the indicative number of nodes in their stems, adults (>25 nodes, N=12), saplings (<20 nodes, N=18) and seedlings (<10 nodes, N=17). Samples were dried in air-conditioned laboratory, ground with mortar and pestle to a fine powder. 2 g of this powder were combined with 10 mL of methanol in a screw cap test tube, sonicated for 10 minutes and then filtered to separate the supernatant. This process was repeated with the spent leaf material, the supernatants were combined and transferred to pre-weighed 20 mL scintillation vials. The solvent was then removed under reduced pressure at 30 °C and the extracts were redissolved in deuterated methanol for ^{1}H NMR analysis.

NMR analyses: All spectra were recorded utilizing a Varian 400 spectrometer (399.78 MHz ^{1}H frequency). For ^{1}H NMR acquisitions we adopted the standard parameters set by the instrument, with spectral accumulations of 64 transients (nt) for the prepared mixtures and 256 transients for the plant extracts. Additionally, ^{1}H NMR spectra were acquired for the individual compounds present in the mixtures (nt=128) and complete peak assignments were performed to aid in the structural annotation of network modules. For compounds whose ^{1}H NMR spectra showed severe degree of peak overlap, such as phytosterols and glycosylated molecules, we also acquired ^{13}C (nt=10000), $^{1}H\{^{1}H\}$ gCOSY (nt=4x128), $^{1}H\{^{13}C\}$ gHSQC (nt=4x256) and $^{1}H\{^{13}C\}$ gHMBC (nt=8x256) spectra to assist in this process. The software MNova (version 10.0, Mestrelab Research, Santiago de Compostela, Spain) was utilized for spectral treatment and data extraction. Each ^{1}H NMR spectrum was referenced by the residual solvent peak at δ_H 3.31, then collectively phased (automatic global method), baseline-corrected (polynomial fit of order 3) and binned into intervals of 0.04 ppm within the spectral range δ_H 0.5–12 (bins integrated by average sum).[24] After binning, the solvent peaks were removed, the spectra were normalized to a total of 100 units and the data was exported as a text file for statistical analyses.

Network analyses: All network analyses were performed in the statistical software R (version 3.2.3) using the package *WGCNA*, which executes all steps required for network construction with a single function.[25] Initially, the nxm data matrix containing m samples and n spectral bins is transformed into a correlation matrix of nxn nodes, in which s_{ij} (Eq. 1.1) represents the degree of similarity between the intensity profiles of spectral bins i and j across samples. A thresholding function was then applied to these values to eliminate baseline correlations from the analysis, resulting in an adjacency matrix, where a_{ij} denotes the connection strength (edge) between two nodes. *Unweighted* networks are obtained by setting a hard threshold value τ beneath which any correlations are eliminated (Eq. 1.2). This type of network is not compatible with the continuous nature of values expressed in NMR data and it may lead to loss of information,[25] so we applied a soft-thresholding power function (Eq. 1.3) to generate a *weighed* network. It has been observed that robust biological networks largely follow a *Scale-free Topology*,[26] in which the frequency distribution of node connectivity follows a decaying power law (Eq. 1.3), so we utilized this criterion to select β. In practical terms, a plot of β *versus* R^2 (the fitting index with a scale-free topology model) is generated and β is chosen at the plateau, the point at which network stability is established (Figure 1.1). We also followed the recommendation that only values of β resulting in $R^2 \geq 0.8$ should be considered.[20]

$$s_{ij} = corr(x_i, x_j) \qquad \text{Eq. 1.1}$$

$$a_{ij} = \begin{cases} 1 \; if \; s_{ij} \geq \tau \\ 0 \; if \; s_{ij} < \tau \end{cases} \qquad \text{Eq. 1.2}$$

$$a_{ij} = \left|s_{ij}\right|^{\beta} \qquad \text{Eq. 1.3}$$

Hierarchical clustering analysis (HCA) was used to detect modules in the networks with a minimum size of three nodes and a merging criterion of 75% similarity between modules. Each module was then assigned an arbitrary color, where "grey" was reserved for unassigned nodes (not an actual module). Through singular value decomposition, an eigenvector was generated for each module (the equivalent of a weighted average of its nodes) and used to obtain a module eigenvalue for each spectrum. With the mixtures data, we calculated the Pearson correlation values between module eigenvalues and compound concentrations to verify specific compound-module associations, and then quantified module coverage of compound signals for significant associations ($p \leq 0.05$). To achieve that, we first estimated the number of *distinguishable* compound peaks by visually investigating their individual spectra. Hydrogen resonances within 0.05 ppm of each other were considered part of the same signal (Figure 1.2) and diastereotopic methylene peaks were only counted once. Following, we calculated the number of peaks detected by a module's nodes, where complete overlaps were attributed a value of 1 and nodes within 0.1 ppm of a compound signal were valued as 0.75 (Figure 1.2). Finally, we calculated module coverage as the ratio of sum of node matches to the total number of distinguishable compound peaks.

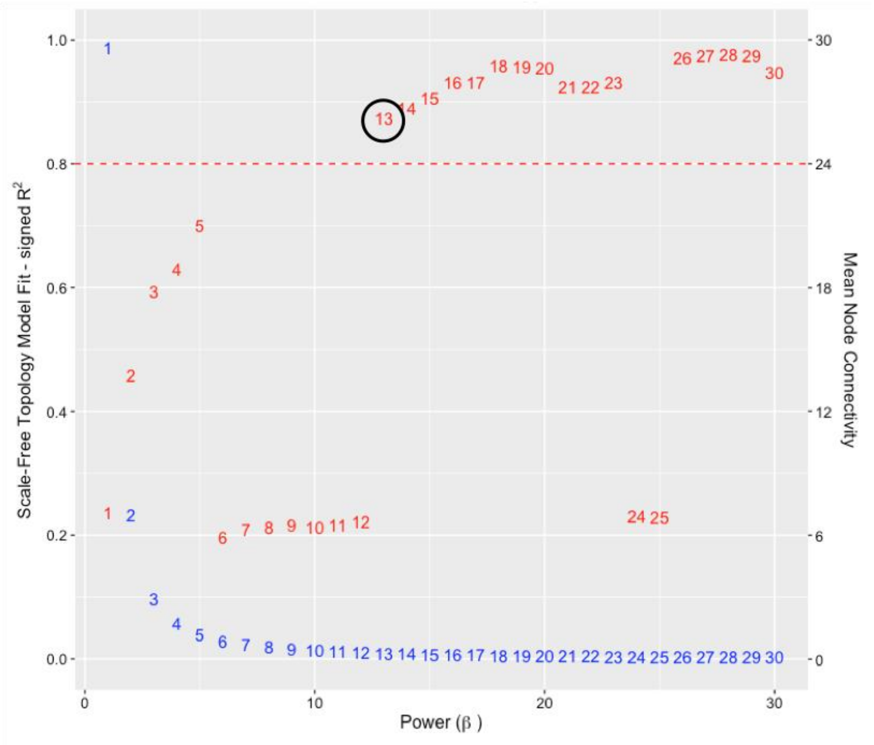


Figure 1.1. Network topology plot. Numbers indicate values for threshold parameter β. Red numbers are positioned according to the resulting fitness with a scale-free topology model (left axis), while blue numbers represent the resulting node connectivity (right axis). In this example, β = 13 is chosen for an appropriate network.

To verify module-to-age associations in the *P. kelleyi* data, we performed a Multivariate Analysis of Variance (MANOVA) with module eigenvalues as dependent variables and developmental stage (adult, sapling, and seedling) as predictor variables. Modules that demonstrated a significant main effect on developmental stage were analyzed using Tukey’s HSD post-hoc tests to determine the developmental stage associated to those modules.

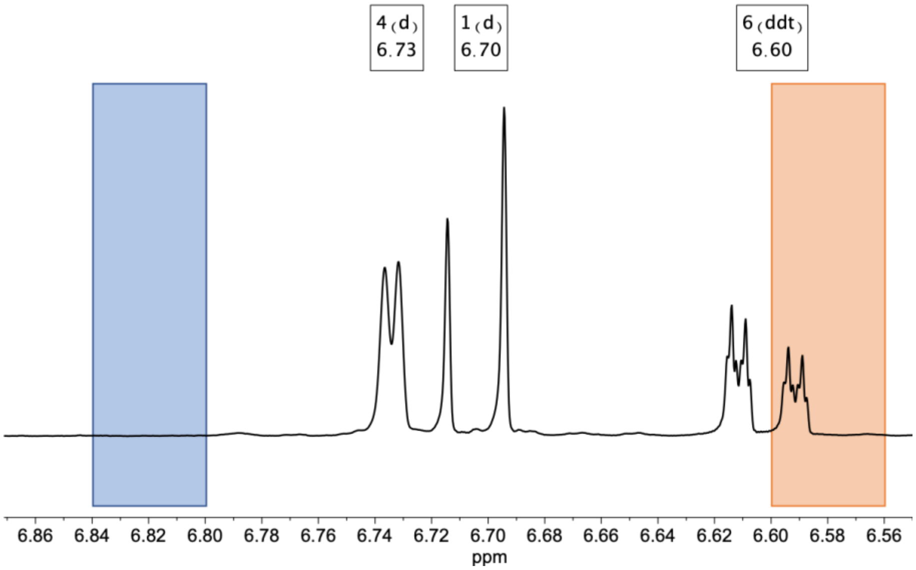


Figure 1.2. Example of module discrimination of compound peaks. Modules are represented by the shaded boxes. In this spectrum, resonances 4 and 1 are considered indistinguishable because they are less than 0.05 ppm apart. The *orange* module has complete overlap with resonance 6, so their association is valued 1. Both *orange* and blue *modules* detect the group of peaks 4/1 by 0.75 because they are within 0.1 ppm of those resonances. Module *blue* is not considered to detect the peak 6.

1.3 Results

Prepared mixtures: To assess the efficacy of the method to detect structural features at different levels of sample complexity, we performed the network analysis in three different sets of samples: intraclass mixtures, interclass with 4-component, and then 4-component only. All analyses resulted in roughly the same number of modules and average number of nodes by module, although the number of unassigned nodes was larger for the 4-component network (Table 1.1). That overall convergence was also expressed by coherent associations between modules and specific compounds or structural features across the three networks, with node-module membership being retained in most cases. As mentioned in the last section, arbitrary colors were assigned to each module, so for the ease of discussion we labeled the modules generated in each analysis according to the consistent structural features they represented (Appendix A.1).

Table 1.1. Network results obtained for each group of prepared mixtures.

	Intraclass	**Interclass and 4-component**	**4-component**
β-parameter	16	6	10
Number of modules	21	23	21
Average nodes/module (± sd)	9 (± 4)	8 (± 4)	8 (± 3)
Unassigned nodes	31.6%	29.8%	39.3%

For the intraclass mixtures, the technique was effective at detecting common structural features among relatively homogeneous samples. The information captured by the modules was strongly influenced by mixture composition, thus emphasizing structural motifs common to compounds of the same class (Figure 1.3). Due to the high degree of compound co-occurrence in these mixtures, we also observed indirect module-compound associations that were not supported by the module's chemical shifts. For example, the module PHP3 described structural features present in the phenylpropanoids eugenol and resveratrol which are not present in PBA (Figure 1.3). However, since PBA was present in all mixtures containing eugenol, it was also significantly associated with PHP3. Notwithstanding a compositional bias towards intra-class associations, the network analysis was still able to recognize structural features shared across classes of compounds, such as the motifs present in prenyl groups of PBA and triterpenes under the module STR2 (Figure 1.3).

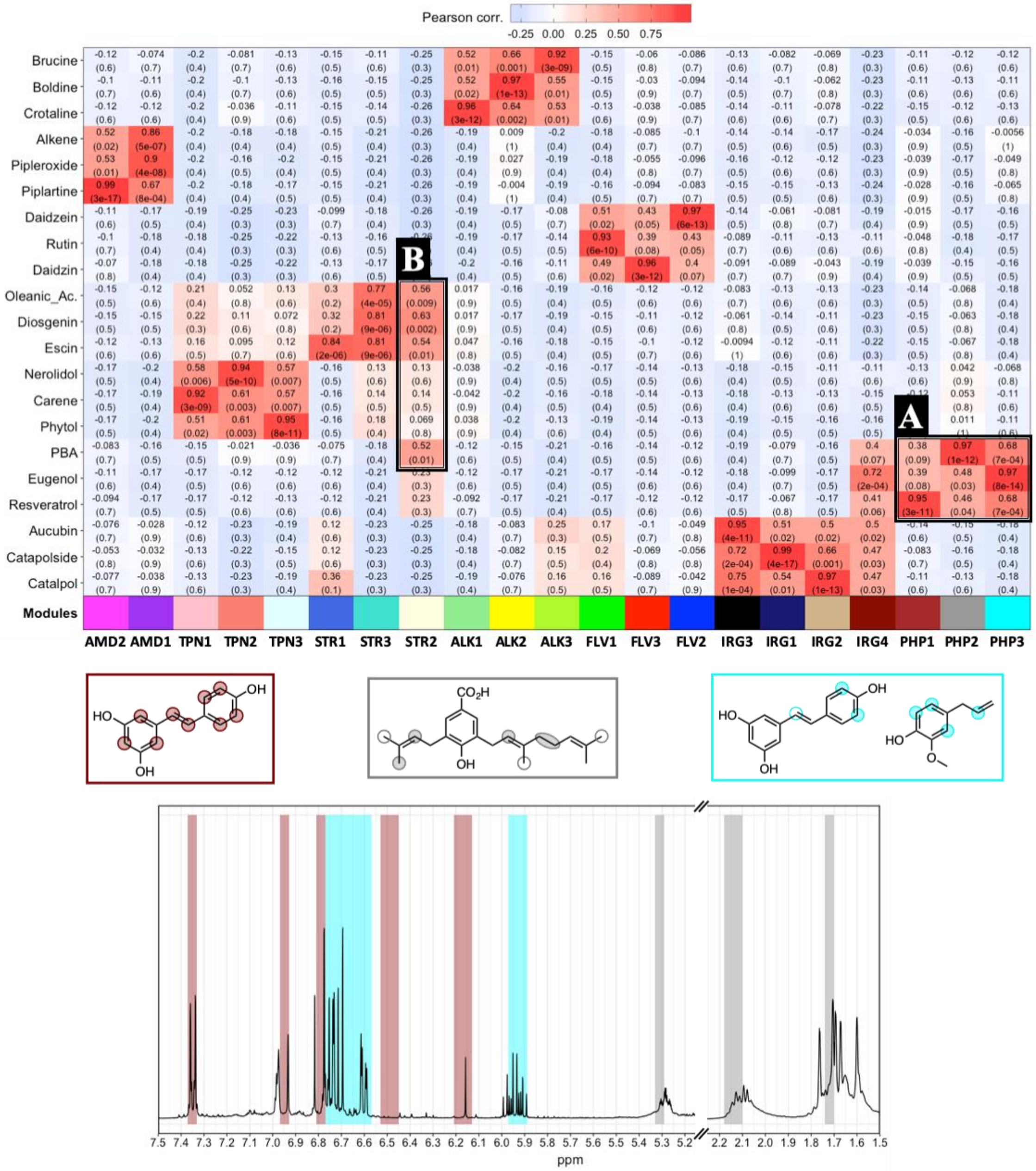


Figure 1.3. Module-compound heatmap from the intraclass network. Insert A (top) shows significant module associations with phenylpropanoids. Modules are represented by the color bar, with consolidated names shown. Structural featured represented in these associations are highlighted in the color-coded boxes (middle) and shaded regions of the spectrum (bottom). Insert B shows the association of module STR2 with compounds from different classes.

Compound co-occurrence had a lesser effect in the network analysis of interclass and 4-component mixtures. These mixtures were more representative of the complexity present in natural extracts and provided a better parameter for assessing the method's sensitivity to compound concentrations and peak overlap. The resulting networks not only retained the compound-class specific modules observed in the intraclass analysis, but they were also characterized by molecular features shared among compounds from distinct and converging biosynthetic pathways. For example, in the interclass and 4-component analysis, the module PHP3 captured the phenylpropanoid-derived aromatic ring present in the flavonoids genistein and daidzein, and in the stilbene resveratrol (Figure 1.4). Flavonoids and iridoids also shared a common module – GLC1 – due to their glycosylated moieties (Appendix A.1).

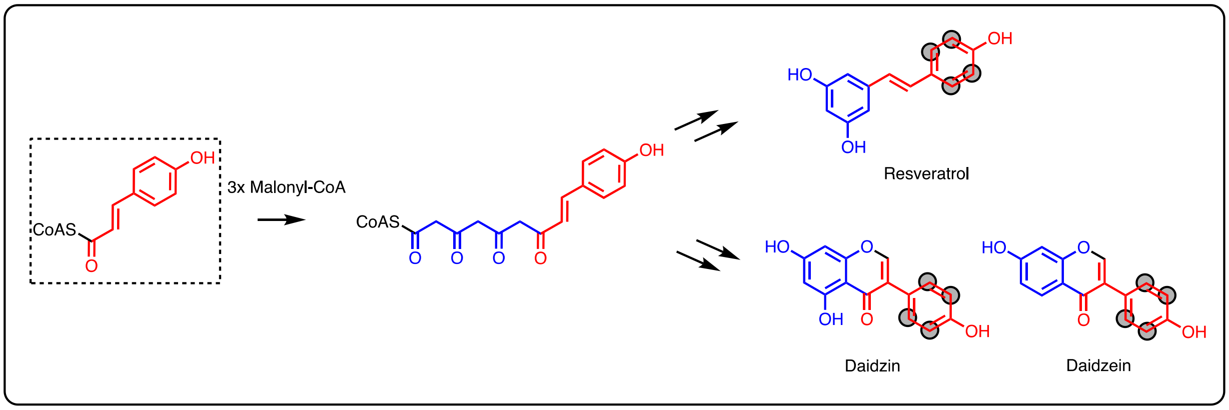


Figure 1.4. Biosynthetic origin of shared structural features from module PHP3. Fragments colored in red originate from the phenylpropanoid pathway, while the blue features come from the polyketide pathway. The black circles represent resonances identified by PHP3.

A particular case of structural similarity was observed in modules involved with amides and alkaloids, which were interconnected in the network through specific proton resonances vicinal to nitrogen atoms (Figure 1.5). In the interclass analysis, three amides and the alkaloid brucine were associated with the module AMD1 due to proton resonances in the α- and β-position of the nitrogen atoms. This module was connected to two alkaloid-related modules, and together they composed a cluster (or meta-module) that was indicative of nitrogen-containing compounds (Figure 1.5). Each of these modules also retained attributes that were particular to their most significant compounds, which accounts for their connections with modules outside of the cluster. For example, ALK3 represents the alkaloid brucine, but it was linked to the phenylpropanoid module PHP2 (green nodes in Figure 1.5) due to the aromatic proton resonances of that compound.

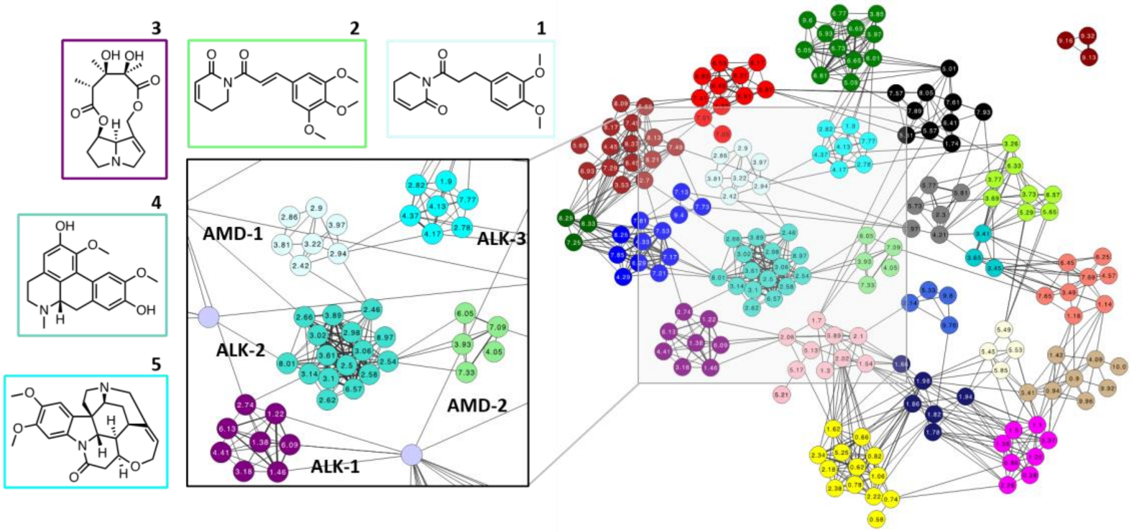


Figure 1.5. Nitrogen-related modules from the interclass network. The expanded region of the network shows the connectivity between modules that represented alkaloids and amides. Their most important compounds are indicated in the corresponding color-coded rectangles: pipleroxide (1), alkene amide (2), crotaline (3), boldine (4) and brucine (5).

Another interesting result demonstrates that the network analysis provided evidence for the interaction between compounds. Particularly, the phenolic peaks in resveratrol were generally broad and undetectable due to proton exchange with the solvent, but the nearly negligible resonances at δ_H 9.10 and δ_H 9.30 from mixtures containing resveratrol, escin and oleanic acid were captured by module PHP4. Phenolic peaks are sensitive to intermolecular interactions based on hydrogen-bonding, as they reduce proton exchange and improve peak sharpness.[27] We verified experimentally that these peaks only appeared in the presence of escin, suggesting that resveratrol has hydrogen-bonding interactions with the glycosylated portions of that compound in solution (Figure 1.6).

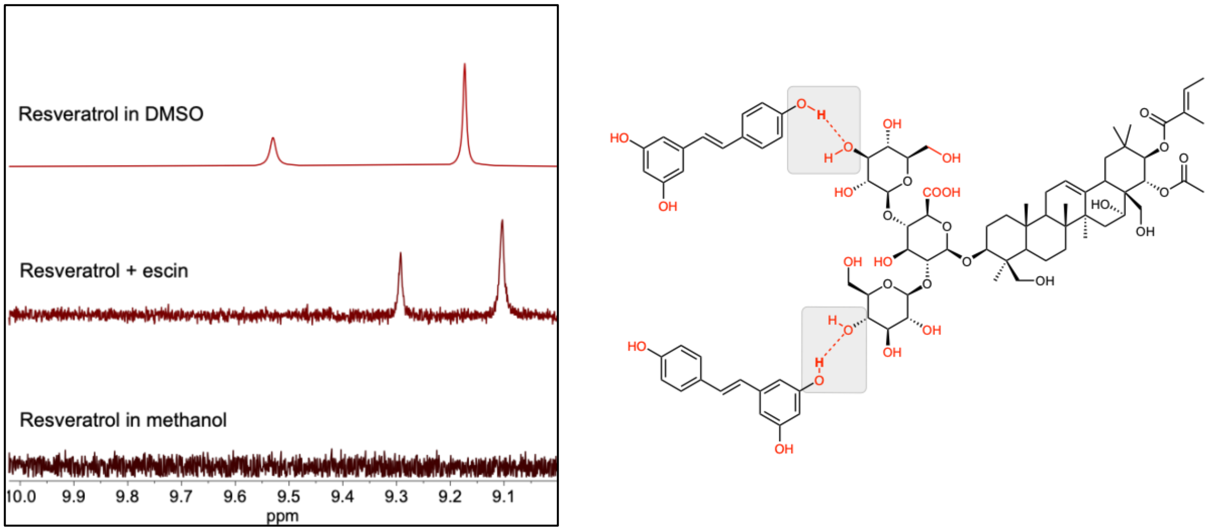


Figure 1.6. Phenolic hydrogen peaks detected by module PHP4. Resonances at δ_{H} 9.10 and δ_{H} 9.30 are absent for resveratrol in a protic solvent (methanol) but become evident in the presence of escin as a result of stabilizing hydrogen-bonding interactions (grey boxes) between the hydroxyl groups of these molecules (highlighted in red). Those peaks are also present when resveratrol is analyzed in an aprotic solvent (dimethyl sulfoxide), which minimizes exchange of labile protons.

Given that each network analysis generated module-compound associations that were described by different collections of peaks, we evaluated the efficacy of the approach to identify a compound across different degrees of sample complexity. However, establishing a definitive measurement of compound identity can be challenging, considering that not every resonance is relevant to identify a compound, particularly because several structural motifs are ubiquitous to the molecules utilized in this study. For that reason, we opted to define compound identity as the complete set of its resonance peaks, and to calculate network accuracy as the fraction of a

Table 1.2. Compound detection accuracy for the three networks of prepared mixtures

		Relative Accuracy		
Compounds		**Intraclass**	**Interclass**	**4 compounds**
Alkaloids	Brucine	0.33	0.35	0.35
	Boldine	0.58	0.80	0.68
	Crotaline	0.32	0.57	-
	Caffeine	-	0.19	0.19
Amides	Alkene amide	0.64	0.56	0.44
	Piplartine	0.88	0.63	0.72
	Pipleroxide	0.58	0.48	0.38
Iridoid glycosides	Aucubin	-	0.34	-
	Catalposide	0.28	0.28	0.41
	Catalpol	0.23	0.36	0.36
Cardiac glycosides	Digitoxin	-	0.21	0.25
Furanocoumarins	Bergapten	-	1.00	0.88
	Imperatorin	-	0.88	-
	Xanthotoxin	-	0.67	0.83
Flavonoids	Rutin	0.54	0.45	0.54
Isoflavonoid	Daidzein	0.95	0.95	0.00
	Daidzin	0.58	0.78	0.25
	Genistein	-	0.60	0.80
Terpenoids	Carene	0.86	0.86	0.86
	Phytol	0.53	0.53	0.72
	Nerolidol	0.69	0.94	0.59
Triterpeinoid saponins	Escin	0.26	0.17	0.31
Saponin	Diosgenin	0.56	0.49	-
	Oleanolic Acid	0.56	0.46	-
Phenylpropenoids	Eugenol	0.43	0.86	0.71
	Resveratrol	1.00	0.83	0.83
	Prenylated Benzoic Acid	0.48	0.33	-
Phytosterols	Sitosterol	-	0.77	0.44
	Stigmasterol	-	0.50	0.38

compound's total peaks that is detected by its most correlated module (Table 1.2). The average overall accuracy was similar across all analyses (0.56 ± 0.05, 0.58 ± 0.05, and 0.52 ± 0.05 for intraclass, interclass, and 4-component mixtures respectively), indicating that about 55% of the signals were captured by the most representative module of a compound in each analysis. In some cases, such as with digitoxin and escin, the modules had a relatively low accuracy because the compounds had a large number of resonances and only the protons from specific structural features were identified.

***P. kelleyi* samples:** Network analysis on the spectra obtained from *P. kelleyi* extracts resulted in 19 modules, with a broad variation in module size (10 ± 9 nodes/module) and only three unassigned nodes. Due to the complexity of crude extracts, more regions of the spectrum were occupied with detectable resonances in comparison with the prepared mixtures, thus contributing for the detection of larger clusters of co-occurring peaks. Through HCA of module eigenvalues, we detected three clusters of samples with apparent enrichment for specific developmental stages (Figure 1.7). These findings were corroborated through MANOVA, which indicated a strong association of about half of the modules with specific life stages (Wilks $\lambda = 0.09$, $p \leq 0.01$). Six of those modules were specifically associated with seedings (Tukey's HSD, $p \leq 0.05$), and they converged with modules from the mixtures networks that described the amide piplartine. Some of the resonances that were not directly associated with piplartine followed the same peak patterns of that compound, suggesting the presence of closely related amides in the seedlings extracts.[28]

A more limited number of modules had significant associations with adults and saplings, and their signature signals were overlapped, indicating that chemical distinction between these two groups was not so evident. Nonetheless, the peaks representing these groups were identified as part of a previously reported chromene compound along with its dimeric benzopyran (Figure 1.7).[29] No piplartine peaks were detected in the adult samples investigated, and complementarily, no chromene peaks were observed in the seedlings. In saplings, however, these compounds were present as a gradient of combinations, demonstrating a clear transition between the metabolic states expressed in the other two extreme developmental stages.

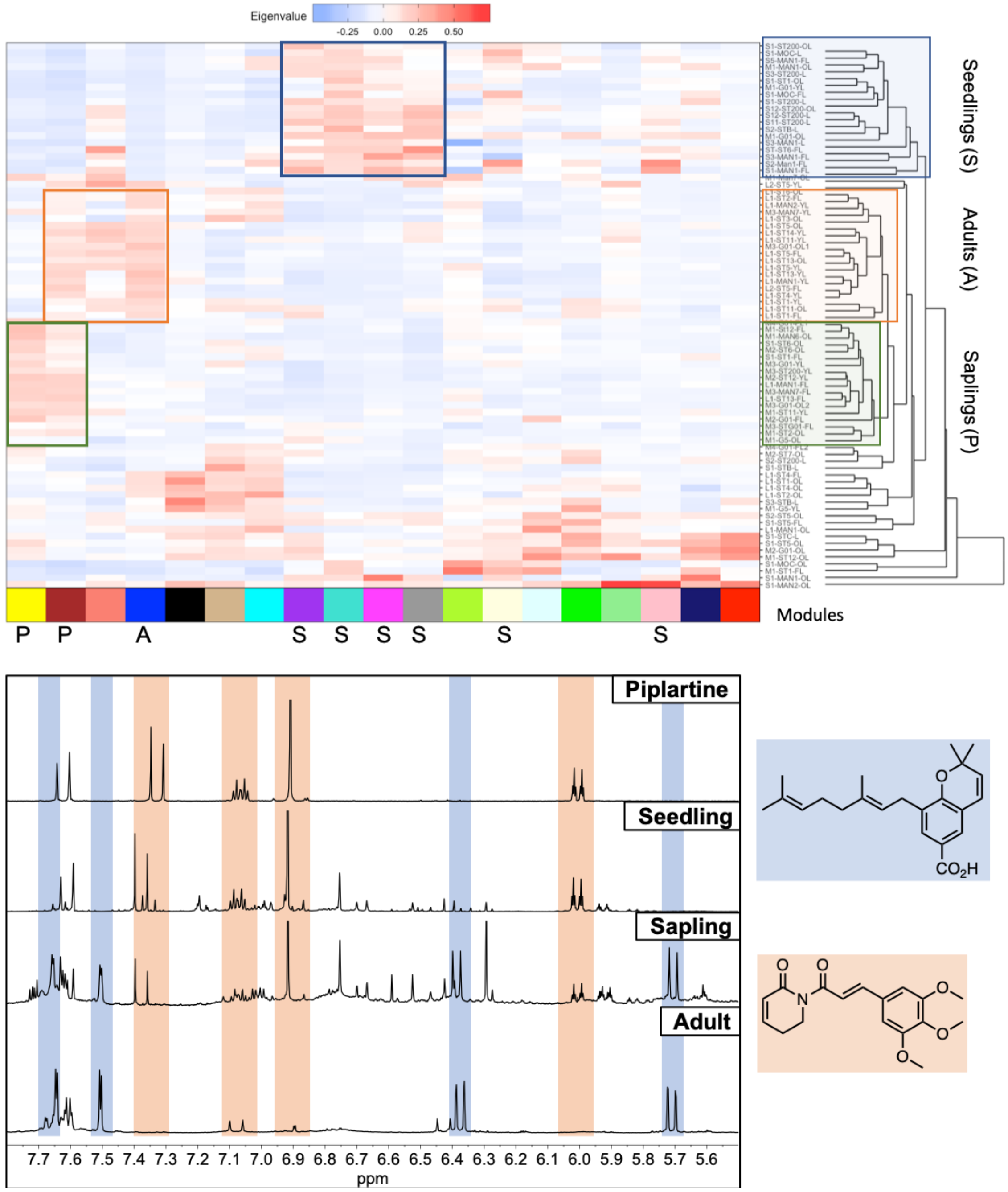


Figure 1.7. Module-to-age associations from the *P. kelleyi* network. Top: heatmap of module eigenvalues for *P. kelleyi* samples, which are organized in the dendrogram according to their module similarity. Significant modules (colored bar) are indicated according to the plant age their mostly significantly represent (A–adults, P–saplings, S–seedlings). Bottom: spectral features indicating the presence of piplartine (orange) and the chromene compound (blue).

1.4 Discussion

The methodology introduced by this study demonstrates the potential of ^{1}H NMR spectroscopy to expose chemical markers of interest from complex mixtures. Through the use of prepared mixtures of specialized metabolites, we verified the reliability of the analysis for linking groups of proton resonances to structurally similar compounds and identifying class-specific modules for metabolites. The results obtained from the intraclass mixtures also demonstrate that the technique identified metabolites connected by co-expression profiles, even with limited structural similarity. This could be a useful feature for the identification of biosynthetically related compounds from complex extracts, especially in cases where the more derivatized compounds are present in minimal concentrations.

By varying the complexity of the prepared mixtures, we gained insight into compound associations driven by shared secondary structural features (prenyl and glycosyl moieties in aromatic compounds) and core structural elements (Nitrogen-vicinal resonances in amides and alkaloids). We also gathered evidence for specific peaks originating from compound associations through hydrogen bonding, which not only shows the degree of sensitivity of this technique, but also demonstrates its utility to the study of the synergistic effects of specialized metabolites. Moreover, the networks produced the same general module-compound associations across varying degrees of mixture complexity, with a reasonable degree of structural coverage for all the compounds. These results were based on a compound set that represents only part of the structural diversity found in plant secondary metabolites, but they demonstrate that this technique could be also effective in the study of natural samples.

In the study of *P. kelleyi*, we combined the variable reduction feature from the network analysis with classic statistical methods to elucidate the effects of plant ontogeny on chemical composition. Due to the spectral complexity presented by the crude extracts, the modules generated in this analysis were more robust and contained a larger fraction of the inputted information comparatively to those obtained from the artificial mixtures. Based on that observation, we predicted that the largest modules would represent the baseline metabolites expressed through plant development, while age-specific information should be represented by the smaller modules. However, the results demonstrated the opposite trend, with the largest modules accounting for age-specific compounds. Upon inspection of spectral data, we verified that this effect was a result of the co-occurrence of closely related compound in the extracts. For instance, *turquoise*, which was the largest module and the most significantly associated with seedlings, included contiguous regions of the spectrum containing peaks representative of piplartine and its closely related analogs. While these results were desirable to gain a holistic perspective of the metabolites present in each stage, we project that with a more rigorous β threshold, the network analysis could generate smaller modules with higher compound specificity.

Ontogeny is a driving element for plasticity in specialized metabolism,[30, 31] and the age-module-compound associations revealed by this study provided strong evidence to attribute plant development as the cause of dramatic changes in the chemical profile of *P. kelleyi*. We initially hypothesized that the observed changes were related to the defensive roles of the metabolites expressed in each stage, thus reflecting the different sources of mortality that challenge these plants. Amides are produced by several species of *Piper*, where they function as antifungal and anti-herbivore defensive compounds.[32-36] Thus, the production of piplartine and other amides is particularly important for seedlings, which are especially vulnerable to generalist herbivores and fungal pathogens. Complementarily, previous studies on *P. kelleyi* found that plants producing PBA and chromene had a lower diversity of specialist caterpillars, suggesting that these compounds play a defensive role against herbivory in adult plants.[37] Following the publication of these results, a study was published that attributed the changes in metabolism of *P. kelleyi* as a result of plant placement in the canopy.[38] Through an elaborated field experiment, the authors demonstrated that access to sunlight drives a growth-defense tradeoff, in which taller plants invest the available resources towards the production of biomass through photosynthesis, while also directing some of its incorporated carbon into photoactive defensive chromenes. With a more limited access to sunlight, plants in the forest understory need to invest more resources to preserve biomass against predation, so their phenylpropanoid metabolism is repurposed for the production of piplartine. By estimating plant's age from the number of nodes in its main stem, we introduced a confounding variable in our study, thus leading to incorrect inferences about ontogeny.

1.5 Conclusion and future directions

We demonstrated that network analysis of ^{1}H NMR spectra can summarize complex phytochemical profiles into a few chemotype-related variables. These variables may represent groups of compounds or shared structural features, and in conjunctions with other statistical techniques, they can help connecting the chemical profile with biological information. Thus, this approach facilitates the examination of the consequences of complete biosynthetic products, as opposed to focusing on the effects of single compounds. Moreover, the experiment with prepared mixtures helped establishing a reference library that can be useful to annotate modules obtained from the network analyses of other sets of samples, as exemplified in the study of *P. kelleyi*. Thus, we expect that by aggregating more data to this analysis, particularly from chemically characterized biological samples, one can create a more robust model that is capable to identify a broader range of structural features. With that perspective, this analysis can be expanded beyond the realm of phytochemistry to incorporate compounds typically produced by other taxa and support the investigation of chemically mediated effects in other systems. From the perspective of natural products chemistry, this approach has the potential to facilitate the prioritization of samples from large field collections, allowing one to select extracts characterized by the most promising modules for subsequent isolation and structure determination. For chemical ecology, it provides a tool for quantifying entire arrays of chemical defenses that can be used as predictors or response variables in statistical models. Module importance and overall network parameters can be examined in response to manipulation of resources, or they can be mapped into phylogenies to address interesting questions about the evolution of metabolism across taxa and the origins of biodiversity.

1.6 References

1. Richards Lora, A.; Dyer Lee, A.; Oliveira, C.; Rumbaugh, A.; Wallace Ian, S.; Dodson Craig, D.; Jeffrey Christopher, S.; Urbano-Munoz, F., Shedding Light on Chemically Mediated Tri-Trophic Interactions: A (1)H NMR Network Approach to Identify Compound Structural Features and Associated Biological Activity. *Front Plant Sci* **2018,** *9*, 1155.

2. Beutler, J. A., Natural Products as a Foundation for Drug Discovery. *Curr Protoc Pharmacol* **2019,** *86* (1), e67.

3. Richards, L. A.; Glassmire, A. E.; Ochsenrider, K. M.; Smilanich, A. M.; Dodson, C. D.; Jeffrey, C. S.; Dyer, L. A., Phytochemical diversity and synergistic effects on herbivores. *Phytochem. Rev.* **2016,** *15* (6), 1153-1166.

4. Fraenkel, G. S., The raison d'etre of secondary plant substances. These odd chemicals arose as a means of protecting plants from insects and now guide insects to food. *Science (Washington, DC, U. S.)* **1959,** *129*, 1466-70.

5. Ehrlich, P. R.; Raven, P. H., Butterflies and Plants: A Study in Coevolution. *18* (4), 586-608.

6. Edger, P. P.; Heidel-Fischer, H. M.; Bekaert, M.; Rota, J.; Glockner, G.; Platts, A. E.; Heckel, D. G.; Der, J. P.; Wafula, E. K.; Tang, M.; Hofberger, J. A.; Smithson, A.; Hall, J. C.; Blanchette, M.; Bureau, T. E.; Wright, S. I.; de Pamphilis, C. W.; Schranz, M. E.; Barker, M. S.; Conant, G. C.; Wahlberg, N.; Vogel, H.; Pires, J. C.; Wheat, C. W., The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (27), 8362-8366.

7. Wishart, D. S., NMR metabolomics: A look ahead. *J. Magn. Reson.* **2019,** *306*, 155-161.

8. Wolfender, J.-L.; Marti, G.; Thomas, A.; Bertrand, S., Current approaches and challenges for the metabolite profiling of complex natural extracts. *J. Chromatogr. A* **2015,** *1382*, 136-164.

9. Liu, W.; Song, Q.; Cao, Y.; Xie, N.; Li, Z.; Jiang, Y.; Zheng, J.; Tu, P.; Song, Y.; Li, J., From 1H NMR-based non-targeted to LC-MS-based targeted metabolomics strategy for in-depth chemome comparisons among four Cistanche species. *J. Pharm. Biomed. Anal.* **2019,** *162*, 16-27.

10. Boiteau, R. M.; Hoyt, D. W.; Nicora, C. D.; Kinmonth-Schultz, H. A.; Ward, J. K.; Bingol, K., Structure elucidation of unknown metabolites in metabolomics by combined NMR and MS/MS prediction. *Metabolites* **2018,** *8* (1), 8/1-8/12.

11. Richards, L. A.; Dyer, L. A.; Forister, M. L.; Smilanich, A. M.; Dodson, C. D.; Leonard, M. D.; Jeffrey, C. S., Phytochemical diversity drives plant-insect community diversity. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (35), 10973-10978.

12. Kim, H. K.; Choi, Y. H.; Verpoorte, R., NMR-based plant metabolomics: where do we stand, where do we go? *Trends Biotechnol.* **2011,** *29* (6), 267-275.

13. Girelli, C. R.; Angile, F.; Del Coco, L.; Migoni, D.; Zampella, L.; Marcelletti, S.; Cristella, N.; Marangi, P.; Scortichini, M.; Fanizzi, F. P., 1H NMR metabolite fingerprinting analysis reveals a disease biomarker and a field treatment response in Xylella fastidiosa subsp. pauca-infected olive trees. *Plants* **2019,** *8* (5), 115.

14. Hu, B.; Yue, Y.; Zhu, Y.; Wen, W.; Zhang, F.; Hardie, J. W., Proton nuclear magnetic resonance- spectroscopic discrimination of wines reflects genetic homology of several different grape (V. vinifera L.) cultivars. *PLoS One* **2015,** *10* (12), e0142840/1-e0142840/16.

15. Sulaiman, F.; Azam, A. A.; Bustamam, M. S. A.; Fakurazi, S.; Abas, F.; Lee, Y. X.; Ismail, A. A.; Faudzi, S. M. M.; Ismail, I. S., Metabolite profiles of red and yellow watermelon (Citrullus lanatus) cultivars using a 1H NMR metabolomics approach. *Molecules* **2020,** *25* (14), 3235.

16. Yun, D.-Y.; Kang, Y.-G.; Yun, B.; Kim, E.-H.; Kim, M.; Park, J. S.; Lee, J. H.; Hong, Y.-S., Distinctive Metabolism of Flavonoid between Cultivated and Semiwild Soybean Unveiled through Metabolomics Approach. *J. Agric. Food Chem.* **2016,** *64* (29), 5773-5783.

17. Crusemann, M.; O'Neill Ellis, C.; Larson Charles, B.; Melnik Alexey, V.; Floros Dimitrios, J.; Jensen Paul, R.; Dorrestein Pieter, C.; Moore Bradley, S.; da Silva Ricardo, R., Prioritizing Natural Product Diversity in a Collection of 146 Bacterial Strains Based on Growth and Extraction Protocols. *J Nat Prod* **2017,** *80* (3), 588-597.

18. Garg, N.; Kapono, C.; Lim, Y. W.; Koyama, N.; Vermeij, M. J.; Conrad, D.; Rohwer, F.; Dorrestein, P. C., Mass spectral similarity for untargeted metabolomics data analysis of complex mixtures. *Int J Mass Spectrom* **2015,** *377*, 719-717.

19. Han, J.-D. J., Understanding biological functions through molecular networks. *Cell Res.* **2008,** *18* (2), 224-237.

20. Zhang, B.; Horvath, S., A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **2005,** *4* (1), No pp given.

21. Futuyma, D. J.; Keese, M. C.; Rosenthal, G. A.; Berenbaum, M. R., Chapter 12 - Evolution and Coevolution of Plants and Phytophagous Arthropods. In *Herbivores: Their Interactions with Secondary Plant Metabolites (Second Edition)*, Academic Press: San Diego, 1992; pp 439-475.

22. Martin, A. C.; Pawlus, A. D.; Jewett, E. M.; Wyse, D. L.; Angerhofer, C. K.; Hegeman, A. D., Evaluating solvent extraction systems using metabolomics approaches. *RSC Adv.* **2014,** *4* (50), 26325-26334.

23. Gershenzon, J.; Fontana, A.; Burow, M.; Wittstock, U.; Degenhardt, J., Mixtures of plant secondary metabolites: metabolic origins and ecological benefits. . In *The Ecology of Plant Secondary Metabolites: From Genes to Global Processes*, Iason, G. R.; Dicke, M.; Hartley, S. E., Eds. Cambridge University Press: Cambridge, 2012; pp 56-77.

24. Verpoorte, R.; Choi, Y. H.; Kim, H. K., NMR-based metabolomics at work in phytochemistry. *Phytochem. Rev.* **2007,** *6* (1), 3-14.

25. Langfelder, P.; Horvath, S., WGCNA: an R package for weighted gene co-expression network analysis. *BMC Bioinf.* **2008,** *9*, No pp given.

26. Bergmann, S.; Ihmels, J.; Barkai, N., Similarities and differences in genome-wide expression data on six organisms. *PLoS Biol.* **2004,** *2* (1), 85-93.

27. Charisiadis, P.; Kontogianni, V. G.; Tsiafoulis, C. G.; Tzakos, A. G.; Siskos, M.; Gerothanassis, I. P., 1H NMR as a structural and analytical tool of intra- and intermolecular hydrogen bonds of phenol-containing natural products and model compounds. *Molecules* **2014,** *19* (9), 13643-82.

28. Maxwell, A.; Rampersad, D., A new dihydropiplartine and piplartine dimer from Piper rugosum. *J. Nat. Prod.* **1991,** *54* (4), 1150-2.

29. Jeffrey, C. S.; Leonard, M. D.; Glassmire, A. E.; Dodson, C. D.; Richards, L. A.; Kato, M. J.; Dyer, L. A., Antiherbivore prenylated benzoic acid derivatives from Piper kelleyi. *J. Nat. Prod.* **2014,** *77* (1), 148-153.

30. Koricheva, J.; Barton, K. E., Temporal changes in plant secondary metabolite production. 34-55.

31. Gaia, A. M.; Yamaguchi, L. F.; Jeffrey, C. S.; Kato, M. J., Age-dependent changes from allylphenol to prenylated benzoic acid production in Piper gaudichaudianum Kunth. *Phytochemistry (Elsevier)* **2014,** *106*, 86-93.

32. Dyer, L. A.; Letourneau, D. K.; Dodson, C. D.; Tobler, M. A.; Stireman, J. O.; Hsu, A., Ecological Causes and Consequences of Variation in Defensive Chemistry of a Neotropical Shrub. *85* (10), 2795-2803.

33. Navickiene, H. M. D.; Alecio, A. C.; Kato, M. J.; Bolzani, V. S.; Young, M. C. M.; Cavalheiro, A. J.; Furlan, M., Antifungal amides from Piper hispidum and Piper tuberculatum. *Phytochemistry* **2000,** *55* (6), 621-626.

34. Vasques da Silva, R.; Debonsi Navickiene, H. M.; Kato, M. J.; Bolzani, V. d. S.; Meda, C. I.; Young, M. C. M.; Furlan, M., Antifungal amides from Piper arboreum and Piper tuberculatum. *Phytochemistry* **2002,** *59* (5), 521-527.

35. Marques, J. V.; Kitamura, R. O. S.; Lago, J. H. G.; Young, M. C. M.; Guimaraes, E. F.; Kato, M. J., Antifungal Amides from Piper scutifolium and Piper hoffmanseggianum. *J. Nat. Prod.* **2007,** *70* (12), 2036-2039.

36. Marques, J. V.; de Oliveira, A.; Raggi, L.; Young, M. C. M.; Kato, M. J., Antifungal activity of natural and synthetic amides from Piper species. *J. Braz. Chem. Soc.* **2010,** *21* (10), 1807-1813.

37. Glassmire Andrea, E.; Jeffrey Christopher, S.; Forister Matthew, L.; Parchman Thomas, L.; Jahner Joshua, P.; Richards Lora, A.; Smilanich Angela, M.; Morrison Colin, R.; Dodson Craig, D.; Miller Jim, S.; Dyer Lee, A.; Leonard Michael, D.; Nice Chris, C.; Wilson Joseph, S.; Walla Thomas, R.; Villamarin-Cortez, S.; Simbana, W.; Salagaje Luis, A.; Tepe Eric, J., Intraspecific phytochemical variation shapes community and population structure for specialist caterpillars. **2016,** *212* (1), 208-19.

38. Glassmire Andrea, E.; Snook Joshua, S.; Philbin, C.; Jeffrey Christopher, S.; Richards Lora, A.; Dyer Lee, A., Proximity to canopy mediates changes in the defensive chemistry and herbivore loads of an understory tropical shrub, Piper kelleyi. *Ecol Lett* **2019,** *22* (2), 332-341.

Chapter 2: ^{1}H NMR Network Analysis Applied to the Identification and Isolation of Antifungal Compounds from Species of the Genus *Piper*

2.1 Introduction

As it was discussed in Chapter 1, plant specialized metabolism is the evolutionary result of a dynamic and complex interplay of selective pressures. Fungal pathogens are an important element in that equation, accounting for more than 13000 unique species and over 75000 plant-fungal combinations only in the United States.[1] Some episodes in history demonstrate the devastating effect of fungal diseases to domesticated crops, such as the potato blight that led to the Great Hunger in an 1840s Ireland,[2] but the effect of these pathogens in natural systems still remains underestimated. A recent meta-analysis revealed that fungal diversity is particularly prominent in higher latitude forests, contrary to the trend observed for plants and arthropods, which are more abundant in the tropical zone.[3] According to the study, fungal distribution was primarily associated with climate factors, particularly temperature, but symbiotic taxa showed a narrower climate niche than the parasitic counterparts. It has also been shown that deforestation has a detrimental effect saprophytical soil fungi, while benefitting pathogenic species.[4] It is thus increasingly evident that climate change and forest disturbance will lead to the prevalence of phytopathogens and to the dramatic alteration in the structure of ecosystems.[5]

The basic mechanisms of defense against fungi in plants are rather unspecific to these pathogens and must have originated in the early stages of adaptation of land plants.[6] For example, the cellular deposition of callose prevents fungal invasion,[7] but it is also a primary constituent responsible for strengthening plant structure against the effects of gravity; flavonoids which are implied in defense against some fungal pathogens may also have been fundamental for the protection of early plants against UV radiation. However, breeches in these mechanisms of defense due to natural variation of pathogens and hosts must have led to events of infection and the consequential development of specific plant-pathogen defenses.[6] As an example of specificity, tomato produces the antifungal saponin α-tomatine, but the tomato pathogen *Septoria lycopersici* produces tomatinase, an extracellular enzyme that hydrolyses α-tomatine into a significantly less toxic compound.[8] Consequentially, we can speculate that specific chemical defenses must be more prevalent in habitats that favor the fungus,[9] and that variations of these chemical traits will be retained in taxa among species occupy the same habitat.

With over 1000 species distributed pantropically and an astounding phytochemical diversity, *Piper* is a formidable model genus for the study of ecological interactions.[10, 11] However, the investigation of plant-pathogen interactions in this genus have remained mostly limited to the characterization of specific antifungal across individual species. A recent review showed the prevalence of amides and phenylpropanoids as the most commonly found antifungal compounds in the genus,[12] which could indicate the conservation of specific modes of chemical defenses. In this study, we adopted the ^{1}H NMR network approach described in Chapter 1 to investigate the biological activity of 30 species of *Piper* against model organisms, including the unicellular fungus *Saccharomyces cerevisae.* We identified specific molecular features associated with yeast inhibition in two species and utilized those targets to guide the process of compound isolation and characterization, thus arriving at three novel compounds, two of which presented antifungal activity. Applied to pathogenic fungi, this methodology could be not only a valuable means to study conserved phytochemical defenses across taxa, but also a useful tool for establishing structure-activity relationships (SAR) of potential targets and optimizing the pharmacological properties of specialized metabolites.

2.2 Methods

Sample preparation: 30 different species of *Piper* (Table 2.1) were collected at La Selva Biological Station in Costa Rica, Heredia Province (10°25′ N, 84°00′ W, 50 m). The most recently expanded leaves were collected from multiple individuals and pooled for each species. The samples were then dried, extracted and prepared for ^{1}H NMR according to the methodology described in Chapter 1.

Bioassays: The crude methanolic extracts were assayed in three different panels: 1) a bacterial growth assay using *Escherichia coli* (Enterobacteriaceae); 2) a yeast growth assay with *Saccaromyces cereviseae* (Saccharomycetaceae); and 3) a root growth assay using the plant *Arabidopsis thaliana* (Brassicaceae).

E. coli strain DH5α cells were grown on Luria Broth (LB) solid media and incubated at 37 °C for 16 hr. Single colonies were then used to inoculate 10 mL LB liquid cultures, which were incubated at 37 °C for 16 h with shaking. Aliquots of the saturated cultures were diluted 100-fold in LB liquid medium. Plant extracts were dissolved in methanol at a standardized concentration of 80 mg/mL, and test extracts were added to the diluted *E. coli* cultures at a concentration of 80 µg/mL. Two hundred microliter samples were arrayed into individual wells of a sterile 96-well plate and sealed with clear adhesive film. The plate was placed in a SpectraMax M2e 96-well plate reader (Molecular Devices, Sunnyvale, CA) equilibrated at 37° C. The absorbance at 600 nm

(OD600) was measured every 5 min for 12 h with an initial shake time of 5 s and 3 s shake prior to each reading.

Table 2.1. *Piper* species included in the study and their clades.

Clade	Species
Schilleria	*P. cabagranum*
	P. scheideanum
Radula	*P. xanthostachyum*
	P. silvivagum
	P. umbricola
	P. concepcionis
	P. biolleyi
	P. aduncum
	P. sanctifecilis
	P. glabrescens
	P. friedrichsthalii
	P. urostachyum
	P. colonense
Pothomorphe	*P. peltatum*
	P. auritum
Peltobyron	*P. nudifolium*
	P. phytolaccifolium
	P. trigonum
	P. augustum
	P. garagaranum
Macrostachys	*P. melanocladum*
	P. biseriatum
	P. holdridgeanum
	P. cenocladum
	P. imperiale
	P. arboreum
	P. peracuminatum
	P. pseudobumbratum
Enckea	*P. reticulatum*
-	*P. perbrevicaule*

S. cerevisiae growth curves were measured in a similar manner. *S. cerevisiae* S288c cells were plated on YPD media (2% [w/v] peptone, 1% [w/v] yeast extract, 2% [w/v] glucose) and incubated for 2 days at 30 °C. A single colony was used to inoculate a 10 mL YPD culture, which was incubated at 30 °C for 18 h with shaking. Saturated cultures were diluted 100-fold into liquid YPD, and extracts were diluted into these cultures as described above. Samples were arrayed into 96-well plated and sealed with adhesive film. A sterile needle was used to puncture a small hole in the adhesive film above each well to prevent gas buildup. The resulting plate was assayed in a SpectraMax M2e 96-well plate reader as described above with OD600 readings taken at 5 min intervals for 18 h with 30 s of shaking before each reading. Following the identification of inhibitory targets, we repeated the assays with the purified compounds, initially at a concentration of 100 μM, and then at serial dilutions in order to calculate the half maximal inhibitory concentration (IC_{50}).

A. thaliana Col-0 seeds were surface sterilized with seed cleaning solution (3% [v/v] sodium hypochlorite, 0.1% [w/v] sodium dodecylsulfate) for 20 min at 25 °C. The seed cleaning solution was removed, and seeds were washed five times in sterile water. Seeds were resuspended in sterile water, incubated at 4 °C for 48 h, then plated on MS-agar media (1/2X Murashige and Skoog salts, MES-KOH pH 5.7, 1% [w/v] sucrose, 1% [w/v] phytoagar) with or without the addition of *Piper* extracts at a final concentration of 80 μg/mL. Plants were grown vertically in a growth chamber at 22 °C with constant light for seven days. The roots of each seedling were straightened, and the resulting plants were imaged on a flatbed scanner. Root lengths were measured using ImageJ (imageJ.nih.gov/ij/).

NMR analyses: Plant extracts were analyzed under the same experimental parameters described in Chapter 1. For chromatographic fractions, sample were resuspended in 300-500 μL of methanol-*d*$_4$, depending on the amount of sample available, and subjected to the same analytical method as the crude extracts. For the characterization of purified compounds, we also collected ^{13}C (nt=15000), ^{1}H{^{1}H} gCOSY (nt=4x128), ^{1}H{^{13}C} gHSQC (nt=4x256) and ^{1}H{^{13}C} gHMBC (nt=8x256), using an experimentally determined optimal pulse width for each compound. In cases where the solvent residual peaks compromised the characterization of the compounds, samples were also prepared and analyzed using deuterated acetonitrile to resolve the assignments.

Spectra from plant samples were processed using MNova according to the methodology described in Chapter 1. In addition, we also evaluated the results obtained through a newly incorporated method for data binning from that software, which automatically recognizes peak regions and attributes zero-values to regions of baseline spectral signal. Spectra acquired for the isolated compounds were processed to improve signal resolution according to established standard protocols.

Statistical analyses: To facilitate the characterization of bioactive molecular targets from the crude extracts, we seeded the network analysis with some of the spectra collected from the prepared mixtures (Chapter 1). After the modules were calculated for the combined set of mixtures and extracts, we calculated the Pearson correlations between module eigenvalues and bioassay data. We then processed a parallel analysis to assign module correlations with compound concentrations from the mixtures. By comparing these results, we were then able to estimate the structural elements present in the bioactive targets. Moreover, by referring to the modules that demonstrated significant correlation with bioactivity, we identified diagnostic peaks in the ^{1}H NMR spectrum that were used as a guide for the isolation of the bioactive compounds.

Compound isolation: For the isolation of target compounds, we sequentially and extensively extracted 6 g of leaf material with *n*-hexane, acetone and methanol at room temperature under mechanical agitation. For extracts that displayed the target peaks upon ^{1}H NMR analysis, 200 mg of the extract were pre-fractionated using preparative RP-LPLC (10 g, FC-C18 60 μm, 2.5 cm x 8 cm), eluted with acetone/H_2O at discrete increments of 10% acetone from the equilibration mixture (e.g., acetone extracts were eluted at 50%–100% Acetone). Two samples of 15 mL were collected for each eluent mixture, then dried under reduced pressure, prepared and analyzed by ^{1}H NMR. When additional steps of purification were necessary, the pre-purified fractions were submitted to RP-LPLC (5 g, FC-C18 60 μm, 1.5 cm x 5 cm) eluting with a continuum gradient of acetone/H_2O.

2.3 Results

Bioassays: Moderate to high biological activity was verified for the collection of extracts across the three assay panels analyzed. *E. coli* growth inhibition was modest for a few species, while *A. thaliana* root growth was affected in variable extents by all *Piper* extracts. Conversely, *S. cereviseae* assays resulted in the most distinct results, with two species–*P. holdrigeanum* and *P. peracuminatum*–presenting nearly total growth inhibition. *P. holdrigeanum* was also the species with the highest inhibitory activity in all panels.

Table 2.2. Inhibitory activities of various *Piper* extracts represented as % growth from control.

	E. coli	*A. thaliana*	*S. cereviseae*
P. aduncum	0.0	62.4	3.7
P. arboreum	0.0	59.3	10.1
P. augustum	6.5	25.6	2.1
P. auritum	9.6	19.6	10.3
P. biolleyi	2.0	13.1	0.0
P. biseriatum	15.5	16.2	1.4
P. cabagranum	1.3	27.6	8.8
P. cenocladum	16.7	19.5	6.9
P. colonense	7.8	30.6	28.9
P. concepcionis	5.0	32.4	18.8
P. friedrichsthalii	0.0	16.7	22.4
P. garagaranum	0.0	48.9	12.1
P. glabrescens	0.5	27.3	19.9
P. holdridgeanum	29.1	82.9	93.7
P. imperiale	0.0	27.5	8.3
P. melanocladum	0.0	29.1	4.4
P. nudifolium	19.6	10.4	17.3
P. peltatum	12.6	21.8	7.7
P. peracuminatum	7.4	24.9	98.7
P. perbrevicaule	1.3	24.3	0.0
P. phytolaccifolium	0.0	14.2	22.6
P. pseudobumbratum	1.8	29.4	63.2
P. reticulatum	0.0	72.8	5.8
P. sanctifecilis	17.1	30.1	15.4
P. scheideanum	0.0	22.1	19.4
P. silvivagum	2.2	15.8	5.0
P. trigonum	0.0	15.4	14.8
P. umbricola	0.0	28.0	3.5
P. urostachyum	5.5	19.7	15.8
P. xanthostachyum	11.1	30.9	12.0
Average	5.8 ± 7.5	29.9 ± 17.8	18.4 ± 24.3

Network analysis: The generated network (β = 11) resulted in 22 modules, with an average of 8 (± 4) nodes/module, excluding the turquoise module which was outstandingly large. Considering only positive eigenvalues, module representation differed between the seeded mixtures and the plant extracts, reflecting the differences in complexity of these two sources of data. The spectra from the mixtures were represented by a smaller number of modules than the extracts (6.4 ± 2.2 *versus* 10.1 ± 2.6, t-value = -6.96, $p<0.01$), but the average eigenvalue for each spectrum was larger for the mixtures (0.12 ± 0.05 *versus* 0.08 ± 0.03, t-value = 4.12, $p<0.01$). Mean eigenvalue variance was also larger for the mixtures set (0.13 ± 0.05 *versus* 0.009 ± 0.005, t-value = 2.92, $p<0.01$).

Significant correlations ($p < 0.05$) between module eigenvalues and biological activity were verified for all the bioactivity assays, and they reflected strong module associations with plant species that displayed enhanced inhibitory activity (Figure 2.1). For instance, the module *lightgreen* had the strongest association with fungal inhibition, and its highest eigenvalue was calculated for *P. peracuminatum* ($\varepsilon = 0.56$), the most potent extract against *S. cereviseae*. The second highest eigenvalue for this module was verified with *P. pseudobumbratum* ($\varepsilon = 0.11$), which also showed distinguishable yeast growth inhibition, although to moderate extent. Interestingly, the second most potent extract, from *P. holdrigeanum*, was not associated with this module, suggesting different mechanisms of inhibition for that species.

By combining mixtures and extracts into the same network analysis, we were able to directly identify the structural features described by each module. Thus, module *lightgreen* was strongly associated with prenylated benzoic acid, and according to its chemical shifts (δ_H 5.33, δ_H 1.82–1.74), it described the prenyl portion of that molecule. Those peaks are present in both *P. peracuminatum* and *P. pseudobumbratum*, where they represented the most dominant signals in the spectrum, suggesting that the corresponding target compounds were the major components of each extract. Given the limited degree of structural overlap revealed by *lightgreen*, we turned to other modules correlated with these *Piper* species to gain more information about the bioactive targets. The second most important module for *P. peracuminatum* was *royalblue* (ε = 0.18), which was associated with pipleroxide and alkene amide in the mixtures through the β-carbonyl benzylic proton resonance (δ_H 2.86). *P. pseudobumbratum* was most strongly associated with the module *lightyellow* (ε = 0.41), which contained chemical shifts representative of the vinylic methyl and allylic methylene resonances from nerolidol and carene in the mixtures (δ_H 1.98–1.94, δ_H 1.58–1.54). With the convergence of these results, we predicted that these two species of *Piper* contained structurally similar prenylated aryls as their most abundant and bioactive compounds (Figure 2.2). *P. pseudobumbratum* presents the least functionalized of these compounds, which is corroborated by its similarity with the basic terpenoid scaffold of nerolidol, while *P. peracuminatum*'s compound has an oxidized moiety similar to that observed in the propenone portion of pipleroxide. We considered these characteristic signals to perform an NMR-guided fractionation of the crude extracts and isolate the active compounds.

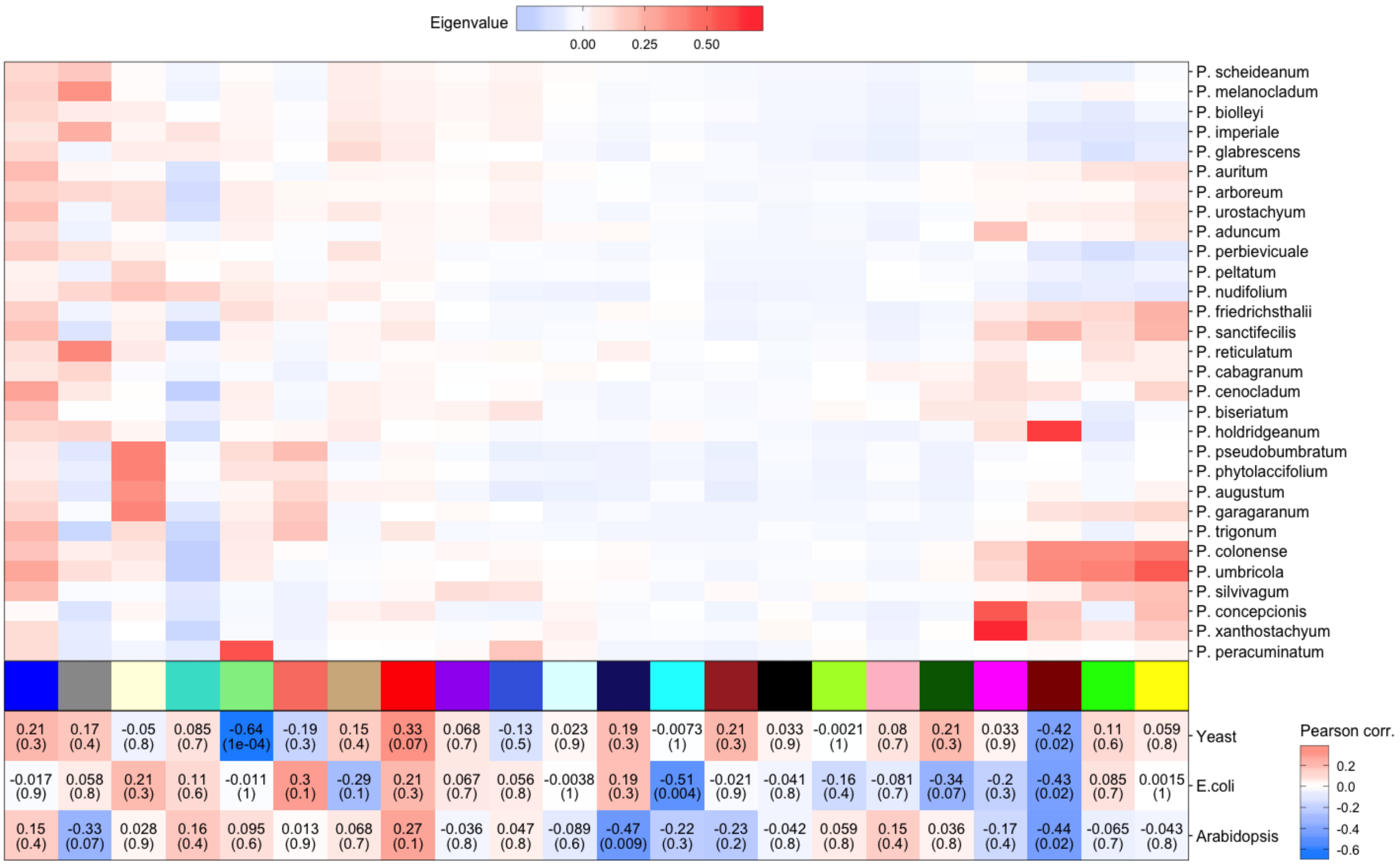


Figure 2.1. Network analysis from *Piper* extracts. The top panel displays module eigenvalues for each species and the bottom panel indicates module correlation to measurements of inhibition across the assay panels. Modules are represented by colors in the middle panel.

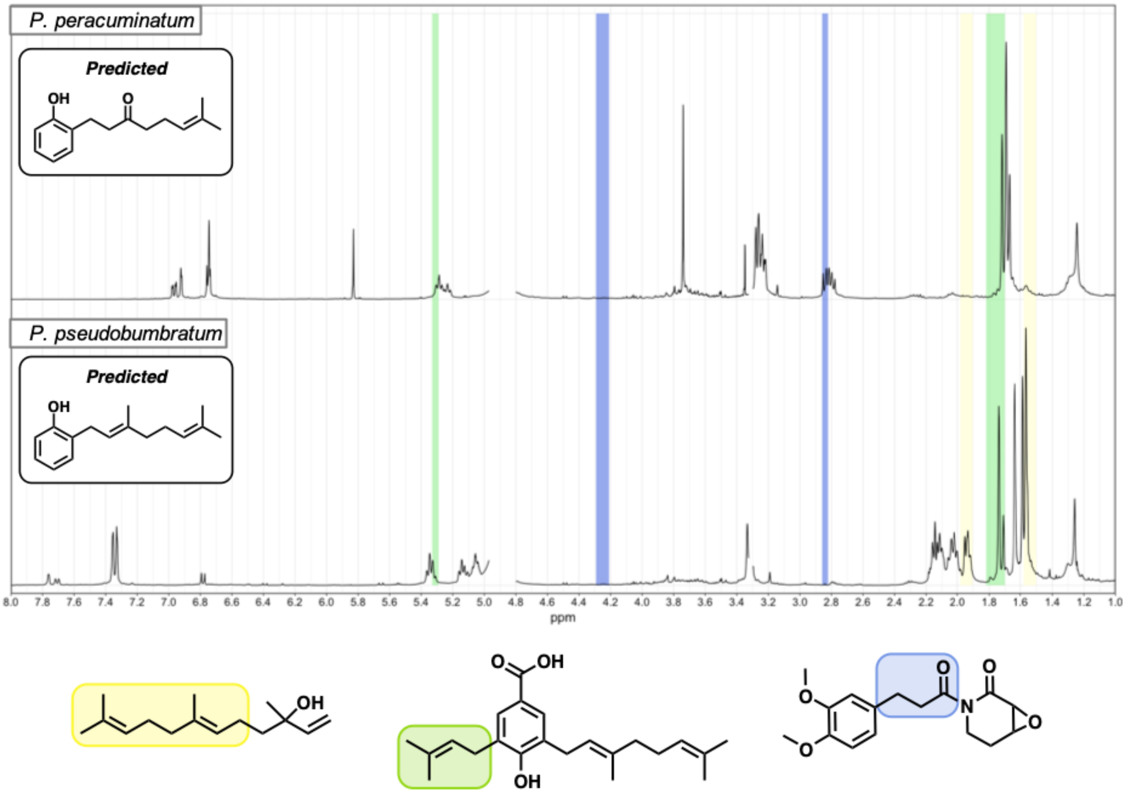
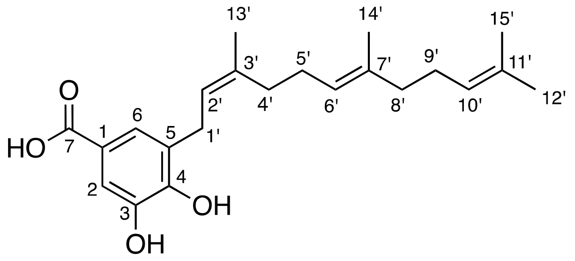


Figure 2.2. Chemical features associated with antifungal activity. In the top panel, chemical shifts for modules *lightgreen*, *lightyellow* and *royalblue* are identified across the spectra obtained from two species of *Piper* by the respective colored bars. These peaks identify structural features present in molecules from the mixtures (bottom, shaded boxes of equivalent color), and helped estimate the structure of the target compounds in each species.

Compound isolation: Targeting the isolation of the bioactive compounds from *P. peracuminatum* and *P. pseudobumbratum*, we identified the crude acetone extracts as the ones containing the signature peaks for each species. Pre-fractionation of the *P. pseudobumbratum* extract yielded the target compound at high concentration from the second fraction eluted with 50% acetone:H_2O, which following purification, resulted in 35 mg of compound ***1*** (Table 2.3). Analysis of the ^{1}H NMR data suggested the presence a 1,2,3,5-tetrasubstitued aromatic ring of a farnesyl moiety, characterized by a sequence of three vinylic resonances (δ_H 5.0–5.4). From the ^{13}C NMR data we concluded that one of the ring substituents was a carboxylic acid (δ_C 170.6) and that the other two were hydroxyl groups (aromatic δ_C 149.4 and δ_C 145.4). Finally, we concluded that since the carboxylic carbon C-7 and one of the hydroxylated carbons (C-4) showed ^{1}H{^{13}C} HMBC correlations with both the aromatic protons, they should stand symmetrically in reference to H-2 and H-6. Placing the protons in the *ortho* position of the carbonyl substituent equilibrated the shielding/deshielding effects of the hydroxyl and carbonyl substituents and was coherent with the aromatic resonances observed at δ_H 7.34 and δ_H 7.48. NOESY data confirmed the configuration of the alkene groups, and the final structure was determined as 3,4-dihydroxy-5-(*Z*,*E*-farnesyl)benzoic acid, which is a novel isomeric form of the equivalent *E*,*Z* compound isolated from *P. auritum*.[13] HRESIMS analysis on negative mode determined a mass of m/z 357.2009 $[M - H]^-$, further supporting this assignment.

Table 2.3. NMR assignments for the isolated compound 1



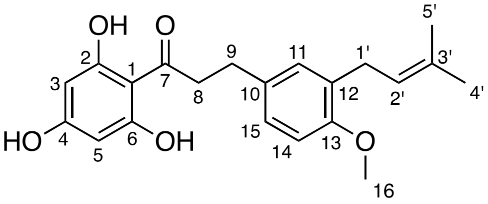
1

Position	δ_C, type	δ_H	^{1}H{^{13}C} HMBC	^{1}H{^{1}H}NOESY
1	122.4, C			
2	115.1, CH	7.32 d (2.0 Hz)	3, 4, 6, 7	
3	145.4, C			
4	149.4, C			
5	137.3, C			
6	124.1, CH	7.35 dt (2.0, 0.5 Hz)	2, 4, 7	1', 2'
7	170.7, C			
1'	28.8, CH_2	3.33 br d (7.1 Hz)	4, 5, 2', 3'	6, 4'
2'	124.0, CH	5.36 br t (7.4 Hz)	1', 3', 4', 13'	6, 13'
3'	129.0, C			
4'	32.9, CH_2	2.21–2.15	5, 2', 5', 6', 13'	1', 6', 13'
5'	27.6, CH_2	2.16–2.09	4', 6', 7'	14'
6'	125.3, CH	5.16 br t (7.4 Hz)	5', 8', 14'	4', 8'
7'	136.2, C			
8'	40.8, CH_2	1.98–1.92	5', 6', 7', 9', 10', 14'	6', 14'
9'	27.8, CH_2	2.08–2.01	6', 8', 10', 11'	14'
10'	125.5, CH	5.07 m	12', 15' (weak)	8', 12'
11'	132.0, C			
12'	25.9, CH_3	1.65 br s	10', 11', 12'	10'
13'	23.8, CH_3	1.75 br q (1.2 Hz)	5, 2', 3', 4'	2', 4'
14'	16.1, CH_3	1.60 br s	6', 7', 8'	5'
15'	17.7, CH_3	1.58 br s	10', 11', 15'	

Pre-fractionation of the acetone extract from *P. peracuminatum* led to the recovery of the target peaks in both fractions eluted with 50% acetone:H_2O and in the first one eluted with 60%. These fractions were pooled for purification and, to our surprise, the resulting collections yielded two novel compounds–**2** (20 mg) and **3** (16 mg)–with proximal eluting times that contained the target resonances for antifungal activity. Both presented resonances characteristic of prenyl groups (identified by module *lightgreen*) across the ^{1}H-NMR spectrum, but they were consistently and slightly shifted upfield for **3** (Table 2.5) compared to the same peaks in **2** (Table 2.4). The compounds also had the signature peak from module *royalblue* (δ_H 2.86, H-9), which was confirmed to be the part of a propenone moiety based on its $^{1}H\{^{1}H\}$ COSY correlations with the methylene resonance at ~δ_H 3.25 (H-8) and on the $^{1}H\{^{13}C\}$ HMBC peaks of both methylenes with the carbonyl resonance at δ_C 206 (C-7). The peak from H-9 was present at an integration ratio of 1:1 with the vinylic resonance (~δ_H 3.25, H-2') in compound **3**, which led us to the conclusion that this compound possessed two equivalent prenyl groups.

The most distinctive proton peak patterns for the two compounds were verified in the aromatic region, where compound **2** displayed three diagnostic resonances for a 1,2,4-trisubstituted ring, while **3** contained a unique singlet indicative of a symmetric 1,2,3,4-tetrasubstituted ring. We confirmed from the $^{1}H\{^{13}C\}$ HMBC peaks between these aromatic protons and the carbons C-9 and C-1' that the prenyl groups are implicated in the ring symmetry of **3**, and that in both compounds the propenone moiety is attached to the ring through the β-carbonyl carbon, at the *meta* position with reference to the prenyl groups. The remaining substituent was identified as a methoxyl group in compound **2** and a hydroxyl group in **3** based upon $^{1}H\{^{13}C\}$ HMBC signals between a hydroxylated aromatic carbon (C-13) and the respective aromatic protons. Finally, we identified the singlet at δ_H 5.86 as being part of a symmetric trihydroxylated aromatic system, which was connected to the propenone linker through the carbonyl carbon ($^1H\{^{13}C\}$ HMBC, C-7 and H-3/5), thus concluding both structures. HRESIMS analysis on negative mode corroborated these assignments with masses of m/z 355.1560 $[M - H]^-$ and m/z 409.2030 $[M - H]^-$ for **2** and **3**, respectively.

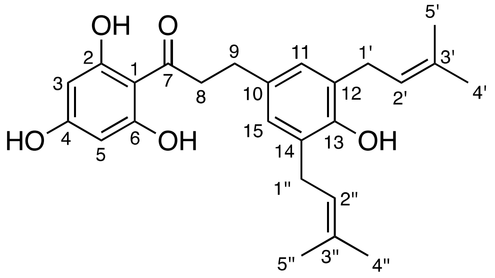
Table 2.4. NMR assignments for the isolated compound 2



2

Position	δ_C**, type**	δ_H	^{1}**H{**^{13}**C} HMBC**
1	105.3, C		
2/6	165.2, C		
3/5	95.9, CH	5.86 s	1, 2/6, 4, 7
4	164.9, C		
7	205.9, C		
8	46.8, CH_2	3.27–3.24	7, 9, 10
9	30.5, CH_2	2.84 dd (8.2, 6.7 Hz)	7, 8, 10, 11, 15
10	134.7, C		
11	130.4, CH	6.98 d (2.0 Hz)	9, 13, 15, 1'
12	130.6, C		
13	156.6, C		
14	111.5, CH	6.82 d (8.3 Hz)	10, 11, 12, 13, 15
15	127.6, CH	7.01 dd (8.2, 2.3 Hz)	11, 13, 14
16	56.1, CH_3	3.77 s	
1'	29.3, CH_2	3.24 br d (7.4 Hz)	11, 12, 13, 2', 3'
2'	123.7, CH	5.25 m	5' (weak)
3'	132.9, C		
4'	25.9, CH_3	1.70 br s	2', 3', 5'
5'	17.8, CH_3	1.70 br s	2', 3', 4'

Table 2.5. NMR assignments for the isolated compound 3



3

Position	**δ_C, type**	**δ_H**	**^{1}H{^{13}C} HMBC**
1	105.3, C		
2/6	165.1, C		
3/5	95.8, CH	5.86 s	1, 2/6, 4, 7
4	164.7, C		
7	206.0, C		
8	46.7, CH$_2$	3.21–3.27	7, 9, 10
9	30.6, CH$_2$	2.79 t (7.6 Hz)	7, 8, 10, 11/15
10	134.4, C		
11/15	128.2, CH	6.79 s	9, 13, 1'/1"
12/14	128.7, C		
13	151.0, C		
1'/1"	29.7, CH$_2$	3.25 br d (7.3 Hz)	11/15, 13, 12/14, 2'/2", 3'/3", 5'/5"
2'/2"	125.5, CH	5.26 m	1'/1", 4'/4", 5'/5"
3'/3"	132.0, C		
4'/4"	23.8, CH$_3$	1.72 br s	2'/2", 3'/3", 5'/5"
5'/5"	17.7, CH$_3$	1.71 br s	2'/2", 3'/3", 4'/4"

2.4 Discussion

The direct network analysis on the combined ^{1}H NMR data from prepared mixtures and plant extracts exhibited the same efficiency at recognizing shared structural features as did the analyses described in Chapter 1. Overall module specificity was different for spectra obtained from seeded mixtures and crude extracts as a result of sample complexity. Thus, modules were more evenly represented (lower eigenvalues and higher module count per spectrum) in the chemically complex crude extracts, while the seeded mixtures, with a significantly smaller number of compounds, were more strongly associated with specific modules. These differences were determinant in recognizing chemical patterns across the data set, as the mixtures served as a common library of well-defined chemical signatures which were then quantitatively affiliated (module eigenvalues) with each of the individual *Piper* extracts. This feature was fundamental for establishing a direct correlation between measurements of biological activity and molecular composition of the crude extracts, ultimately leading to the guided and accurate annotation of the biologically active targets. With this metabolomic evaluation, compounds were purified targeting the resonances that were identified as relevant to bioactivity. Isolation led to three compounds, two of which retained the inhibitory activity verified from the original extracts (Table 2.6), thus validating the effectiveness of this method in the early-stage identification of bioactive molecules from complex natural mixtures.

Table 2.6. Inhibitory activity of isolated compounds against *S. cereviseae*

Compound	*IC-50 (μM)*
1	46.8 (± 19.2)
2	3.02 (± 1.05)
3	**

** not significantly different from control

Several *Piper* extracts exhibited moderate to high inhibitory activity in the yeast and plant root assays, but we prioritized *P. peracuminatum* and *P. pseudobumbratum* because the differentiated antifungal activity of the extracts from these two species offered a formidable insight into how the network analysis can be applied to establishing SAR for target samples. Extracts of both plant species were associated with the module that showed the strongest correlation (r = 0.64) with *S. cereviseae* inhibition, and which represented a prenyl moiety, suggesting that this recurring structural feature was determining for compound activity in the investigated extracts. This result resonates with a collection of studies that encountered prenylated benzoic acid derivatives with antifungal activity in several species of *Piper*.[14] It is noteworthy, however, that although this particular structural feature was correlated with increased inhibitory activities of compounds **1** and **2**, the presence of a second prenyl unit in compound **3** resulted in significantly reduced potency, suggesting that the spacial arrangement of this functional group is also important.

The enhanced inhibitory effect of *P. peracuminatum* was associated with another module that was not statistically significant, but that still had a correlation of 0.13 with yeast inhibition. This module represented part of a dihydrochalcone motif, which was also present in antifungal compounds isolated from *P. mollicomum*[15] and *P. aduncum*[16], and that can be postulated as an important element for yeast inhibition. Conversely, the module that was most strongly associated with *P. pseudobumbratum* was only marginally correlated with inhibitory activity as it represented structural features of an elongated farnesyl group that was not as important. We predict that the evaluation of a more complete library of *Piper* extracts using this methodology may further elucidate the structural effect of elongation and functionalization of the prenyl chain in modulating the antifungal activities of these compounds. Complementarily, the application of this approach with yeast deletion collections could provide a powerful method to identify the specific mechanisms of action of a compound.[17] Gene knockouts that result in altered drug resistance may offer direct cues to structure-cellular target affiliations, and thus help to discern whether varying levels of inhibition result from differentiated compound-specificity towards a common target, or if these compounds target distinct cellular processes. The refined structural information obtained from the chemical modules can then be evaluated against susceptible/resistant phenotypes to highlight the role of specific molecule motifs in dictating biological activity.

Finally, it is noteworthy that for the two *Piper* species investigated, the active compounds were also the major components of the extract, so the distinction of the target structures was facilitated by a more simplified spectrum. Less informative results may be expected in cases where the crude extract is far more convoluted or when a minor compound is responsible for the bioactivity. We suspect that this scenario is applicable to *P. holdrigeanum*, where the analysis did not highlight specific peaks in the ^{1}H NMR spectrum despite a potent inhibitory activity of the extract across all assays. The only module associating this species with bioactivity accounted for a minor peak in the spectrum at δ_H 8.61-8.53, offering limited insight into potential structural features. However, attempts to isolate the active compound led to the disappearance of the target peak and the consequent loss of inhibitory activity. The investigation of this phenomenon goes beyond the scope of this work, but it demonstrates how even in a more complex scenario the network analysis can distinguish the spectral signals associated with modes of bioactivity.

2.5 Conclusions and future directions

Plant-pathogen interactions have shaped the phytochemical landscape for as early as the transition from aquatic to terrestrial life forms. The mechanisms of defense identified in modern plants offer a snapshot of this convoluted story of chemical character selection, and by probing plant-bacteria or plant-fungi interactions through diverse assays one can discriminate plant taxa with enrichment for phytochemical defenses against such pathogens. Complementarily, NMR-based network analysis allows the deconvolution of phytochemical profiles into quantifiable chemical characters that can be mapped directly into bioassay data. The immediate product of this process is the identification of diagnostic peaks in the spectrum that can be utilized to monitor the isolation of bioactive compounds, thus replacing the recursive assay of chromatographic fractions. Moreover, with a suitable library of annotated spectra from pure compounds that can be analyzed in conjunction with the extracts, one can dereplicate known compounds and estimate the identity of novel molecules. However, the utility of this methodology is not limited to optimizing the process of compound isolation and identification. For example, in combination with inhibition assays against knockout libraries of model organisms, such as *S. cereviseae*, the analysis of chemical modules can support the optimization of structural features for enhancing compound activity against specific cellular targets. Additionally, activity screening against other species of fungi might reveal additional inhibitory compounds and provide evidence for the implication of structural features in compound selectivity towards different strains. The information contained by the modules can ultimately help assessing the taxonomical recurrence of molecular motifs associated with specific modes of biological activity, as demonstrated with the genus *Piper*, and help retracing the evolutionary history of chemical defenses in plants and other organisms.

2.6 References

1. Madden, L. V.; Wheelis, M., The threat of plant pathogens as weapons against U.S. crops. *Annu. Rev. Phytopathol.* **2003,** *41*, 155-176.

2. Yoshida, K.; Schuenemann Verena, J.; Cano Liliana, M.; Pais, M.; Mishra, B.; Sharma, R.; Lanz, C.; Martin Frank, N.; Kamoun, S.; Krause, J.; Thines, M.; Weigel, D.; Burbano Hernan, A., The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *Elife* **2013,** *2*, e00731.

3. Větrovský, T.; Kohout, P.; Kopecký, M.; Machac, A.; Man, M.; Bahnmann, B. D.; Brabcová, V.; Choi, J.; Meszárošová, L.; Human, Z. R.; Lepinay, C.; Lladó, S.; López-Mondéjar, R.; Martinović, T.; Mašínová, T.; Morais, D.; Navrátilová, D.; Odriozola, I.; Štursová, M.; Švec, K.; Tláskal, V.; Urbanová, M.; Wan, J.; Žifčáková, L.; Howe, A.; Ladau, J.; Peay, K. G.; Storch, D.; Wild, J.; Baldrian, P., A meta-analysis of global fungal distribution reveals climate-driven patterns. **2019,** *10* (1), 5142.

4. Shi, L.; Dossa, G. G. O.; Paudel, E.; Zang, H.; Xu, J.; Harrison, R. D., Changes in fungal communities across a forest disturbance gradient. *Appl. Environ. Microbiol.* **2019,** *85* (12), e00080.

5. Velásquez, A. C.; Castroverde, C. D. M.; He, S. Y., Plant-Pathogen Warfare under Changing Climate Conditions. *Current biology : CB* **2018,** *28* (10), R619-R634.

6. Heath, M. C., Evolution of resistance to fungal parasitism in natural ecosystems. *New Phytologist* **1991,** *119* (3), 331-343.

7. Maor, R.; Shirasu, K., The arms race continues: battle strategies between plants and fungal pathogens. *Curr. Opin. Microbiol.* **2005,** *8* (4), 399-404.

8. Martin-Hernandez, A. M.; Dufresne, M.; Hugouvieux, V.; Melton, R.; Osbourn, A., Effects of targeted replacement of the tomatinase gene on the interaction of Septoria lycopersici with tomato plants. *Mol. Plant-Microbe Interact.* **2000,** *13* (12), 1301-1311.

9. Huntley, N.; Inouye, R., Diseases and Plant Population Biology. Jeremy J. Burdon. *The Quarterly Review of Biology* **1988,** *63* (2), 241-242.

10. Dyer, L. A.; Palmer, A. D. N., *Piper : a model genus for studies of phytochemistry, ecology, and evolution*. Kluwer Academic/Plenum Publishers: New York, 2004; p xiii, 214 pages.

11. Jaramillo, M. A.; Manos, P. S., Phylogeny and patterns of floral diversity in the genus Piper (Piperaceae). *Am J Bot* **2001,** *88* (4), 706-16.

12. Xu, W.-H.; Li, X.-C., Antifungal Compounds from Piper Species. *Current bioactive compounds* **2011,** *7* (4), 10.2174/157340711798375822.

13. Ampofo, S. A.; Roussis, V.; Wiemer, D. F., New prenylated phenolics from Piper auritum. *Phytochemistry* **1987,** *26* (8), 2367-70.

14. Xu, W.-H.; Li, X.-C., Antifungal compounds from Piper species. *Curr. Bioact. Compd.* **2011,** *7* (4), 262-267.

15. Lago, J. H. G.; Young, M. C. M.; Reigada, J. B.; Soares, M. G.; Roesler, B. P.; Kato, M. J., Antifungal derivatives from Piper mollicomum and P. lhotzkyanum (Piperaceae). **2007,** *30*, 1222-1224.

16. Orjala, J.; Wright, A. D.; Erdelmeier, C. A. J.; Sticher, O.; Rali, T., New monoterpene-substituted dihydrochalcones from Piper aduncum. *Helv. Chim. Acta* **1993,** *76* (4), 1481-8.

17. Hoon, S.; St.Onge, R. P.; Giaever, G.; Nislow, C., Yeast chemical genomics and drug discovery: an update. *Trends in Pharmacological Sciences* **2008,** *29* (10), 499-504.

Chapter 3: Chemical Networks Reveal Conserved Metabolite Patterns in the *Piper* clade *Schilleria*

3.1 Introduction

Beyond the realm of bioactive targets, the study of specialized metabolites is driven by the discovery of novel molecules and by the exploration of nature-driven transformations that inspire biomimetic organic synthesis,[1] of which Robinson's preparation of tropinone is a landmark.[2] However, without a guide for sample prioritization, this untargeted approach incurs into some of the same limitations encountered in the prospection of bioactive compounds, which have been discussed in chapter 1. In that regard, phylogenetic information provides a formidable reference to phytochemical studies, as it correlates to the summation of the evolutionary processes that amount to metabolite diversity.[3] It can thus help prioritizing the investigation of taxa which shows enrichment for a specific class of compounds,[4] or taxa in which high rates of speciation may indicate a hotspot for structurally diverse compounds. Complementarily, phytochemical data can be mapped into phylogenies to help reconstructing the evolution of metabolic pathways,[5] and to assess the conservation of chemical traits within taxa. Inferences from the chemical similarity between known and unexplored species can then be drawn to facilitate compounds dereplication and, supported by phylogenetic associations, assist the identification of novel molecules.

Efforts to bridge secondary metabolism and phylogeny have largely focused on the study of biosynthetic gene clusters,[6, 7] which makes use of specific sequence tags to evaluate accumulated modifications in target enzymes. Alternatively, comparisons of unspecific sequence markers provide a holistic description of species phylogeny, which can accommodate the untargeted study of complete phytochemical profiles. The caveat for this approach is the risk of conflicting phylogenetic signals,[6] thus demanding a metabolomics methodology that can reliably reproduce the degree of structural refinement present at a given taxonomic level. As an insightful example, Ernst *et al* demonstrated the use MS-similarity approaches to assess chemo-evolutionary relationships in the hyper-diverse plant genus *Euphorbia*, which ultimately led to a better understanding of the biogeographic history of this genus.[8] Following the promising results obtained in chapter 2, we project that chemo-phylogenetic comparisons utilizing ^{1}H NMR data should more accurately reveal the varying levels of chemical diversification across taxa. Notably, the refined structural resolution inferred from the ^{1}H NMR spectrum is advantageous to discriminate compound features that result from different stages of the biosynthetic process (Chapter 1, Figure 1.4), and which may highlight more complex chemo-taxonomical relationships than when considering generalized classes of metabolites. Network analysis can thus consolidate the phytochemical data into quantifiable chemotypes, which may be used to assess species similarity at various phylogenetic scales.

In this study, we revisit the plant genus *Piper* to assess the presence of conserved specialized metabolites in infrageneric groups. According to molecular phylogenetic based on ITS sequences, Neotropical *Piper* species form a segregated group from Old World species, comprising eight distinct clades.[9] *Radula* and *Macrostachys* are the most species rich (~450 and ~200 species, respectively) and also the most recently diverged clades (ca. 7-11 Ma), suggesting that current diversity in the genus is a result of recent diversification events.[10] We sampled over 90 species representing six of the Neotropical *Piper* clades, and by utilizing the ^{1}H NMR-based network approach, we evaluated the overlap between phytochemical profiles and species phylogenetic distribution. We then focused our efforts on the clade *Schilleria*, and by identifying modules that most consistently characterized species within this clade, we encountered chemical commonalities that suggested the conserved biosynthesis of lignans. These compounds are formed from the oxidative coupling of two C3-C6 phenylpropanoid units, involving the direct, or oxygen-bridged, linkage between carbons 3 or 8, thus giving rise to an incredible diversity of structural motifs.[11] Lignans are implicated in several mechanisms of plant defense,[12] which is a statement for their evolutionary relevance. Altogether, this chapter establishes a framework for the integration of chemical and phylogenetic data, supplementing an evolutionary argument for the discovery of new compounds.

3.2 Methods

Data preparation: We gathered ^{1}H NMR data from a digital library of *Piper* samples maintained by the Hitchcock Center for Chemical Ecology at UNR. Samples were collected over the last decade from a broad geographical range of the Central and Southern American tropics, including collection sites in Brazil, Panama, Ecuador, Costa Rica and French Guyana, then processed and extracted according to the general protocol described in Chapter 1. After data filtering for adequate signal resolution and sample-to-solvent signal ratio, 223 ^{1}H NMR spectra from *Piper* specimens with taxonomical identification were selected for the network analysis (Table 3.1). That included 94 species, which were assigned into clades according to published *Piper* phylogeny,[9] *Radula* being the most well represented (88 samples) and *Enckea* the least (5 samples). Compared to the natural distribution of Neotropical *Piper* species,[10] *Radula* was the only clade with a proportional number of species (35), while the remaining ones were over or underrepresented in varying extents.

The spectra were individually referenced by the residual solvent peak (methanol-d4, δ 3.31), collectively phased and baseline-corrected (polynomial fit of order 3). For spectral binning, we opted to utilize a different method that became available in the most recent versions of the processing software MNova, and that automatically recognizes peak regions before integrating the ranges. A visual comparison of the results obtained with this "peak" method and the regular "average sum" approach showed that the former is more efficient at distinguish peaks from baseline signals or undefined regions with peak overlap (Figure 3.1). For example, the residual solvent peak region at ~δ 4.85 is considerably reduced with peak-binning because the areas flanking that peak are not considered for integration, unless a detectable sample signal is present. These differences ultimately led to a decreased influence of the baseline on the integrated data and reduced the fraction of spectrum area that is represented by the aliphatic region (δ 0.5-2.5).

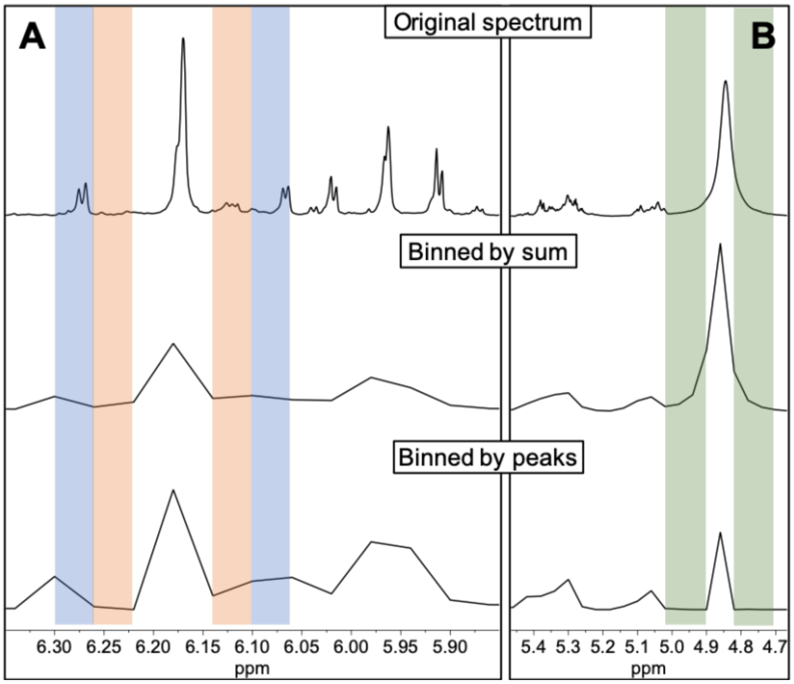


Figure 3.1. Effect of binning method on spectral data. Panel A compares method performance in differentiating peak intensities. The "binned by peaks" method provided a better ratio between areas containing defined peaks (blue rectangles) compared to baseline-intensity or undefined peaks (orange rectangles). In panel B "binned by peaks" also performed better at reducing the influence of the areas flanking the residual solvent peaks.

Statistical analyses: Network analysis was performed on the entire selection of *Piper* spectral data. Following, HCA was applied to module eigenvalues to evaluate the clustering of samples according to shared chemical featured, where distances between samples were calculated based on pairwise Pearson correlations, and clusters were defined by complete linkage. We then evaluated the overall success to distinguish clade-specific chemical features by calculating and comparing average distances within clade, outside of clade, and within species. MANOVA was performed to verify module-to-clade associations, and modules with significant interaction were further probed using Tukey's HSD post-hoc tests to verify the clade they most significantly distinguished. For method comparison, we also analysed intra- and inter-clade distances obtained directly from the filtered ^{1}H NMR data.

Table 3.1. Piper species utilized in the study.

Clade	Species	Samples	Country of Origin
Enckea	*P. laevigatum*	LAE1	Ecuador
	P. reticulatum	RET1–RET4	Ecuador and Costa Rica
Macrostachys	*P. arboreum*	ARB1–ARB2	Ecuador and Costa Rica
	P. auritifolium	ARF1–ARF4	Costa Rica
	P. bellidifolium	BEL1	Ecuador
	P. biseriatum	BIS1–BIS5	Costa Rica
	P. cenocladum	CEN1–CEN5	Costa Rica
	P. holdrigeanum	HOL1–HOL5	Costa Rica
	P. imperiale	IMP1–IMP4	Costa Rica
	P. marsupiiferum	MAR1–MAR4	Ecuador
	P. melanocladum	MEL1–MEL2	Costa Rica
	P. obliquum	OBL1	Ecuador
	P. obtusilimbum	OBT1–OBT6	Ecuador
	P. peracuminatum	PER1–PER2	Costa Rica
Peltobyron	*P. andreanum*	AND1	Ecuador
	P. caucaense	CAU1	Ecuador
	P. cirratum	CIR1–CIR2	Ecuador
	P. conejoense	CNJ1	Ecuador
	P. conispicum	CON1	Ecuador
	P. cyanophyllum	CYA1–CYA3	Ecuador and Costa Rica
	P. cyphophyllum	CYP1	Costa Rica
	P. garagaranum	GRG1	Costa Rica
	P. generalense	GEN1–GEN5	Costa Rica
	P. lanceolatum	LAN1	Ecuador
	P. macerispicum	MAC1	Ecuador
	P. maranyonense	MRN1	Ecuador
	P. musteum	MUS1–MUS6	Ecuador
	P. nudifolium	NUD1–NUD2	Costa Rica
	P. nudilimbum	NDL1	Ecuador
	P. paludosum	PAL1–PAL2	Ecuador
	P. pubinervulum	PBN1–PBN2	Ecuador
	P. prismaticum	PRI1–PRI3	Costa Rica
	P. stelipilum	STE1–STE3	Ecuador
	P. terrabanum	TER1–TER3	Panama and Costa Rica

	P. trigonum	TRG1–TRG3	Costa Rica
Pothomorphe	*P. auritum*	AUR1	Costa Rica
	P. multiplinervium	MUL1–MUL2	Costa Rica
	P. peltatum	PEL1–PEL3	Costa Rica
	P. umbellatum	UBL1–UBL2	Ecuador and Costa Rica
Radula	*P. amoenum*	AMO1	Ecuador
	P. arcteacuminatum	ARC1–ARC4	Costa Rica
	P. bioleyi	BIO1	Costa Rica
	P. baezense	BZS1–BZS2	Ecuador
	P. chiminanthifolium	CHI1	Brazil
	P. concepcionis	CNP1–CNP2	Costa Rica
	P. colonense	COL1–COL5	Costa Rica
	P. coruscan	COR1	Ecuador
	P. crassivervium	CRA1	Brazil
	P. culebranum	CUL1–CUL2	Costa Rica
	P. epigynium	EPI1–EPI2	Costa Rica
	P. friedrichsthalii	FRI1–FRI2	Costa Rica
	P. gaudichaudianum	GAU1	Brazil
	P. glabracens	GLA1–GLA3	Costa Rica
	P. goesii	GOE1	Brazil
	P. hispidum	HIS1	Ecuador
	P. "hispidum complex"	HISc1–HISc4	Costa Rica
	P. immutatum	IMM1–IMM2	Ecuador
	P. lacunosum	LAC1–LAC2	Ecuador
	P. lanceifolium	LCF1	Ecuador
	P. longicaudatum	LON1–LON2	Ecuador
	P. malacophyllum	MAL1	Brazil
	P. mosenii	MOS1	Brazil
	P. napo-pastazanum	NAP1	Ecuador
	P. otophorum	OTO1	Costa Rica
	P. pseudofuligineum	PSF1	Panama
	P. sancti-felicis	SAN1–SAN6	Costa Rica
	P. schuppii	SCH1–SCH2	Ecuador
	P. silvivagum	SIL1–SIL6	Costa Rica
	P. tabanicidum	TAB1	Costa Rica
	P. tonduzii	TON1–TON5	Costa Rica
	P. umbricola	UMB1–UMB5	Costa Rica
	P. urostachyum	URO1–URO7	Costa Rica
	P. villaloboense	VIL1	Ecuador
	P. xanthostachyum	XAN1–XAN8	Costa Rica
Schilleria	*P. aequale*	AEQ1–AEQ3	Ecuador

	P. amphioxys	AMP1	Panama
	P. asymmetricum	ASY1–ASY3	Costa Rica
	P. cabagranum	CAB1	Costa Rica
	P. paulownifolium	PAU1–PAU3	Costa Rica
	P. perlaense	PRL1–PRL2	Panama
	P. scutilimbum	SCU1	Ecuador
	P. subscutatum	SUB1–SUB3	Ecuador
	P. urophyllum	URP1–URP5	Costa Rica
Undeterminated	*P. baezanum*	BAE1	Ecuador
	P. barbatum	BAR1	Ecuador
	P. boquetense	BOQ1	Panama
	P. eriocladum Sodiro	ERI1	Ecuador
	P. japurense	JAP1	Ecuador
	P. lucigaudens	LUC1	Panama
	P. pseudobumbratum	PSB1–PSB3	Costa Rica
	P. pseudogaragaranum	PSG1–PSG2	Panama
	P. pseudovariabile	PSV1	Panama
	P. puberulescens	PUB1	Ecuador
	P. sinugaudens	SIN1–SIN2	Panama

For the investigation of the clade *Schilleria*, we first performed an HCA including the species from that clade exclusively, but utilizing the same modules obtained from the network analysis of all samples. Clusters of chemically similar samples were then evaluated, and shared chemical features were identified through the comparison of cluster-associated modules and the original spectra. Two of these modules were identified as relevant targets for compound isolation, leading to the phytochemical investigation of *P. cabagranum* and *P. paulownifolium*.

Compound isolation: Leaf samples of *P. cabagranum* and *P. paulownifolium* collected at La Selva Biological Station in Costa Rica were sequentially and exhaustively extracted with hexanes and acetone. We then subjected 200 mg of the crude extracts to pre-fractionation through RP-LPLC and subsequent purification by RP-MPLC to recover the target compounds, following the procedure described in Chapter 2. Compound characterization was supported by one- and two-dimensional NMR analytical techniques, according to the parameters also specified in Chapter 2.

3.3 Results

Network analysis of *Piper*: The collection of *Piper* spectra analyzed in this study represented a matrix of chemical data that was far more diverse than any of the analyses described in the previous chapters. As a result, a much lower threshold parameter ($\beta = 4$) was required to construct the network, which resulted in 31 modules with an average composition of 7 ± 2 nodes/module. HCA on module eigenvalues distinguished three clusters predominantly composed of samples pertaining to *Radula* and one enriched with samples of the *Peltobyron* clade (Figure 3.2, shaded sectors). The majority of the clusters, however, contained a combination of samples belonging to multiple clades, and no other clade was uniformly represented within a cluster.

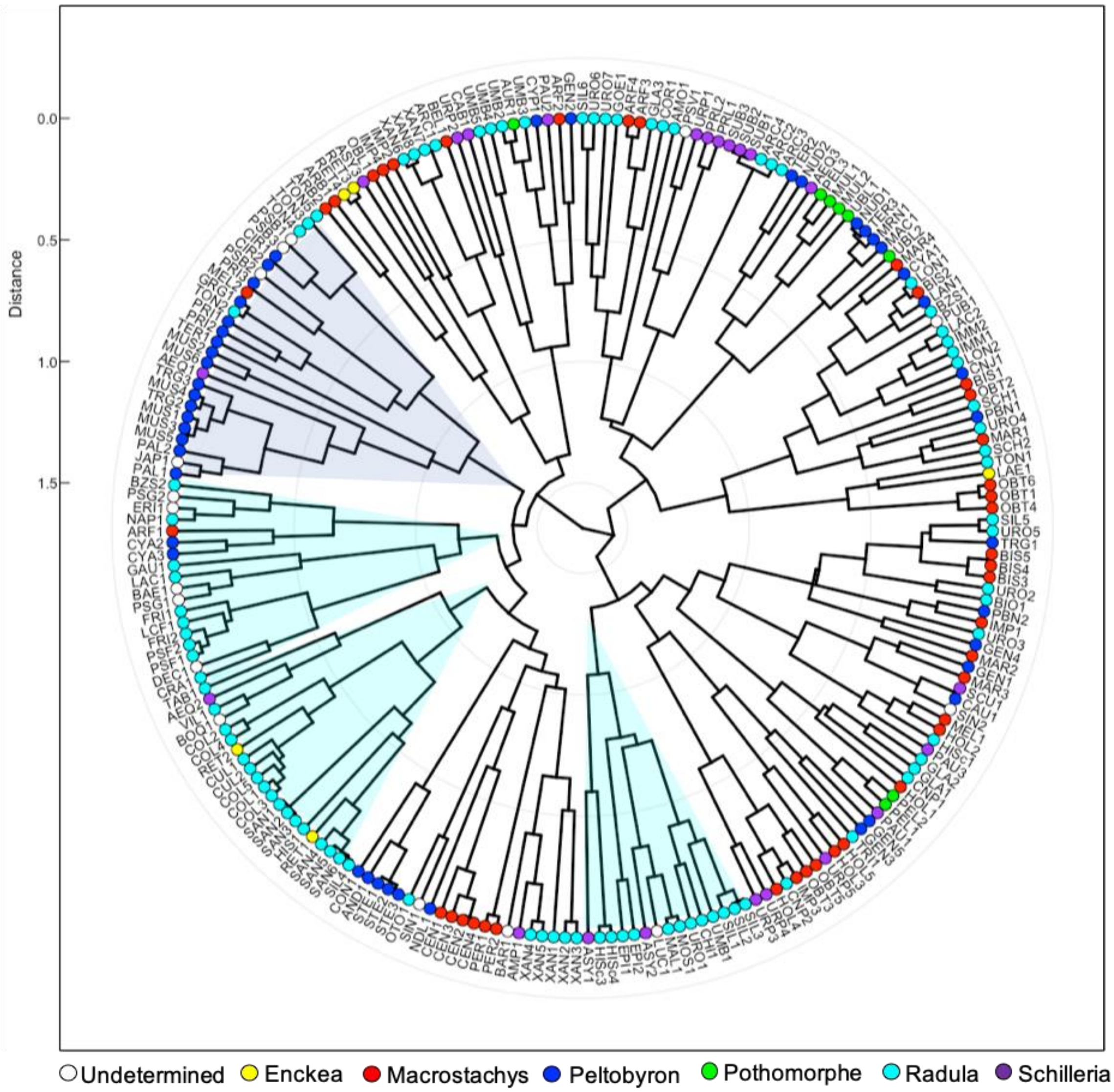


Figure 3.2. Hierarchical clustering obtained from the analysis of module eigenvalues of *Piper* samples. The colored circles represent clade assignments for each sample. Shaded sectors in the dendrogram show clusters with enrichment for the clades *Peltobyron* (blue) and *Radula* (cyan). Other clusters display a more heterogeneous clade composition.

In order to obtain a quantitative assessment of these results, we estimated sample similarity as the pairwise Pearson correlation distance of module eigenvalues. If two samples express the same general module profiles, their distance is small, and they are considered to be chemically similar. We found that distances to samples outside of the clade were consistent across all groups, and marginally higher than the distances to samples within-clade (Table 3.2). Intra-clade similarity generally decreased with species diversity, and for *Radula* (s = 35) that diversity led to a closer overlap between intra- and inter-clade distances. However, intraspecific variation in that clade was comparable with the least diverse clades *Enckea* (s = 2) and *Pothomorphe* (s = 4). Conversely, *Schilleria* (s = 9) showed the lowest degree of intraspecific similarity, despite having a higher intra-clade similarity than *Radula*. Moreover, intra-clade distances obtained directly from the ^{1}H NMR data were higher than those generated from the module eigenvalues, while inter-clade distances were conserved, demonstrated that the modules were more effective at distinguishing clade-associated features.

Table 3.2. Average sample distances calculated from module eigenvalues and ^{1}H NMR data.

	Module eigenvalues			***^{1}H NMR data***
	Intra-clade	**Inter-clade**	**Intra-species**	**Intra-clade***
Enckea	0.57 (± 0.12)	0.98 (± 0.22)	0.41 (± 0.16)	0.74 (± 0.31)
Macrostachys	0.87 (± 0.26)	1.00 (± 0.21)	0.50 (± 0.25)	0.95 (± 0.16)
Peltobyron	0.91 (± 0.29)	1.01 (± 0.22)	0.36 (± 0.12)	0.94 (± 0.20)
Pothomorphe	0.55 (± 0.27)	1.01 (± 0.21)	0.43 (± 0.37)	0.77 (± 0.22)
Radula	0.97 (± 0.26)	1.01 (± 0.20)	0.46 (± 0.17)	0.99 (± 0.15)
Schilleria	0.91 (± 0.23)	1.01 (± 0.21)	0.63 (± 0.14)	0.94 (± 0.18)

***** Inter-clade measures from ^{1}H NMR data were consistently around 1.01 with a standard deviation of approximately 0.11.

To verify the presence of clade-specific chemotypes, we tested module association with clade identity. MANOVA indicated the significance of this relationship (Wilks $\lambda = 0.09$, $p \leq 0.01$), and 9 of the 31 modules showed specific connection with one of the clades (Tukey's HSD, $p \leq 0.01$). The clade *Schilleria* was most significantly identified by the module *darkgreen*, so we subsequently performed an HCA for samples within that clade to evaluate how this module associated with different species. Samples were distinguished into three main clusters (*S1–S3*), but the *darkgreen* module was most consistently associated with *S3*, predominantly due to the species *P. subscutatum* (Figure 3.3). Visual analysis of the ^{1}H NMR spectra from SUB1 revealed the presence of few dominant resonance peaks that coincided with those represented by *darkgreen*. Based on reported phytochemical studies of *P. subscutatum*, were identified those peaks as pertaining to the lignan grandisin (Figure 3.6).[13] This compound was present at high concentrations in all *P. subscutatum* samples, but only as a trace compound in other samples that were distinguished into *S3*. Interestingly, the *darkgreen* module was also modestly associated with the species *P. cabagranum*, not due to the presence of grandisin, but for the partial overlap with a similar group of resonances in the aromatic and methoxylated regions of the spectrum. Given that *P. cabagranum* was present in another cluster (*S1*), we considered this evidence a prognostic for the presence of structurally related compounds in other *Schilleria* species.

Targeting modules that were shared by samples within the same cluster, we identified *grey60* as a relevant module that was most strongly associated with *P. cabagranum*. This module contained a group of resonance peaks that altogether indicated the presence of a C6-C3 phenylpropanoid system with a hydroxylated aromatic ring (δ_H

6.78 and δ_H 6.66). The most characteristic peak for this structural motif, a *ddt* pattern at approximately δ_H 6.00, was also identified in other species present in *S1*, but those displayed different patterns across the aromatic region (Figure 3.3). For instance, CAB1 contained two apparent pairs of 1,3-coupled protons at δ_H 6.60–7.00 (J ~ 2 Hz), URP2 showed a single peak at δ 6.49, and PAU3 contained a major pair of 1,2-coupled resonances at δ_H 6.80 and δ_H 7.25 (J ~ 8 Hz). These patterns suggested the presence of diverse lignan structures in *Schilleria*, so we targeted some of these compounds for isolation. We selected *P. cabagranum* and *P. paulownifolium* for this investigation primarily due to sample availability, but also because they represented the most dissimilar samples in the cluster *S1*, thus likely to contain the most distinct lignan motifs.

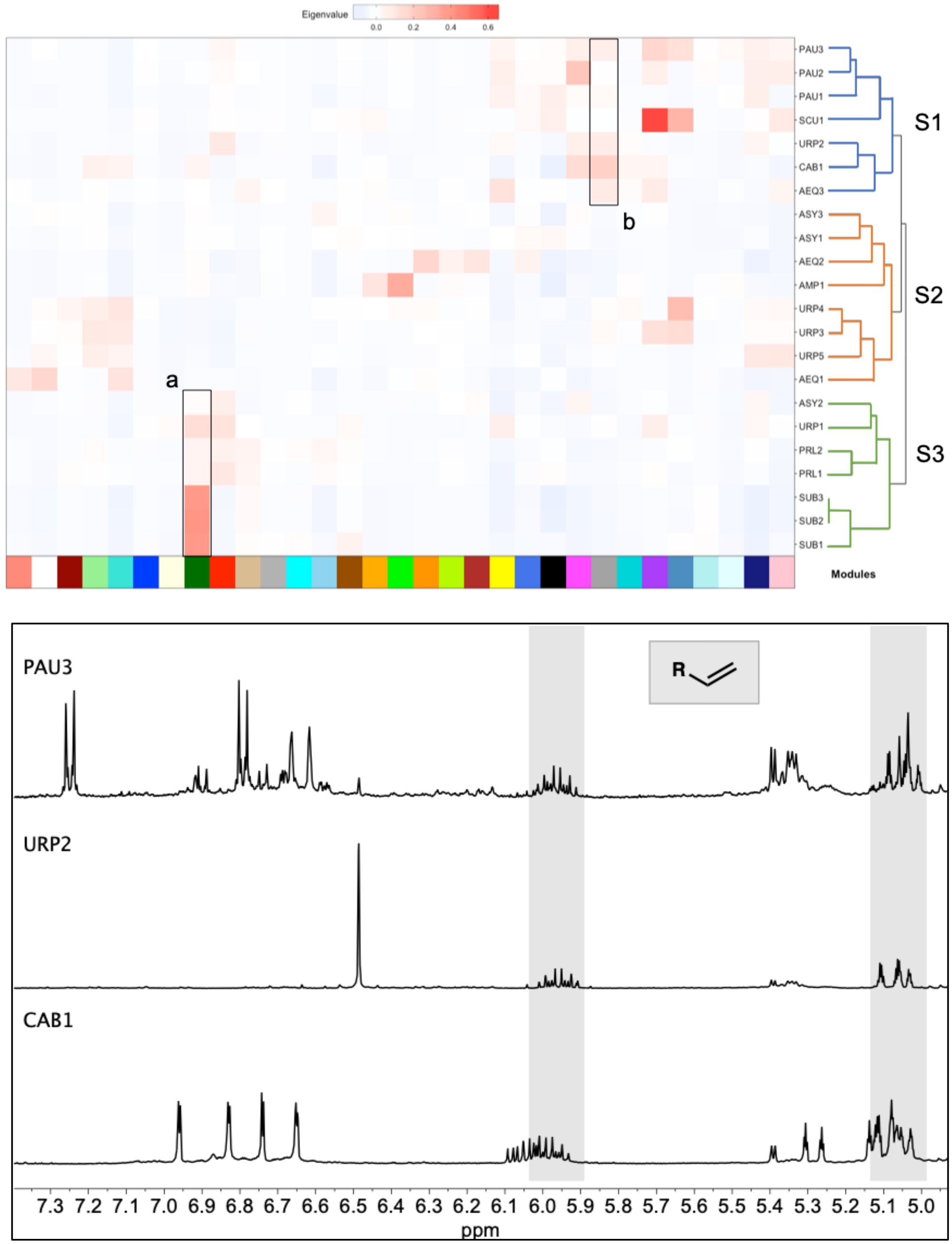
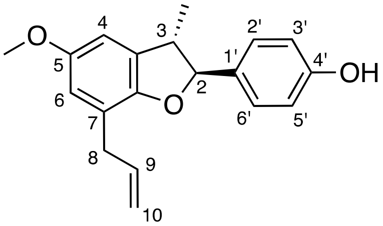


Figure 3.3. HCA of the clade *Schilleria*. Top: heatmap of module eigenvalues, with samples organized according to chemical similarity into three main clusters (*S1*-*S3*). Cluster associations with modules *darkgreen* (a) and *grey60* (b) are highlighted. Bottom: spectral features from *grey60*.

Compound isolation: Targeting the recovery of the resonance peak at ~δ_H 6.00, we identified the acetone extracts as the source for the desired compounds in each species. Pre-fractionation of the *P. paulownifolium* extract yielded the target peaks in the fractions eluted with 60% acetone:water, which upon purification resulted in 12 mg of compound **4**. Analysis of the ^{1}H NMR spectrum of this compound confirmed the presence of a 2-propenyl unit, and suggested the presence of a 1,2,3,4-tetrasubstitued ring (δ_H 6.62 and δ_H 6.59, *J* ~ 2 Hz), in addition to a 1,4-disubstitued ring system (δ_H 7.31 and δ_H 6.32). A linear sequence of resonances was also identified that included a methyl doublet coupled with a methine doublet of quartets (δ_H 1.37 and δ 3.43, *J* ~ 7 Hz), which in turn was coupled with a more deshielded methine doublet (δ_H 5.11, *J* = 9.5 Hz). Guided by the original hypothesis of structurally diverse lignans, we predicted that this group of resonances indicated a modified 1-propenyl unit that integrated the disubstituted ring and connected the two ring systems. The large coupling constant values between the two methines suggested a *trans* ring system, while the presence of the deshielded methine at δ_H 5.11 indicated an O–linked carbon, thus strongly supporting the hypothesis of a benzofuran lignan. Similarity searches against our in-house database of published *Piper* compounds (Jennifer McCracken, Hitchcock Center for Chemical Ecology) led us to a neolignan isolated from *P. aequale* (Table 3.3) whose proton resonances closely matched our experimental results.[14] We found enough overlap between the data to assume that this is the identity of the isolated compound.

Table 3.3. ^{1}H NMR assignments for the isolated compound 4 (400 Mz, $CDCl_3$) and comparison with reported data[14]



4

Position	δH, type	Δ δH (exp. - lit.[14])
2	5.11 d (9.4 Hz)	0
3	3.43 dq (9.3, 7.0 Hz)	-0.01
3-Me	1.37 d (6.8 Hz)	0.02
4	6.59 m	-0.02
5-OMe	3.87 s	0
6	6.62 m	-0.01
8	3.35 dt (6.7, 1.5 Hz)	0
9	5.98 ddt (16.8, 10.0, 6.7 Hz)	0.01
10	5.05–5.14 m	0
2'/6'	7.31 d (8.6 Hz)	0.05
3'/5'	6.82 d (8.6 Hz)	0.04

Pre-fractionation of the *P. cabagranum* extract yielded the target peaks in the fractions eluted with 50% acetone:water, and subsequent purification resulted in 28 mg of compound **5** (Figure 3.4 and Table 3.4). It was immediately apparent from the ^{1}H NMR spectrum that this molecule contained two distinct propenyl units, one of which was substituted at the position 1, resulting in a *ddd* pattern for the vinyl-methine resonance (H-8'). We verified from the ^{1}H{^{1}H} COSY spectrum that the corresponding benzyl methine (H-7') displayed a resonance peak at δH 5.07, thus strongly suggesting a hydroxyl substituent. The compound also showed three methyl singlets at δ_H 3.5–4.0, suggesting the presence of methoxyl groups. Proton resonances in the aromatic region indicated the presence of two pairs of meta-coupled protons (δ_H 6.96/6.74 and δ_H 6.83/6.65, *J* ~ 2 Hz), each pair displaying $^1H\{^{13}C\}$ HMBC signals (Table 3.4) with two aromatic O-substituted carbons (δ_C 144–154) and one of the benzylic carbons (δ_C 76.0 and δ_c 40.8, respectively). We concluded from this data and from the absence of otherwise indicative resonances, that the two ring systems must be directly connected, although we found no supporting evidence from the $^1H\{^{13}C\}$ HMBC spectrum for the long range $^1H\{^{13}C\}$ coupling between the two rings. Still, $^1H\{^1H\}$ NOESY correlations between the least deshielded protons in each ring supported the proximity of the two rings through direct linkage (Table 3.4). We believe that the two rings are oriented with a dihedral angle of 0–90º, which causes the inner protons H6/H6' to experience anisotropic shielding. Lastly, through $^1H\{^1H\}$NOESY we located three methoxyl groups across the aromatic rings, which supported that the lone phenol was *para* to the modified propenyl moiety. HRESIMS on positive mode produced an ion of mass m/z 379.1551 $[M + Na]^+$, which is coherent with the determined structure.

Table 3.4. NMR assignments for the isolated compound 5 (400 Mz, CD_3OD)

Position	δ_C, type	δ_H	$^1H\{^{13}C\}$ HMBC*	$^1H\{^1H\}$ NOESY
1	126.7, C			
2	146.2, C			
2-OMe	61.1, CH_3	3.58 s	2	3-OMe, 6'
3	153.9, C			
3-OMe	56.3, CH_3	3.86 s	3, 4	2-OMe, 4
4	113.1, CH	6.83 d (2.1 Hz)	2, 3, 5, 6, 7	3-OMe, 7
5	136.9, C			
6	124.4, CH	6.65 d (2.2 Hz)	1, 2, 3, 4, 7	7, 6'
7	40.8, CH_2	3.35 br d (6.5 Hz)	4, 5, 6, 8	4, 6, 8, 9
8	138.9, CH	5.98 ddt (16.8, 10.0, 6.7 Hz)	5, 7	7
9	116.0, CH_2	5.01–5.12		
1'	133.9, C			
2'	144.3, C			
3'	149.1, C			
3'-OMe	56.5, CH_3	3.90 s	3', 4'	4'
4'	109.9, CH	6.96 d (2.0 Hz)	2', 3', 5', 6', 7'	3'-OMe, 7'
5'	135.2, C			
6'	122.6, CH	6.74 d (2.1 Hz)	1', 2', 3', 4', 7'	6, 2-OMe, 7'
7'	76.0, CH	5.07 d (5.6 Hz)	8', 9' (weak)	4', 6'
8'	142.3, CH	6.05 ddd (17.1, 10.3, 5.9 Hz)	5', 7'	4', 6'
9'	114.5, CH_2	5.28 dt (17.1, 1.6 Hz) 5.15–5.11	8' (weak)	

* HMBC correlations are from the proton(s) to the indicated carbon.

We also verified the resonances from the target propenyl moiety in other fractions collected from the purification of compound **5**, as well as 70% acetone:water fraction from the pre-fractionation of the acetone extract. The purification of these fractions resulted in four compounds (Figure 3.4 and Table 3.5) that retained most of the structural features present in compound **5**, but that also captured a sequence of modifications in the propenyl moiety connected to the hydroxylated ring (Figure 3.6). For instance, compound **6**, isolated from the 70% acetone:water eluent, represented the reduced precursor of compound **5**, lacking the benzylic hydroxyl group. Compound **7** and **8**, isolated concomitantly from the purification of compound **5**, were characterized as the ketone and aldehyde functionalized products. Compound **9** was also isolated from the the 70% acetone:water eluent, but it was later found to be a predominant component of the hexane extract. It displayed a more complex spectrum, containing a series of multiplets in the δ_H 0.5–2.0 range that suggested the presence of a terpenoid-like motif. The methine at δ_H 5.49 indicated that this group contained a trisubstituted alkene and from the pair of methyl doublets at ~δ_H 0.85 that it must also contain an isopropyl moiety. ^{1}H{^{1}H} COSY correlations from these two units suggested a [2.2.2]bicyclic motif, and further support from a key ^{1}H{^{13}C} HMBC correlation between the methine at δ_H 3.50 and the carbon at δ_C 202.4 revealed that this 12-carbon system was connected to the aromatic ring through a ketone group (Figure 3.5). The configuration of the carbons C-8' and C-12' was assumed from the NOESY correlation between their corresponding protons, which must be in closer proximity (Figure 3.5). Further 2-D NMR correlations led to the assignment of compound **9**, which is postulated to be the product of Diels-Alder coupling between the enone **7** and the monoterpene α-phellandrene (Figure 3.6). To the best of our knowledge, this new molecule represents a novel late-stage merger between terpene and lignan biosynthesis through a unique Diels-Alder reaction.

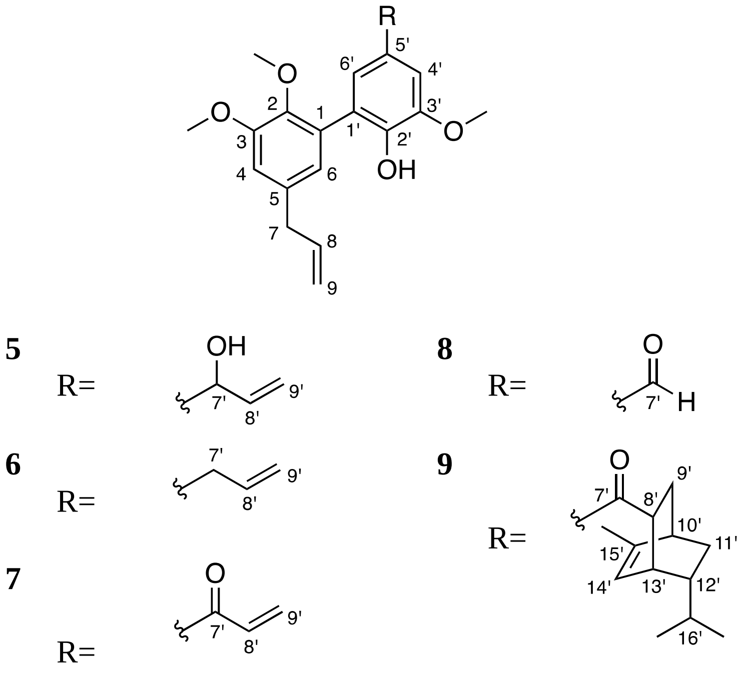


Figure 3.4. Compounds isolated from *P. cabagranum*.

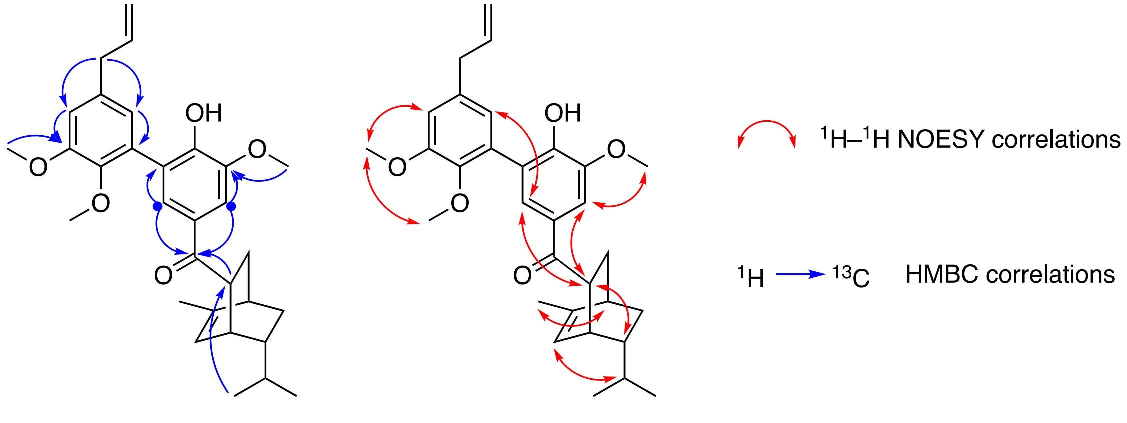

OH
O
O
O
O
OH
O
O
O
O
^{1}H–^{1}H NOESY correlations
^{1}H → ^{13}C HMBC correlations

Figure 3.5. HMBC and NOESY correlations identified for compound **9**.

Table 3.5. Partial NMR assignments for the isolated compounds 6–9 (400 Mz, CD$_3$OD)

	6		**7**		**8**		**9**	
Position*	δ_C, type	δ_H (*J* in Hz)	δ_C, type	δ_H (*J* in Hz)	δ_C, type	δ_H (*J* in Hz)	δ_C, type	δ_H (*J* in Hz)
1'	133.8, C		132.9, C		132.5, C		133.1, C	
2'	144.1, C		151.1, C		152.1, C		150.3, C	
3'	149.0, C		149.3, C		149.5, C		149.1, C	
3'-OMe	56.5, CH$_3$	3.87 s	56.6, CH$_3$	3.98 s	56.7, CH$_3$	3.98 s	56.6, CH$_3$	3.95 s
4'	112.0, CH	6.76 d (2.0)	110.9, CH	7.59 d (2.1)	109.5, CH	7.46 d (1.9)	110.8, CH	7.48 d (2.0)
5'	131.9, C		129.4, C		129.5, C		128.8, C	
6'	124.1, CH	6.56 d (2.0)	127.4, CH	7.52 d (2.1)	129.9, CH	7.36 d (1.9)	126.5, CH	7.44 d (2.0)
7'	40.9, CH$_2$	3.34–3.31 d	191.1, C		193.0, C	9.76 s	202.4, C	
8'	139.4, CH	5.92–6.04 ddt	133.4, CH	7.33 dd (17.0, 10.6)			48.4, CH	3.50 ddd (9.4, 5.8, 1.9)
9'	115.6, CH$_2$	5.00–5.13	129.6, CH$_2$	6.37 dd (17.0, 2.0) 5.87 dd (10.6, 2.0)			29.5, CH$_2$	1.66–1.76 m
10'							37.5, CH	2.40 m
11'							32.7, CH$_2$	1.80, 0.97 m
12'							48.5, CH	1.48 m
13'							38.7, CH	2.93 dt (6.5, 2.0)
14'							121.9, CH	5.49 br d (6.2)
15'							144.6, C	
15'-Me							20.0, CH$_3$	1.76 d (1.7)
16'							34.5, CH	1.08 m
16'-Me							21.7, CH$_3$	0.88 d (6.5)
							20.9, CH$_3$	0.82 d (6.6)

* Positions 1–9 are identical or very similar to **5** (Table 3.4).

3.4 Discussion

In this study, we surveyed a remarkably diverse collection of phytochemical data, which by means of network analysis was consolidated into a few descriptors (modules) of commonly expressed molecular features. By comparing module expression across samples, we identified patterns of chemical diversification in *Piper* that were far more complex than the phylogenetic classification into clades, suggesting that phytochemical traits are dispersed. For instance, we verified clusters of chemically similar samples that were most consistently aligned with the clades *Radula* and *Peltobyron* (Figure 3.2), although samples identified with those clades were also distributed in clusters that had a more heterogeneous species composition. Uckele *et al* arrived at a similar conclusion from the analysis of ^{1}H NMR profiles of 65 Neotropical *Piper* species, supported by a direct and highly resolved phylogenomic analysis of the investigated species.[15] They detected no phylogenetic signal at the level of compound structural resolution obtained from ^{1}H NMR data, but clearer patterns of trait distribution across *Radula* emerged when phytochemical profiles were characterized by the qualitative presence/absence of general metabolite classes. Likewise, we observed that samples were more chemically dissimilar when considering then entire ^{1}H NMR profile than when utilizing module expression profiles (Table 3.2). This divergence suggests that while early biosynthetic products are moderately conserved, chemical traits emerging from the late-stage modification of common precursors are phylogenetically labile.[15] We thus speculate that the network modules distinguished compound traits from both early- and late-stage biosynthetic steps, but that modules containing broader compound class features were more important for the clustering of *Piper* species of the same clade. Structural annotation of the modules might validate these hypotheses, but a yet stronger evidence for the conserved distribution chemical traits might be constructed by incorporating complete phylogenetic data into this analysis.

Our results partially differ from those obtained by Salazar and collaborators who found no phylogenetic signal of chemical composition for a smaller collection of *Piper* species from Costa Rica.[16] In that study, phytochemical profiles were surveyed by GC-MS, and that is likely a determining element to explain the different results obtained in each study. MS is a useful technique to reveal similarities in compound composition, but that type of information represents a limited snapshot of the chemical overlap between different species and does not distinguish between isomeric molecules and lacks the functional group-level resolution required to make phylogenetic references. In contrast, ^{1}H NMR data describes varying levels of structural complexity, from conserved class-specific motifs to generalized compound modifications. It should be noted, however, that differences in the sampling effort (geographical location, replicates per species and total number of species) may also have an influential effect in the conclusions obtained from each study.

We verified that intra-clade chemical similarity decreased with the number of species, but that trend was particularly exacerbated in *Schilleria*, which showed an average intra-clade sample distance comparable to *Peltobyron* with just half the number of species. *Schilleria* was also the clade with the most chemically dissimilar samples at the intraspecific level, despite the fact that species representation was consistent across all clades (2.5 ± 0.4 sample/species). A possible reason for this outstanding diversity comes from uncertain taxonomical identification of some *Piper* species, which might be particularly prominent in *Schilleria*. For example, *P. amphyoxis*, *P. asymmetricum*, *P. cabagranum* and *P. perlaense* are listed as synonyms of *P. aequale*,[17] and the HCA revealed that AEQ samples were assigned to distinct clades of the dendrogram, where they were chemically more similar to other synonym species (Figure 3.3). It should also be considered, however, that these differences might reflect an actual diversity of chemotypes within the same population. The *Schilleria* species sampled in this study were represented by individuals collected from the same geographical locations, and as it has been demonstrated by Richards et al, phytochemical diversity has a beneficial effect on plant communities by reducing herbivory.[18] While these inferences can only be fully supported by a direct phylogenetic analysis of the samples in question, the intraspecific variation observed in *Schilleria* demonstrates how phytochemical data can also supplement and help resolving conflicting taxonomical classifications.

Although the HCA on module profiles only partially grouped samples from the same clade, MANOVA revealed significant module associations for all *Piper* clades. This divergence indicates that the phylogenetic signal of secondary metabolites was rather defined by a few chemical characters. Thus, statistical methods that evaluate the individual contribution of modules should be prioritized for the detection of phylogenetically relevant markers. This strategy allowed us to identify the predominance of lignan motifs in samples from the clade *Schilleria* (Figure 3.3), and we project that the same approach can also reveal conserved classes of specialized metabolites from other *Piper* clades. It is noteworthy that although the most significant module for *Schilleria* (*darkgreen*) was highly specific to grandisin, its association with samples of the cluster *S1* (Figure 3.3) resulted from a different compound. Therefore, the relative magnitude of module eigenvalues can be interpreted as (i) the differential expression of a particular compound, or (ii) the degree of structural similarity between different compounds. Effect (i) dictated the relationship between samples in the cluster *S3*, which contain grandisin in varying concentrations, while the association between *darkgreen* and CAB1 was driven by (ii). We deem this latter interpretation particularly useful for the prospection of structurally diverse compounds within a clade, as evidenced by the successful isolation of lignans in this study.

Our results reinforce the evolutionary relevance of lignans by suggesting the conservation of their biosynthesis in the *Piper* clade *Schilleria*, and principally by demonstrating the structural diversification of these compounds within the clade. We encountered lignans formed from three distinct linkages in three different *Piper* species (Figure 3.6), and that is certainly an understatement to the chemical diversity present in the clade. For example, URP2 contains a major lignan whose spectrum combines elements from both *P. cabagranum* and *P. subscutatum* (Figure 3.3), and the cluster *S2* possibly contains structural motifs that are still more distinct from those represented in *S1* and *S3*. Moreover, the minor compounds isolated from *P. cabagranum* demonstrate that further modifications from the main lignan motifs contribute to increase compound diversity and complexity, which might be important for intraspecific chemical variations within the clade. We cannot ascertain the evolutionary implications of these findings without the support of direct phylogenetic data, but by characterizing the compounds represented by other *Schilleria*-related modules, we might be able to provide a stronger phytochemical basis for this argument.

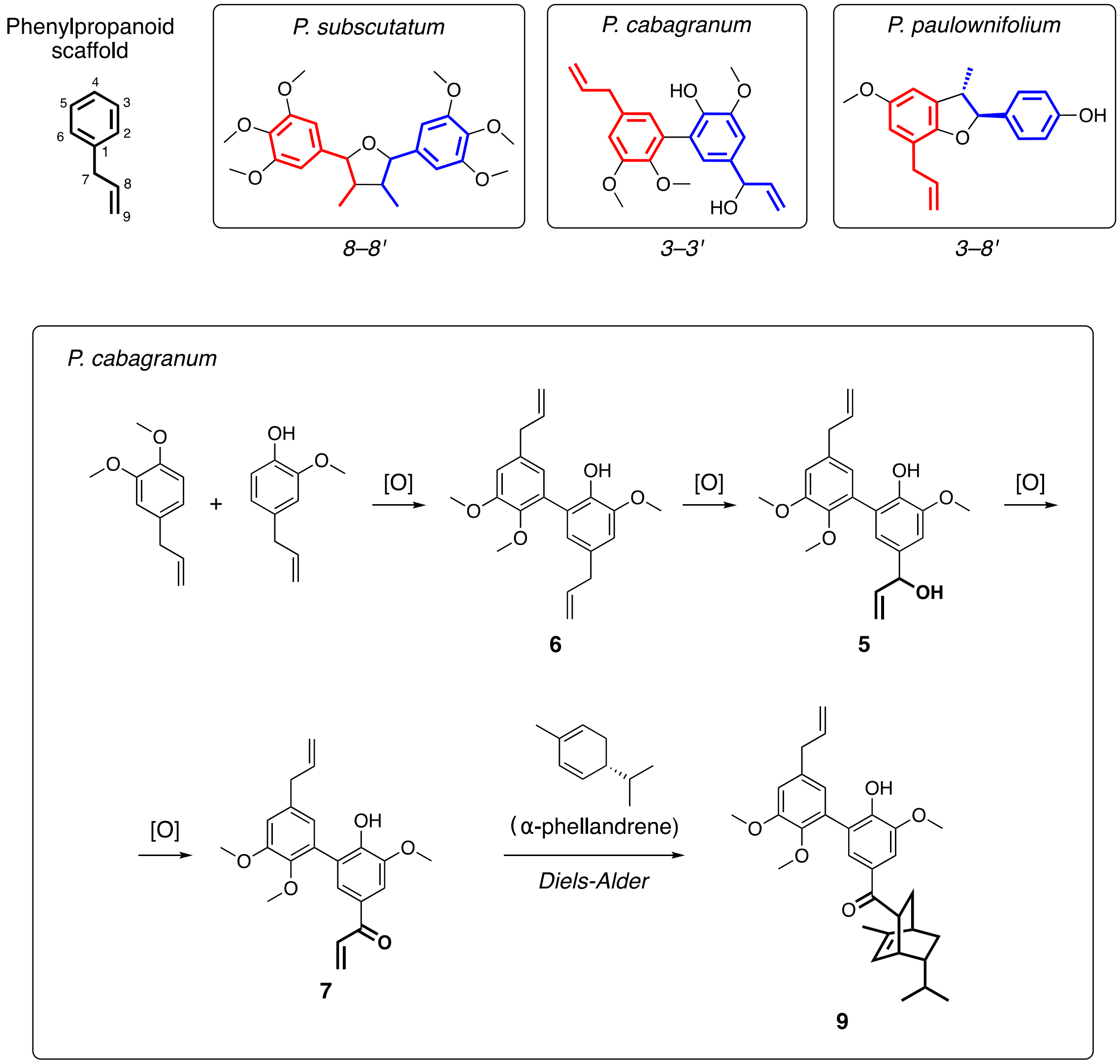


Figure 3.6. Lignans identified in the clade *Schilleria*. Top: structural variation observed in the lignans from three *Piper* species as a result of different linkage modes between the phenylpropanoid units. Bottom: proposed biosynthetic steps for the lignans isolated from *P. cabagranum*.

3.5 Conclusions and future directions

Specialized metabolites account for a large portion of the evolutionary history impressed into phylogenies. As such, the integration between genetic and phytochemical data provides a powerful means to retrace the diversification of plant taxa and to expand our understanding of chemically mediated adaptations. We herein demonstrated that our ^{1}H NMR-based network approach can support this integration, revealing conserved chemical traits even at the high level of intra- and interspecific diversity present in *Piper*. The main advantage of this methodology is that it permits the quantification of chemotypes into modules eigenvalues, which can be used to estimate chemical diversity across taxa. This feature can be particularly useful in the context of ecological and populational studies that consider phytochemical diversity as an explaining variable for community dynamics. A second advantage, as shown with the clade *Schilleria*, is that it facilitates the characterization of chemo-phylogenetic signals, whether that is through the comparative analysis of significant modules, or the isolation of compounds guided by specific spectral features.

This study was mostly focused on the chemical characterization of the clade *Schilleria*, but we also found significant module interactions for the other clades. We thus anticipate that future directions may undertake the investigation of those phylogenetic associations, which ultimately will contribute to consolidate a chemical phylogeny for the genus *Piper*. A promising modification to the method could include pre-fractionation of the extracts to emphasize minor compounds which would otherwise not be detected by the analysis, as observed with *P. cabagranum*. Lastly, the conclusions from our study are based on indirect inferences from a generic phylogeny of the genus *Piper*. Paired chemical analysis using direct phylogenetic data from species subsets should further qualify the generalization of these results.

3.5 References

1. de la Torre, M. C.; Sierra, M. A., Comments on Recent Achievements in Biomimetic Organic Synthesis. *Angewandte Chemie International Edition* **2004,** *43* (2), 160-181.

2. Robinson, R., LXIII.—A synthesis of tropinone. *Journal of the Chemical Society, Transactions* **1917,** *111* (0), 762-768.

3. Firn, R. D.; Jones, C. G., Natural products – a simple model to explain chemical diversity. *Natural Product Reports* **2003,** *20* (4), 382-391.

4. Larsson, S., The “new” chemosystematics: Phylogeny and phytochemistry. *Highlights in the Evolution of Phytochemistry: 50 Years of the Phytochemical Society of Europe* **2007,** *68* (22), 2904-2908.

5. Edger, P. P.; Heidel-Fischer, H. M.; Bekaert, M.; Rota, J.; Glockner, G.; Platts, A. E.; Heckel, D. G.; Der, J. P.; Wafula, E. K.; Tang, M.; Hofberger, J. A.; Smithson, A.; Hall, J. C.; Blanchette, M.; Bureau, T. E.; Wright, S. I.; de Pamphilis, C. W.; Schranz, M. E.; Barker, M. S.; Conant, G. C.; Wahlberg, N.; Vogel, H.; Pires, J. C.; Wheat, C. W., The butterfly plant arms-race escalated by gene and genome duplications. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (27), 8362-8366.

6. Adamek, M.; Alanjary, M.; Ziemert, N., Applied evolution: phylogeny-based approaches in natural products research. *Nat Prod Rep* **2019,** *36* (9), 1295-1312.

7. Kang, H.-S., Phylogeny-guided (meta)genome mining approach for the targeted discovery of new microbial natural products. **2017,** *44* (2), 285-293.

8. Ernst, M.; Nothias, L. F.; van der Hooft, J. J. J.; Silva, R. R.; Saslis-Lagoudakis, C. H.; Grace, O. M.; Martinez-Swatson, K.; Hassemer, G.; Funez, L. A.; Simonsen, H. T.; Medema, M. H.; Staerk, D.; Nilsson, N.; Lovato, P.; Dorrestein, P. C.; Rønsted, N., Did a plant-herbivore arms race drive chemical diversity in <em>Euphorbia?</em>. *bioRxiv* **2018**, 323014.

9. Jaramillo, M. A.; Ricardo, C.; Christopher, D.; James, F. S.; Angela, C. S.; Eric, J. T., A Phylogeny of the Tropical Genus *Piper* Using ITS and the Chloroplast Intron *psbJ–petA*. *Systematic Botany* **2008,** *33* (4), 647-660.

10. Martínez, C.; Carvalho, M. R.; Madriñán, S.; Jaramillo, C. A., A Late Cretaceous Piper (Piperaceae) from Colombia and diversification patterns for the genus. *American Journal of Botany* **2015,** *102* (2), 273-289.

11. Teponno, R. B.; Kusari, S.; Spiteller, M., Recent advances in research on lignans and neolignans. *Natural Product Reports* **2016,** *33* (9), 1044-1092.

12. Gottlieb, O. R. In *Chemistry of Neolignans with Potential Biological Activity*, Berlin, Heidelberg, Springer Berlin Heidelberg: Berlin, Heidelberg, 1977; pp 227-248.

13. Ramirez, J.; Gilardoni, G.; Gozzini, D.; Boiocchi, M.; Malagon, O.; Finzi, P. V.; Vidari, G. In *Estudo fitoquimico de las plantas ecuatorianas: Piper subscutatum C.DC. y Lepechinia mutica BENTH*, XXII Congresso Italo-Latinomericano de Etnomedicina, Puntarenas, Costa Rica, Revista clínica de la escuela de medicina UCR-HSJD: Puntarenas, Costa Rica, 2013; pp 143-144.

14. Maxwell, A.; Dabideen, D.; Reynolds, W. F.; McLean, S., Neolignans from Piper aequale. **1999,** *50* (3), 499-504.

15. Uckele, K. A.; Jahner, J. P.; Tepe, E. J.; Richards, L. A.; Dyer, L. A.; Ochsenrider, K. M.; Philbin, C. S.; Kato, M. J.; Yamaguchi, L. F.; Forister, M. L.; Smilanich, A. M.; Dodson, C. D.; Jeffrey, C. S.; Parchman, T. L., Evidence for both phylogenetic conservatism and lability in the evolution of secondary chemistry in a tropical angiosperm radiation. *bioRxiv* **2020**, 2020.11.30.404855.

16. Salazar, D.; Jaramillo, M. A.; Marquis, R. J., Chemical similarity and local community assembly in the species rich tropical genus Piper. *Ecology* **2016,** *97* (11), 3176-3183.

17. Tropicos.org. http://www.tropicos.org/Name/25001129 (accessed 05 Nov 2020).

18. Richards, L. A.; Dyer, L. A.; Forister, M. L.; Smilanich, A. M.; Dodson, C. D.; Leonard, M. D.; Jeffrey, C. S., Phytochemical diversity drives plant-insect community diversity. *Proc. Natl. Acad. Sci. U. S. A.* **2015,** *112* (35), 10973-10978.

Conclusions

The metabolome is a composite variable that encapsulates the evolutionary history of an organism, and for that reason, it is an important source of natural data for the study of ecosystems. For example, the study of *P. kelleyi* highlighted in Chapter 1 demonstrated how resource availability and herbivore/parasite pressures are simultaneously implicated in metabolic plasticity through plant development, and how the phytochemical products resulting from these adaptations influence predator community. ^{1}H NMR analysis enables the access to a high-degree of structural information from these complex biological mixtures, thus allowing us to establish organismal comparisons that could not be as feasibly attained with a pure focus on compositional analysis. Thus, partial structural overlap between metabolic profiles may be informative of chemo-phylogenetic relationships at varying taxonomic levels, as shown in Chapter 3, but more importantly, it can highlight a biochemical convergence of different organisms towards the production of compounds with similar biological properties, as demonstrated in Chapter 2.

Our network approach addresses two major limitations that emerges from the informational content of the ^{1}H NMR spectrum, namely, intensified peak overlap and variable redundancy. By consolidating patterns of peak co-variance into modules, the analysis primarily operates a variable reduction transformation, but this process also helps resolving certain convoluted regions of the spectra, such as the glycosylated region (δ_H 3-4.5). Another important aspect of this pattern recognition approach is the filtering of uncorrelated peaks, which are irrelevant in the context of comparative analyses of large sample collections. Altogether, the modules are a simplified and more meaningful representation of the metabolomic space. They can serve as composite variables in discriminant and exploratory analyses to verify sample segregation from which inferences may be drawn on the association of certain modules with specific groups. Moreover, the chemical identity of each module can be more easily determined based on the specific peak combinations they represent, which in some instances may even preclude compound isolation for the complete characterization.

Trait quantification through module eigenvalues is arguable the most important result from this approach, as it enables the direct association between chemotype and biological information. That can be a categorical value (*P. kelleyi* developmental stages in Chapter 1 and *Piper* clades in Chapter 3) or a quantifiable variable (mixture composition in Chapter 1 and assay activities in Chapter 2). Similarity measurements can also be obtained from the module eigenvalues, not only as a means to determine the chemical proximity between samples, but also as a potential measurement of chemical heterogeneity across different groups. For example, average intra-group distances may be calculated based on module profile for organisms two different physiological states, where a significantly higher distance could imply a state that leads to higher diversity. Not addressed in this work but a potential future direction is also the employment of module eigenvalues for deriving chemical diversity measurements within sample.

The study cases highlighted in this work demonstrate that the network approach can resolve the spectral complexity of samples collections with varying levels of chemical diversification, from single species (*P. kelleyi*) to the broad survey of a chemically diverse genus (*Piper*). It is important to recognize, however, that as an analytical technique that operates *via* pattern recognition, it might encounter some limitations in systems with minimal structural overlap, as with species from broader taxonomic levels. We postulate that in this scenario module identity will be largely descriptive of chemical traits expressed at lower taxonomical ranges (e.g., within a genus), with only a few modules that are more broadly represented across samples (e.g., different genus or families). However, even if limited, the detection of structural overlap may provide evidence for concerted chemical strategies, particularly when strongly associated with biological or environmental factors. For example, diverse plant species within an experimental plot might produce similar phytochemical responses to seasonal fluctuation in precipitation, or to predation by arthropod herbivores, in which case we may anticipate the detection of commonly expressed traits in modules that have higher eigenvalues for samples subjected to these altered states.

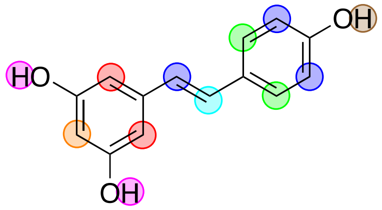
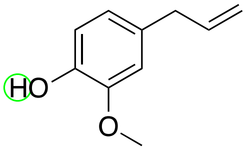
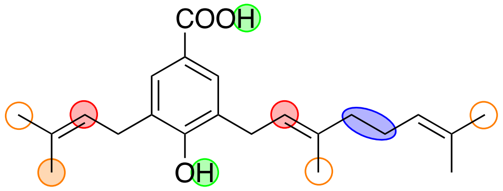
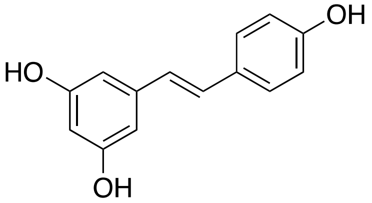
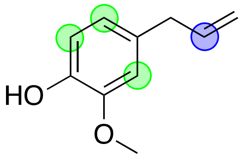
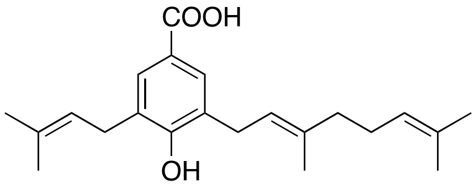
A promising future direction for this approach is the cross-platform integration with LCMS data to further improve compound deconvolution in the generated modules. Consequentially, module-based measurements of chemical diversity may be factorized into compositional and structural determinants to discern whether diversity arises from complex mixtures of compounds or from increased structural complexity of a few metabolites. This integration can also expedite the process of compound dereplication and characterization of novel molecules. Another compelling development might arise from the analysis of two-dimensional NMR data, particularly $^{1}H\{^{13}C\}$ correlation spectra, utilizing this approach. Notwithstanding the technical challenges associated with handling higher-dimensionality, network analysis on correlational data may provide an unprecedent level of structural coverage from complex mixtures and unveil more subtle and intricate chemical associations between organisms.

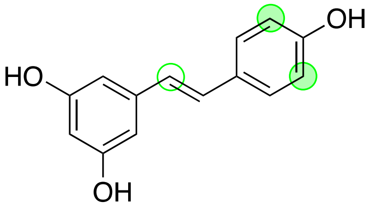
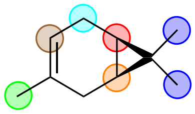
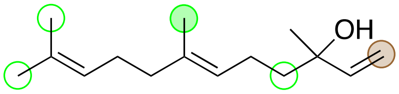
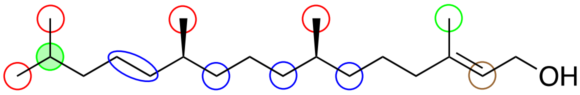
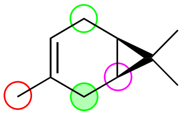
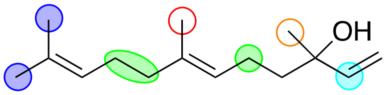
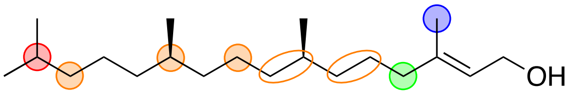
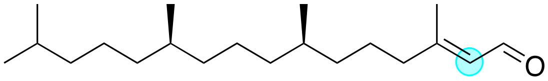
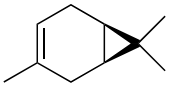
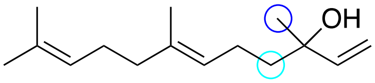
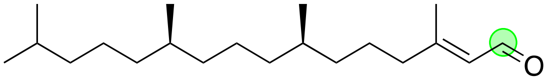
Appendix

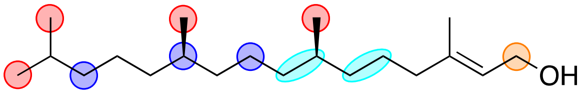
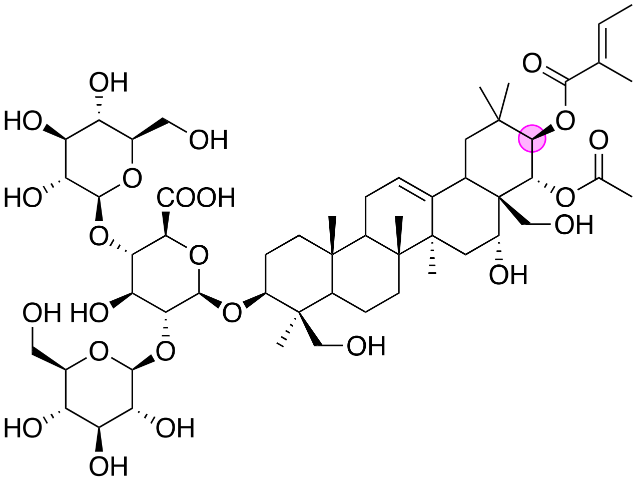
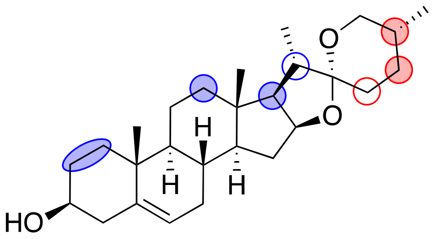
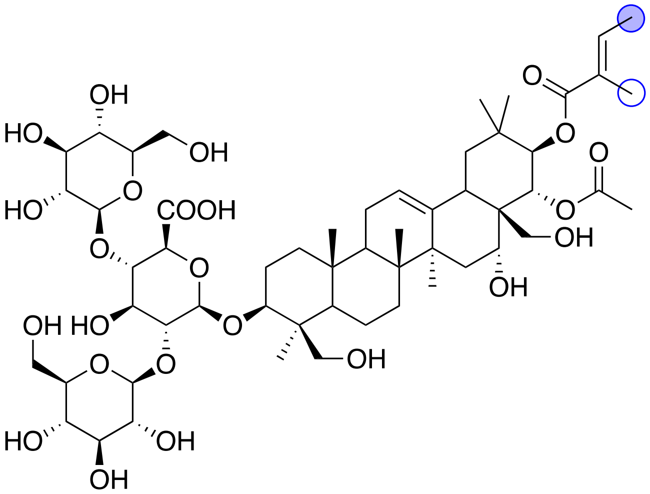
A.1. Module-compound associations from the network analysis of prepared mixtures.

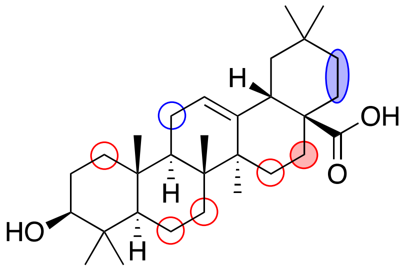
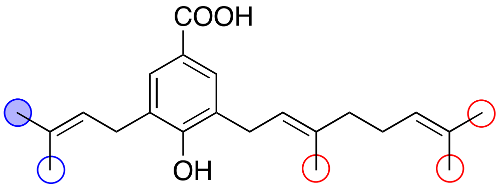
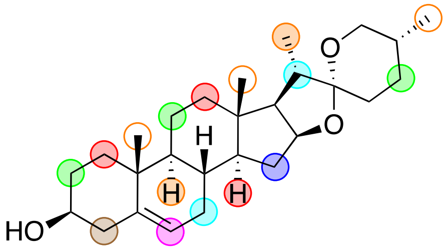
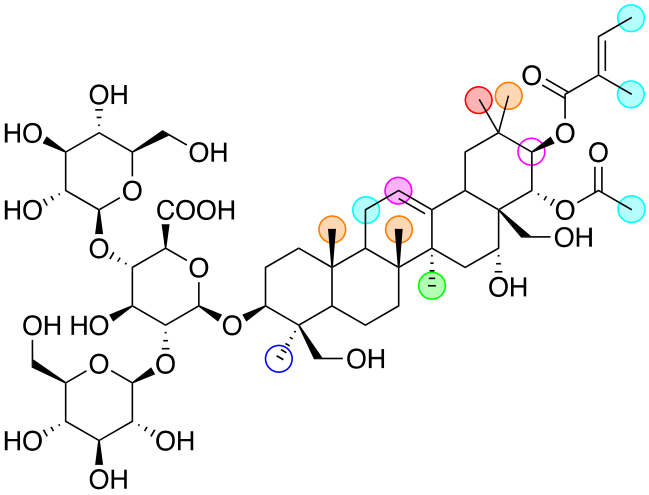
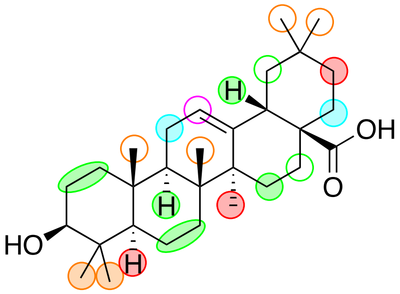
For each table, modules are named accordingly to the color code generated in the analysis, and the unified code (in parenthesis) that best describes the highlighted structural features. The representative compounds for each module are shown with their respective correlation value. The colored circles indicate proton resonances depicted by the module, whose values in ppm are displayed under the module name. Unfilled circles identified resonances within 0.1 ppm of an identified bin. Chemical shift values with no correspondence to the molecules of the module are indicated in black.

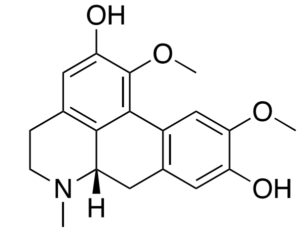
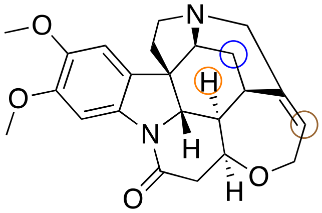
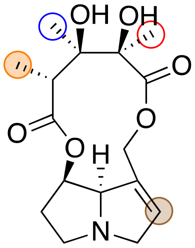
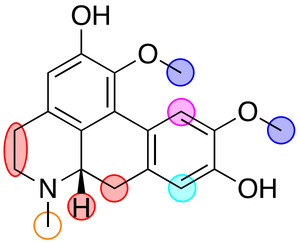
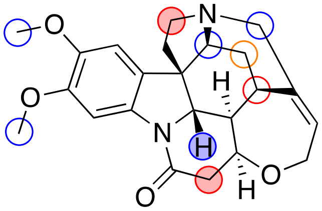
Table A.1.1. Module identity, chemical shifts and compound correlations obtained from the network analysis of intraclass mixtures.

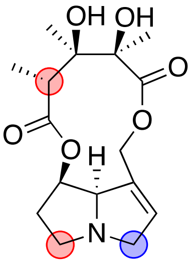
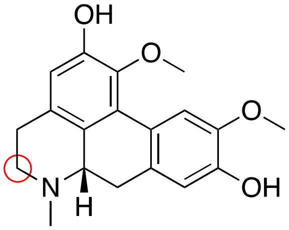
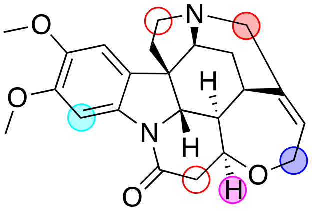
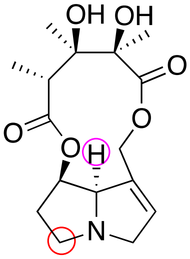
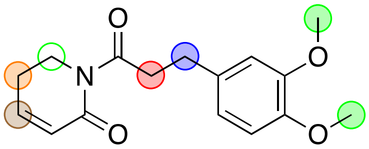
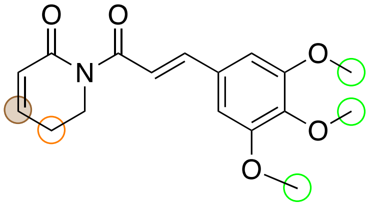
Module (δ)	Compounds (Pearson's correlation)
BROWN (PHP-1) **6.17 6.21 6.49 6.53** **6.81 6.97 7.37 9.05** **9.09 9.13 9.20 9.28** 9.48	**Resveratrol (0.93)** 
GREY 60 (PHP-2) **1.74 2.14 2.18 5.33** **9.76 9.80**	**Eugenol (0.48)** 
	PBA (0.97) 
	Resveratrol (0.46) 
CYAN (PHP-3) **5.93 5.97 6.61 6.65** **6.69 6.73 6.77**	**Eugenol (0.97)** 
	PBA (0.68) 

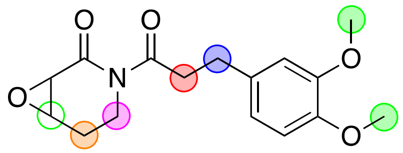
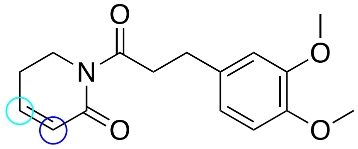
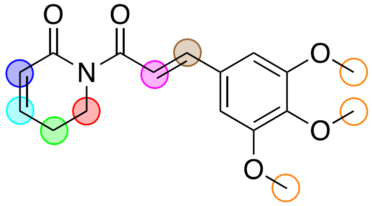
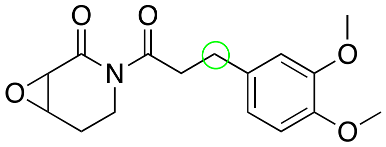
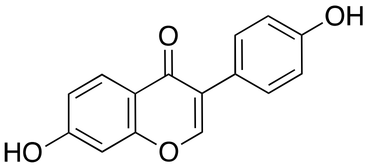
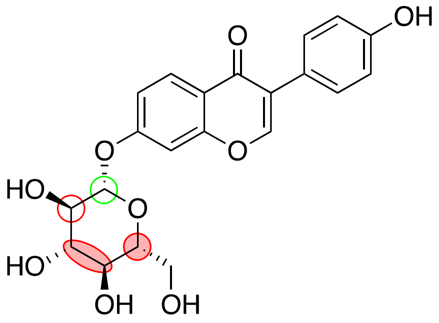
	Resveratrol (0.68) 
PINK (TPN-1) 0.62 0.66 0.70 0.74 0.78 0.82 1.06 1.62 2.22 2.34 5.25	**Carene (0.92)** 
	Nerolidol (0.58) 
	Phytol (0.51) 
SALMON (TPN-2) 0.58 1.30 1.54 1.70 2.02 2.06 2.10 5.89	**Carene (0.61)** 
	Nerolidol (0.94) 
	Phytol (0.61) 
	Phytenal 
LIGHT CYAN (TPN-3) 0.90 0.94 1.34 1.42 4.09 9.96 10.0	**Carene (0.57)** 
	Nerolidol (0.57) 
	Phytenal 

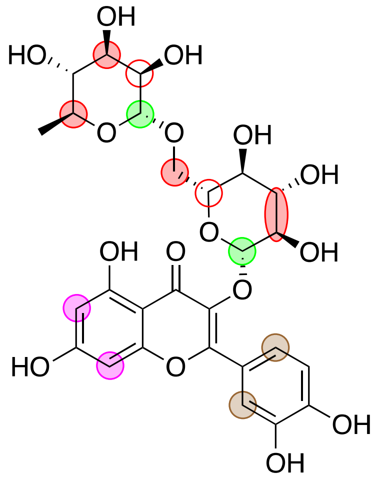
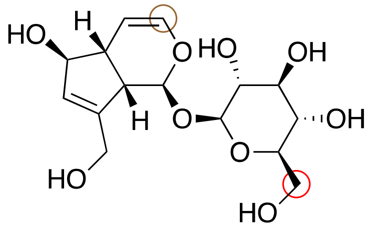
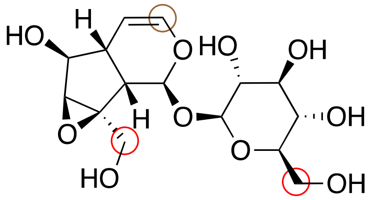
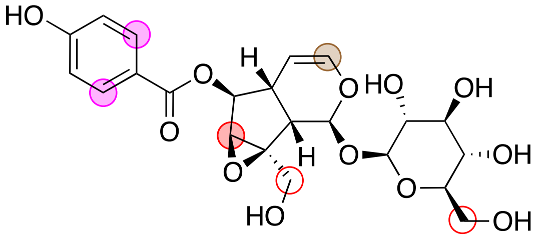
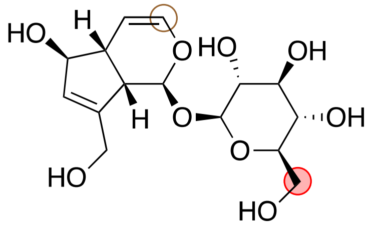
	Phytol (0.95) 
ROYAL BLUE **(STR-1)** 5.49 5.53 7.89	**Escin (0.84)**  OH HO OH HO O COOH O O OH HO O O O OH HO OH OH O O O O O OH OH
LIGHT YELLOW **(STR-2)** 1.66 1.78 1.82	**Diosgenin** **(0.63)** 
	Escin (0.54)  OH HO OH HO O COOH O O OH HO O O O OH HO OH OH O O O O O OH OH

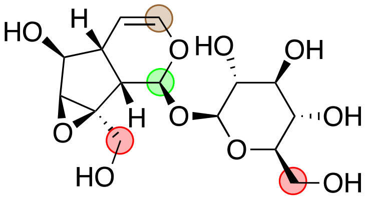
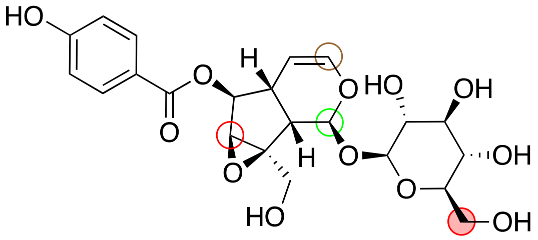
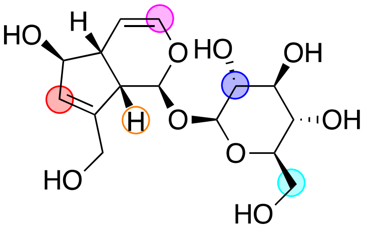
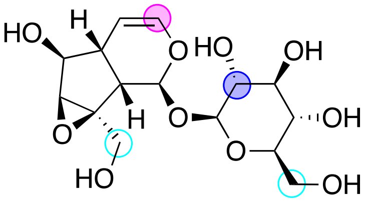
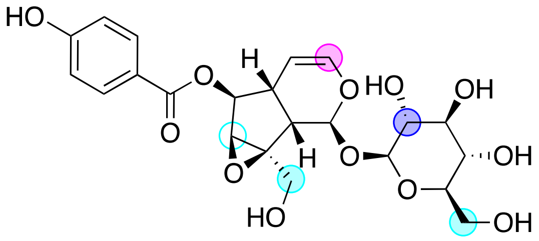
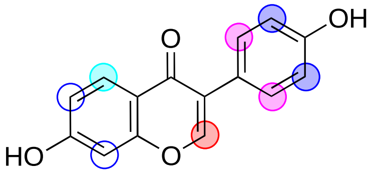
	Oleanic acid (0.56) 
	PBA (0.52) 
TURQUOISE (STR-3) **0.86 0.98 1.02 1.10** **1.18 1.26 1.50 1.58** **1.86 1.94 1.98 2.26** **5.37 5.41** 5.45 7.81	**Diosgenin (0.81)** 
	Escin (0.81) 
	Oleanic acid (0.77) 

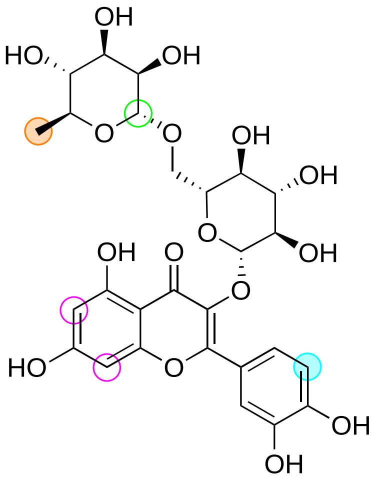
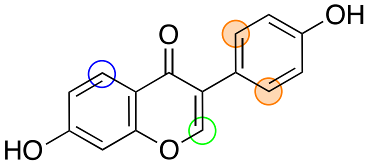
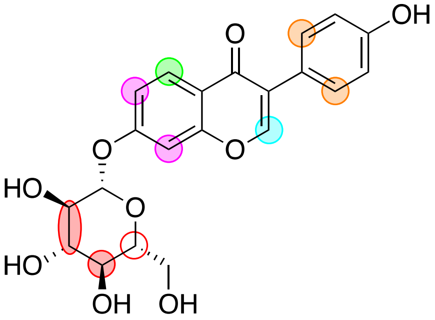
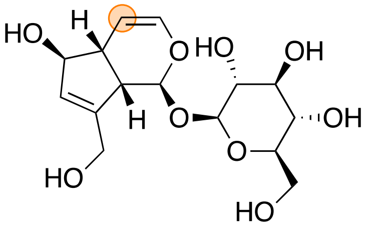
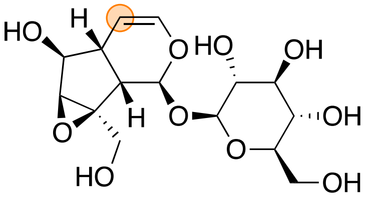
LIGHT GREEN (ALK-1) **1.22 1.38 1.46 6.09** **6.13**	**Boldine (0.52)** 
	Brucine (0.52) 
	Crotaline (0.96) 
YELLOW (ALK-2) **2.46 2.50 2.58 2.98** **3.02 3.06 3.10 3.14** **3.18 3.61 3.89 6.57** **8.01**	**Boldine (0.97)** 
	Brucine (0.66) 

	Crotaline (0.64) 
GREEN YELLOW (ALK-3) 2.74 2.78 2.82 4.13 4.17 4.33 4.37 7.77	**Boldine (0.55)** 
	Brucine (0.92) 
	Crotaline (0.53) 
PURPLE (AMD-1) 2.42 2.86 2.90 2.94 3.22 3.81 4.25 4.29 7.05	**Alkene amide (0.86)** 
	Piplartine (0.67) 

	Pipleroxide (0.90) 
MAGENTA (AMD-2) **2.54 3.93 4.01 4.05** **6.05 7.09 7.13 7.33** **7.61 7.65**	**Alkene amide (0.52)** 
	Piplartine (0.99) 
	Pipleroxide (0.53) 
GREEN (FLV-1) **3.41 3.45 4.49 3.57** **4.49 4.53 4.57 5.21** **6.25 6.45 7.69 7.73** 9.64	**Daidzein (0.51)** 
	Daidzin (0.49) 

	Rutin (0.93) 
MIDGNIGHT BLUE (IRG-1) 3.77 5.61 6.41 7.53 7.57 7.93 7.97	**Aucubin (0.51)** 
	Catalpol (0.54) 
	Catapolside (0.99) 
TAN (IRG-2) 3.69 3.73 5.29 5.65 6.33 8.57 10.68 10.96 11.0 11.28 11.84 11.88	**Aucubin (0.50)** 

	Catalpol (0.97) 
	Catapolside (0.66) 
BLACK (IRG-3) 2.30 3.26 3.65 4.21 5.77 5.81 5.85 6.33 6.37	**Aucubin (0.95)** 
	Catalpol (0.75) 
	Catapolside (0.72) 
BLUE (FLV-2) 2.70 4.45 6.29 6.93 7.29 7.45 7.49 8.09 8.21 8.29 8.33 8.37 8.45	**Daidzein (0.97)** 

	Rutin (0.43) 
RED (FLV-3) 2.62 2.66 **3.53** 4.41 **7.21** **7.25** **7.41** **8.05** **8.13** **8.17** **8.25** 8.49 8.97	**Daidzein (0.43)** 
	Daidzin (0.96) 
DARK RED (IRG-4) **5.01** **5.05** **5.09**	**Aucubin (0.5)** 
	Catalpol (0.47) 

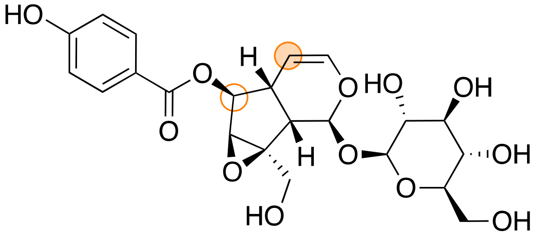
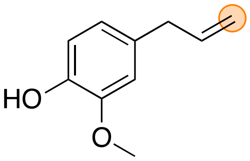
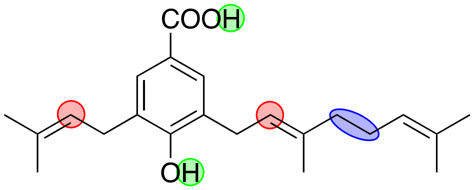
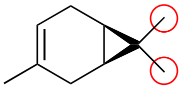
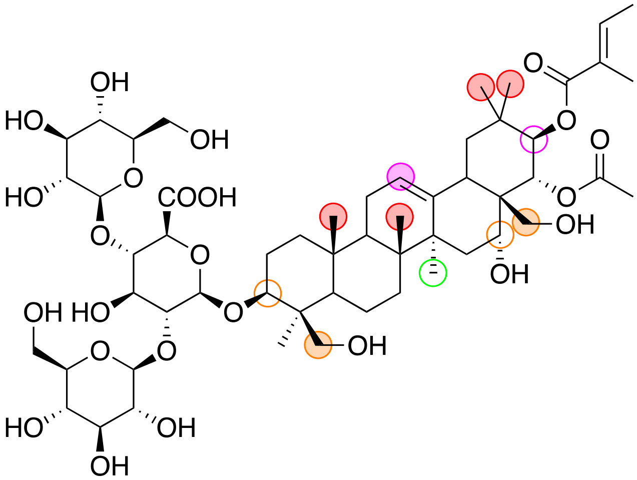
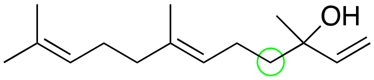
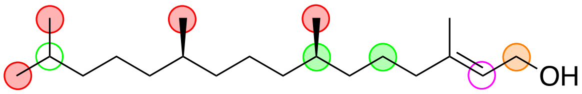
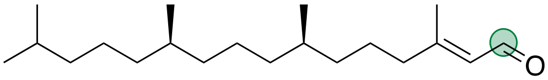
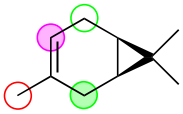
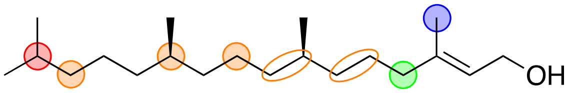
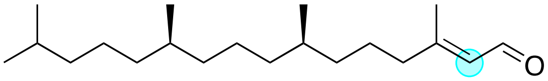
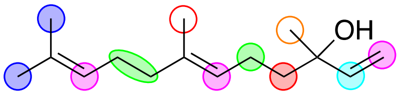
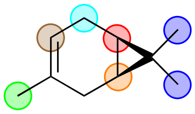
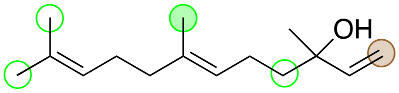
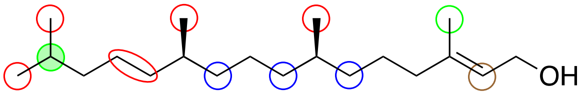
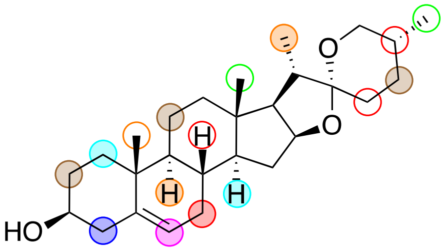
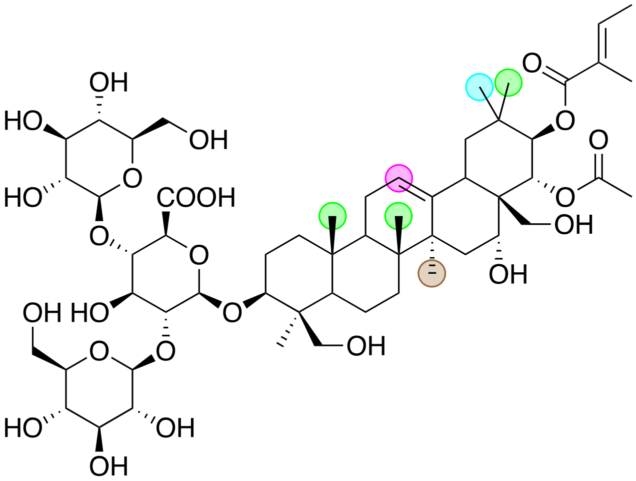
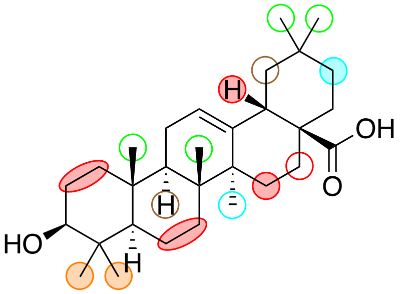
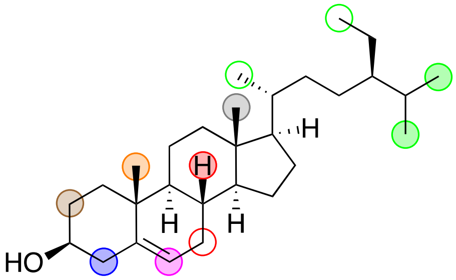
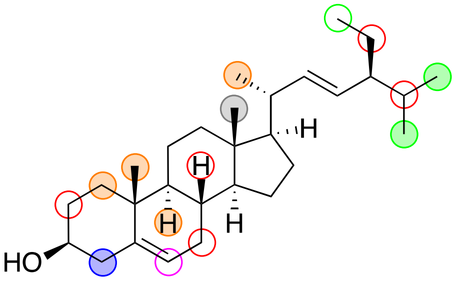
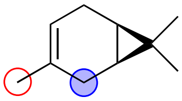
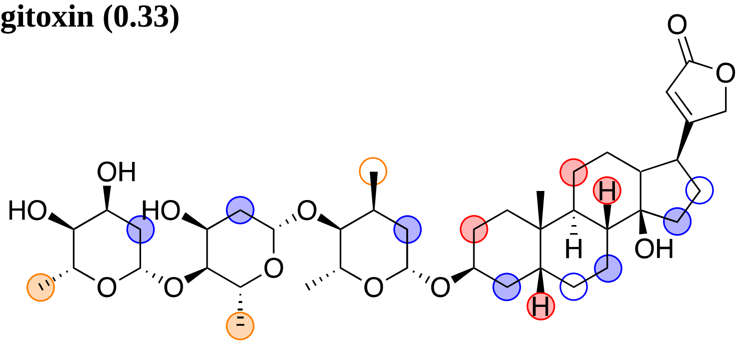
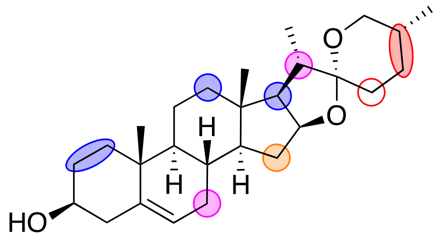
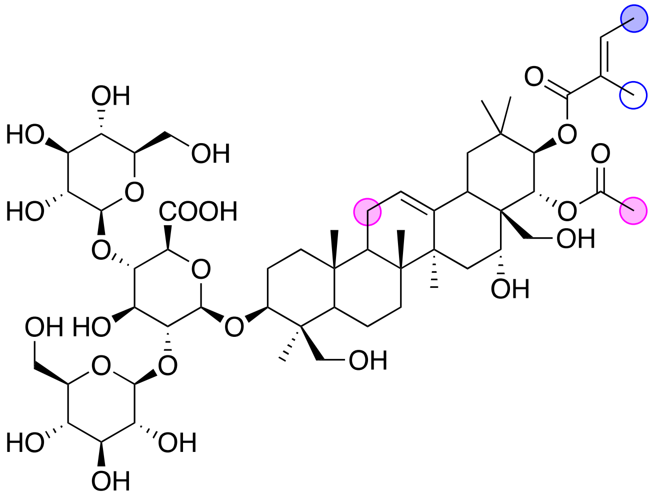
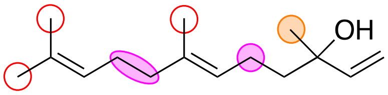
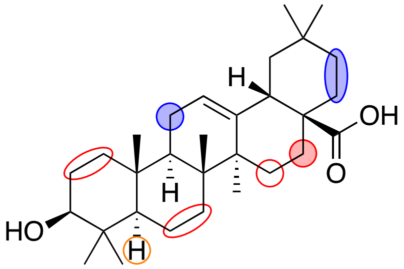
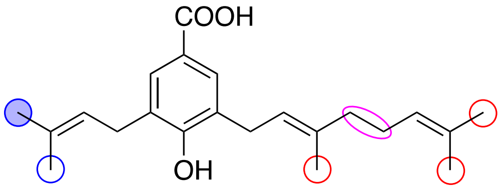
	Catapolside (0.47) 
	Eugenol (0.72) 

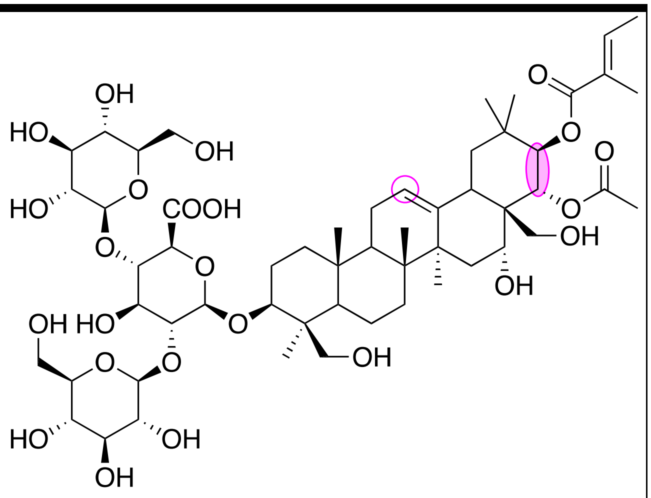
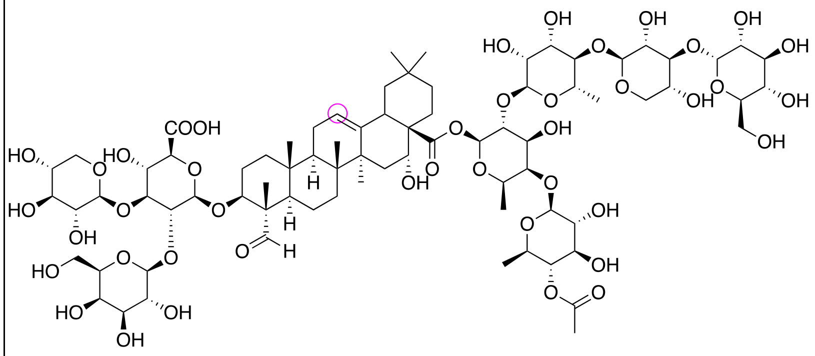
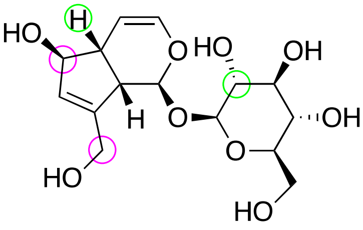
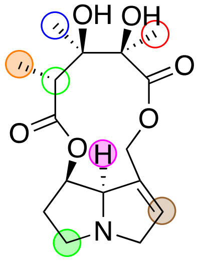
Table A.1.2. Module identity, chemical shifts and compound correlations obtained from the network analysis of interclass mixtures.

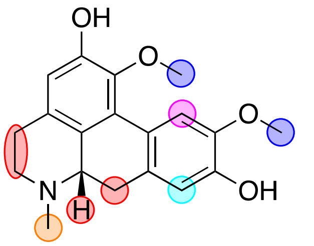
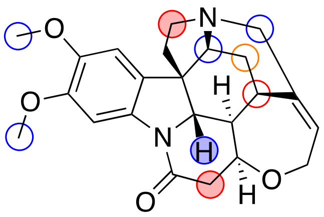
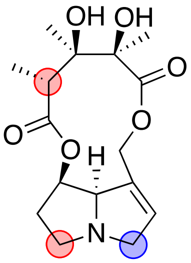
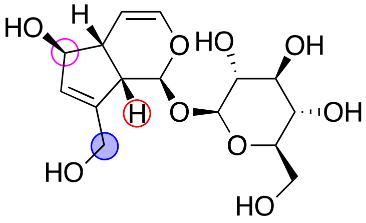
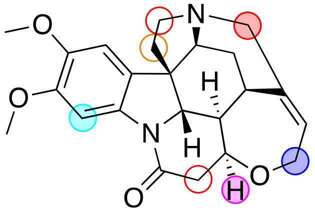
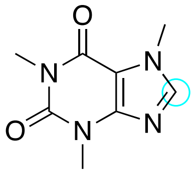
Module (δ)	Compounds (Pearson's correlation)
ROYAL BLUE (PHP-2) **2.14 5.33 9.76 9.80**	**PBA (0.95)** 
TAN (TPN-3) **0.90 0.94 1.42 4.09** **5.41 9.92 9.96 10.0**	**Carene (0.19)** 
	Escin (0.26)  OH HO OH HO O COOH O O OH HO O O O HO OH OH O O O O OH OH OH
	Nerolidol (0.18) 
	Phytol (0.89) 
	Phytenal 
PINK (TPN-2) **1.30 1.54 1.70 2.02** **2.06 2.10 5.13 5.17** **5.21 5.89**	**Carene (0.22)** 
	Phytol (0.24) 

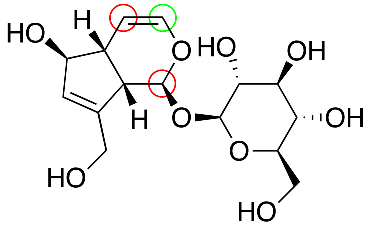
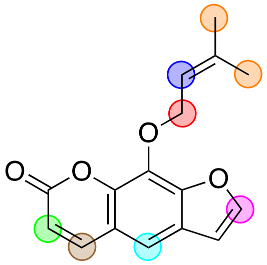
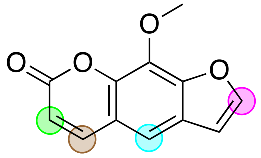
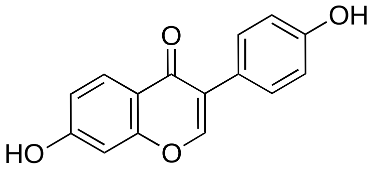
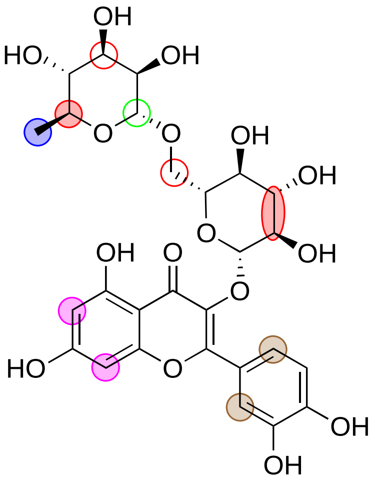
	Phytenal 
	Nerolidol (0.9) 
YELLOW (TPN-1) 0.58 0.62 0.66 0.74 0.78 0.82 1.06 1.62 2.18 2.22 2.34 2.38 5.25	**Carene (0.94)** 
	Nerolidol (0.3) 
	Phytol (0.19) 
MAGENTA (STR-3) 0.70 0.86 0.98 1.02 1.10 1.50 1.58 2.26 5.37	**Diosgenin (0.63)** 
	Escin (0.35) 

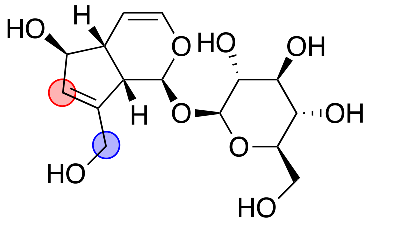
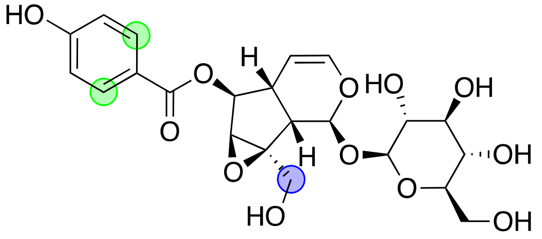
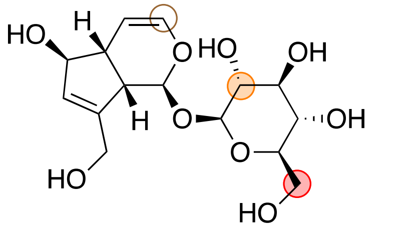
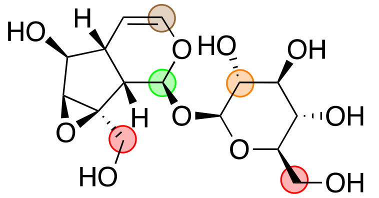
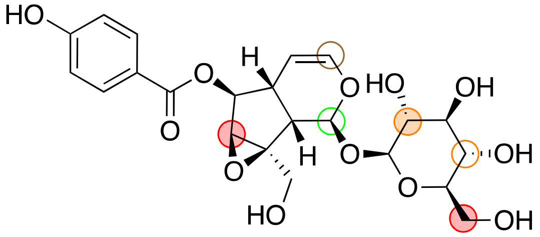
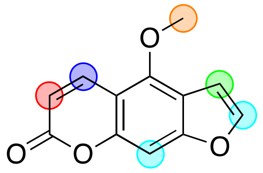
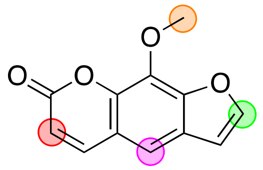
	Oleanic Acid (0.5) 
	Sitosterol (0.35) 
	Stigmasterol (0.21) 
MIDNIGHT BLUE (STR-2) **1.26 1.66 1.78 1.82** **1.86 1.94 1.98**	**Carene (0.33)** 
	Digitoxin (0.33) 

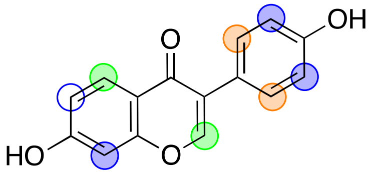
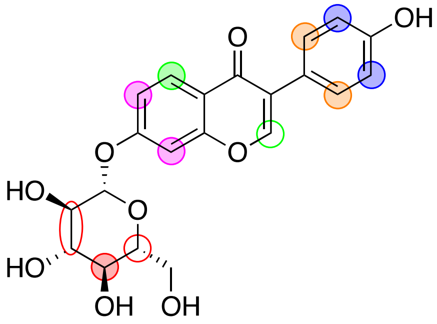
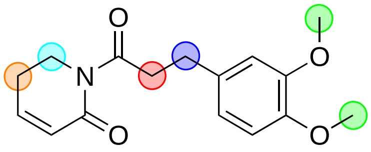
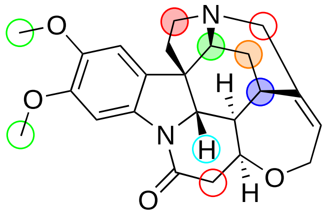
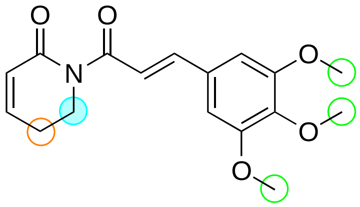
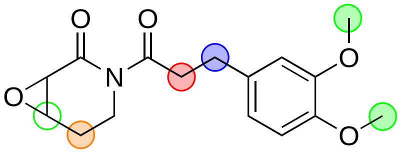
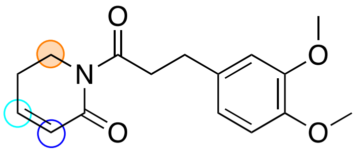
	Diosgenin (0.5) 
	Escin **(0.52)**  OH HO OH HO O COOH O O OH HO O O O HO OH OH O O O O OH OH OH
	Nerolidol (0.15) 
	Oleanic Acid (0.32) 
	PBA (0.22) 

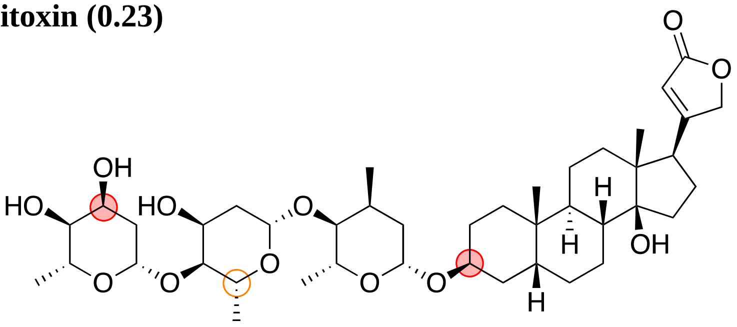
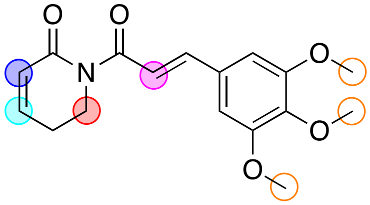
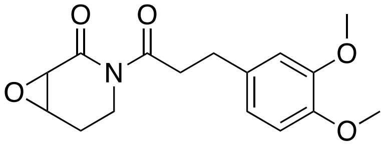
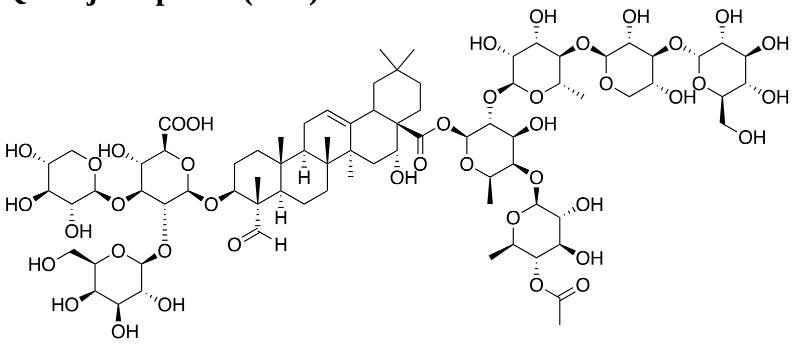
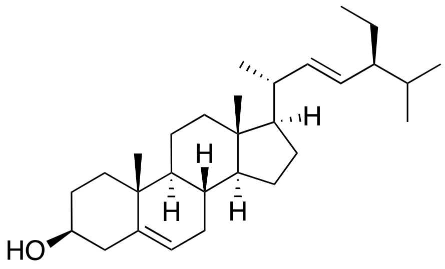
LIGHT YELLOW (STR-1) 5.45 5.49 5.53 5.85	**Escin (0.91)** 
	Quillaja (0.33) 
PURPLE (ALK-1) 1.22 1.38 1.46 2.74 3.18 4.41 6.09 6.13	**Aucubin (0.16)** 
	Crotaline (0.98) 

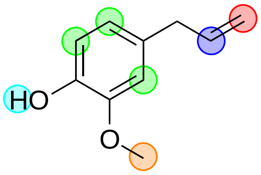
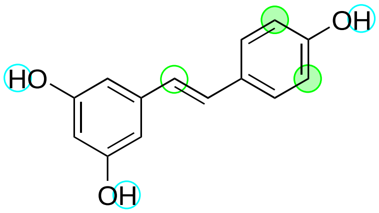
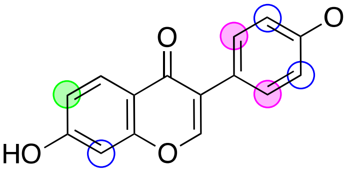
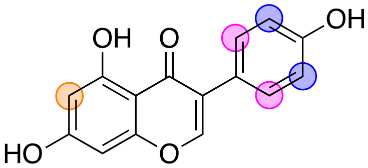
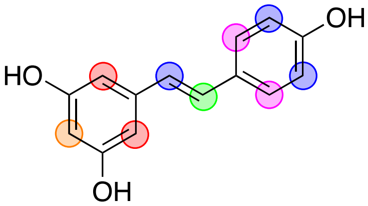
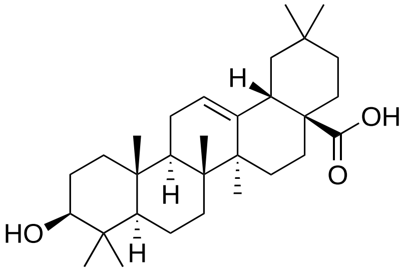
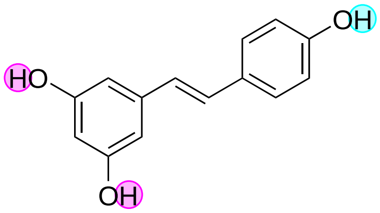
TURQUOISE (ALK-2) **2.46 2.50 2.54 2.58** **2.62 2.66 2.98 3.02** **3.06 3.10 3.14 3.61** **3.89 6.57 8.01** 8.97	**Boldine (0.96)**	
	Brucine (0.29)	
	Crotaline (0.17)	
CYAN (ALK-3) **1.90 2.78 2.82 4.13** **4.17 4.37 7.77**	**Aucubin (0.25)**	
	Brucine (0.9)	
	Caffeine (0.2)	

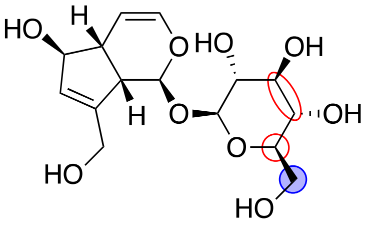
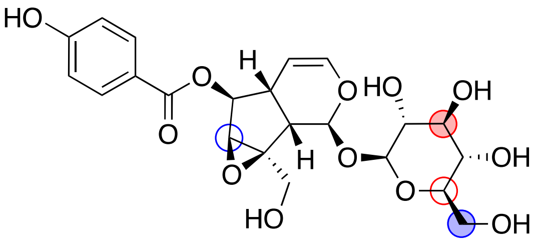
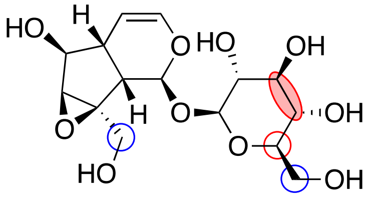
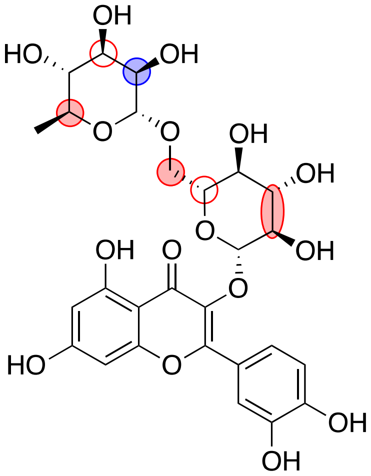
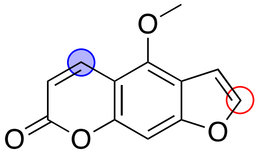
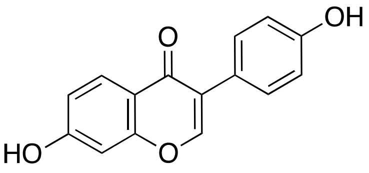
BLACK (FRC-1) 1.74 5.01 5.57 5.61 6.41 7.57 7.61 7.89 7.93 8.05	**Aucubin (0.15)** 
	Imperatorin (0.91) 
	Xanthotoxin (0.15) 
SALMON (FLV-1) 1.14 1.18 3.49 4.57 6.25 6.45 7.65 7.69	**Daidzein (0.16)** 
	Rutin (0.92) 

GREY 60 **(IRG-3)** 2.30 **4.21** **5.73** **5.77** **5.81** **7.97**	**Aucubin (0.77)** 
	Catapolside (0.43) 
GREEN YELLOW **(IRG-2)** **3.26** **3.69** **3.73** **3.77** **5.29** 5.65 **6.33** 8.57	**Aucubin (0.23)** 
	Catalpol (0.88) 
	Catapolside (0.34) 
BLUE (FRC-2) **4.25** **4.29** **4.33** **6.29** **6.37** **7.13** **7.17** **7.21** **7.53** **7.73** **7.81** **7.85** **8.25** 8.49 9.40	**Bergapten (0.96)** 
	Xanthotoxin (0.77) 

BROWN **(FLV-2)** 2.70 **3.53** 4.45 5.69 **6.89 6.93 7.29 7.45** **7.49 8.09 8.13 8.17** **8.21** 8.37 8.45	**Daidzein (0.8)** 
	Daidzin (0.55) 
LIGHT CYAN **(AMD-1)** **2.42 2.86 2.90 2.94** **3.22 3.81 3.97**	**Alkene Amide (0.77)** 
	Brucine (0.19) 
	Piplartine (0.26) 
	Pipleroxide (0.53) 
LIGHT GREEN **(AMD-2)** **3.93 4.05 6.05 7.09** **7.33**	**Alkene Amide (0.18)** 

	Digitoxin (0.23) 
	Piplartine (0.9) 
	Pipleroxide (0.38) 
	Quillaja Saponin (0.19) 
	Stigmasterol (0.16) 

GREEN (PHP-3) 3.85 5.05 5.09 5.93 5.97 6.01 6.61 6.65 6.69 6.73 6.77 9.60	**Eugenol (0.91)** 
	Resveratrol (0.25) 
RED (PHP-1) 4.53 6.17 6.21 6.49 6.53 6.81 6.85 6.97 7.01 7.05 7.37 7.41	**Daidzein (0.18)** 
	Genistein (0.23) 
	Resveratrol (0.93) 
DARK RED (PHP-4) 9.13 9.16 9.32	**Oleanic Acid (0.17)** 
	Resveratrol (0.34) 

DARK TURQUOISE (GLC-1) **3.41 3.45 3.65**	**Aucubin (0.31)** 
	Catapolside (0.4) 
	Catalpol (0.24) 
	Rutin (0.6) 
DARK GREEN (FLV-3) **7.25 8.29** 8.33	**Bergapten (0.96)** 
	Daidzein (0.39) 

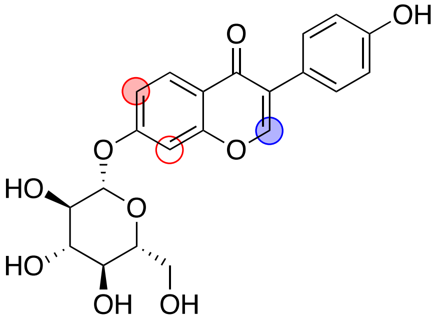
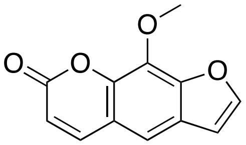
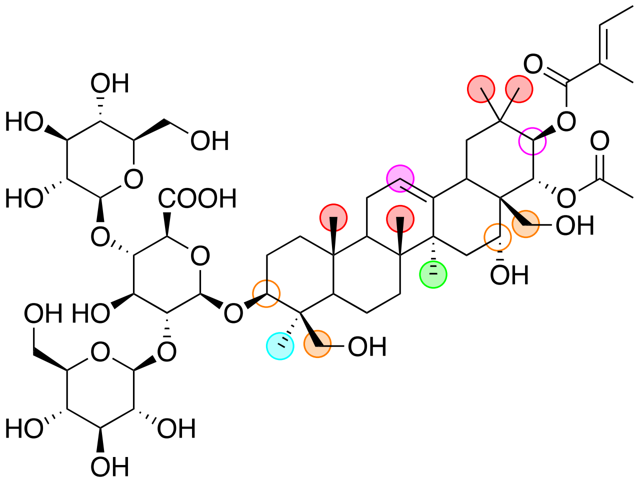
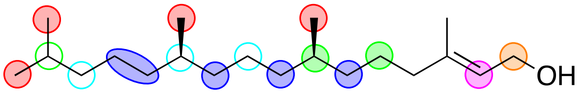
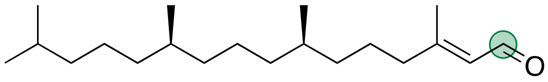
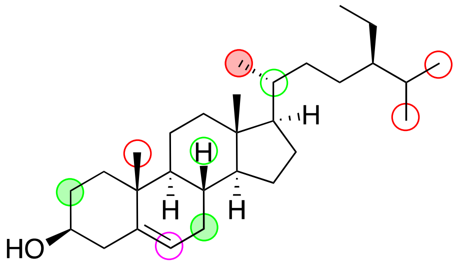
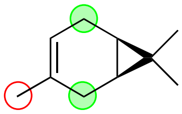
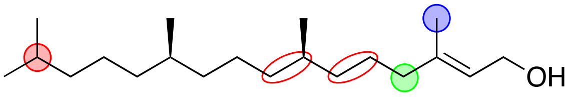
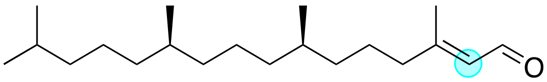
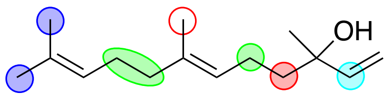
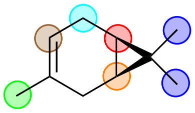
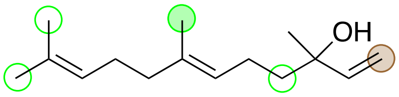
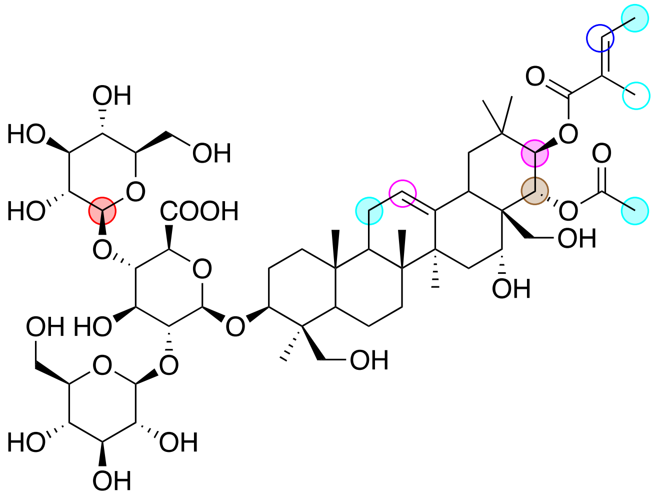
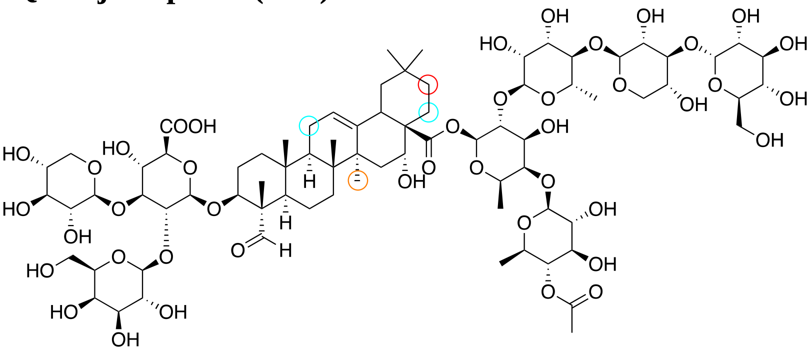
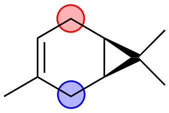
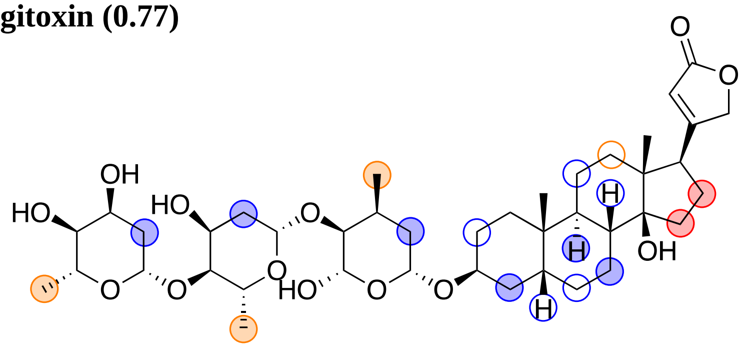
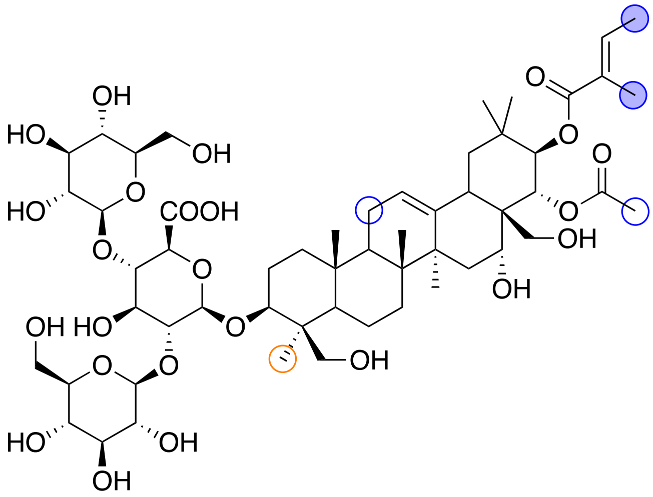
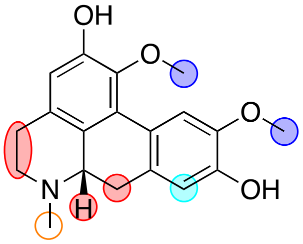
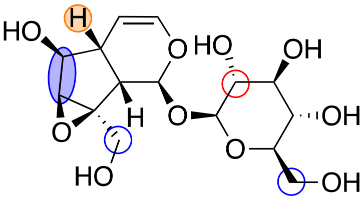
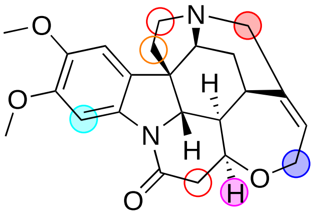
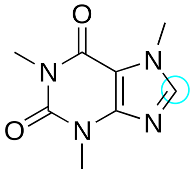
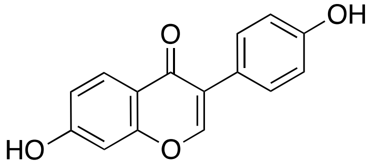
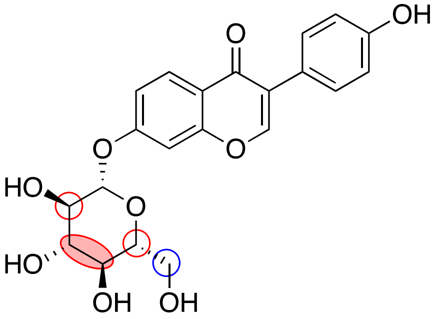
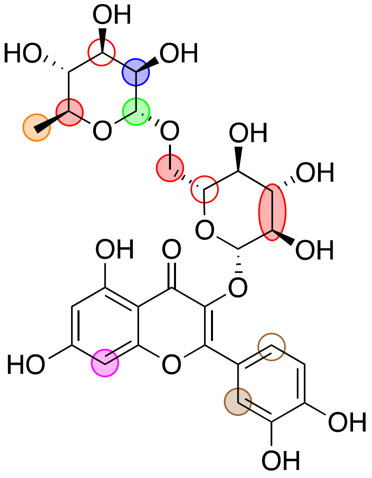
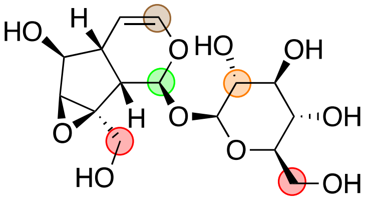
	Daidzin (0.55) 
	Xanthotoxin (0.4) 

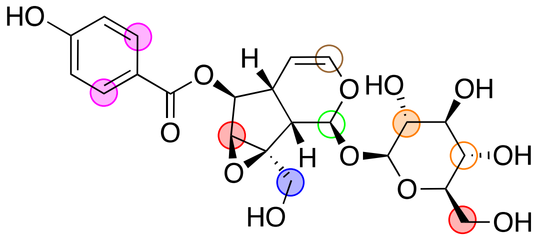
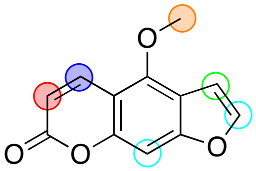
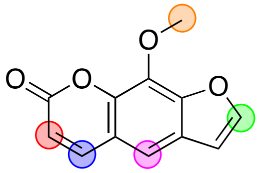
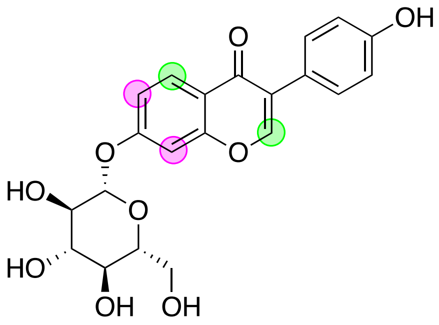
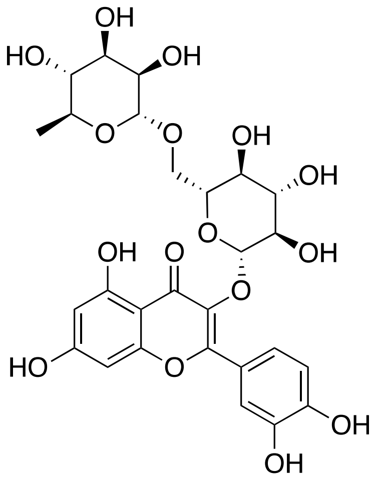
Table A.1.3. Module identity, chemical shifts and compound correlations obtained from the network analysis of complex mixtures.

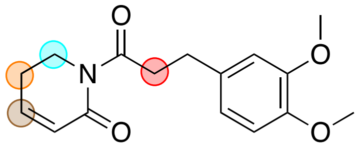
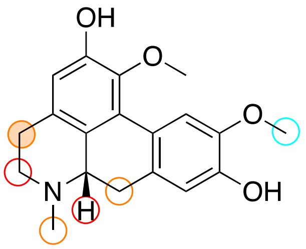
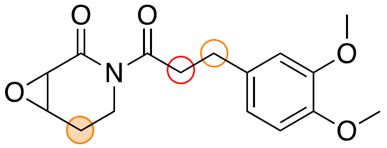
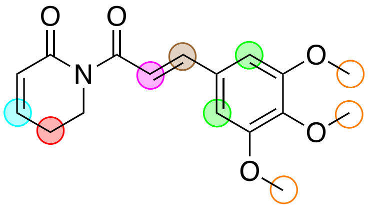
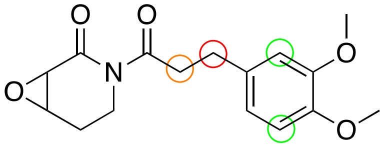
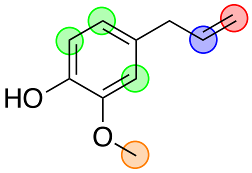
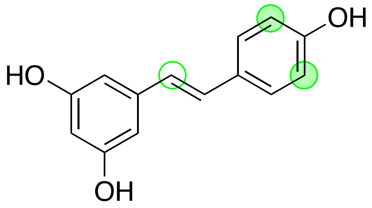
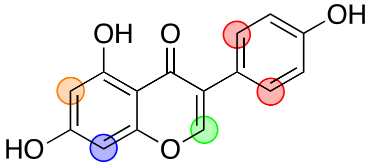
Module (δ)	Compounds (Pearson's correlation)
GREEN (TPN-3) 0.90 0.94 0.98 1.10 1.22 1.42 1.46 1.50 4.09 5.37 5.41 9.96	**Escin (0.31)** 
	Phytol (0.76) 
	Phytenal (N/A) 
	Sitosterol (0.25) 
PURPLE (TPN-2) 1.54 1.58 1.66 1.70 2.02 2.06 2.10 2.14 5.89	**Carene (0.23)** 
	Phytol (0.32) 
	Phytenal 

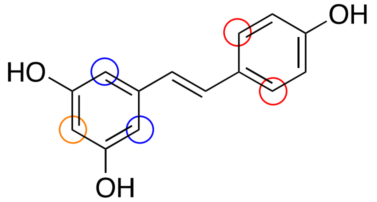
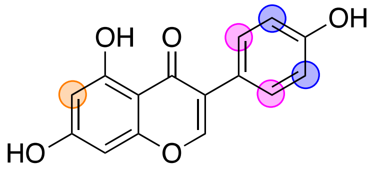
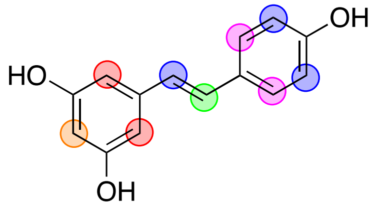
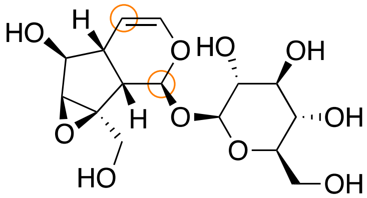
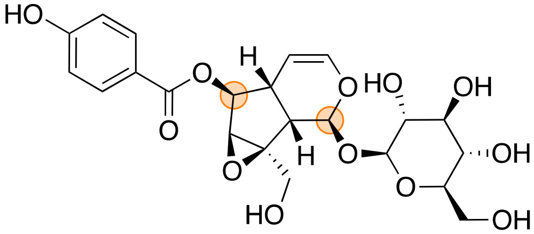
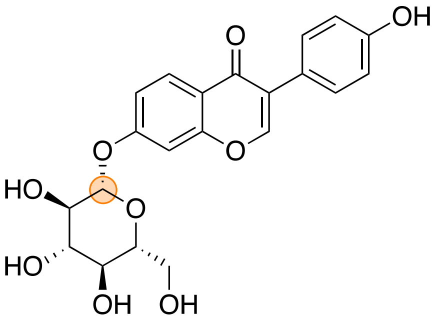
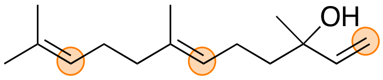
	Nerolidol (0.93) 
PINK (TPN-1) 0.62 0.66 0.78 0.82 1.06 1.62 2.34 2.38 5.25	Carene (0.95) 
	Nerolidol (0.25) 
BLACK (STR-1) 1.94 1.98 4.45 5.45 5.49 5.53 5.77 5.81 5.85 6.13	Escin (0.88) 
	Quillaja Saponin (0.55) 
MIDNIGHT BLUE (STR-2) 1.26 1.74 1.78 1.82 1.86 2.22	Carene (0.47) 

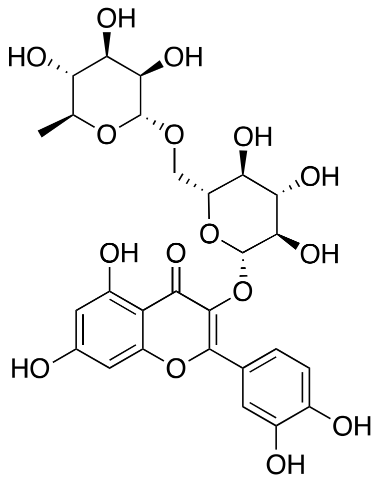
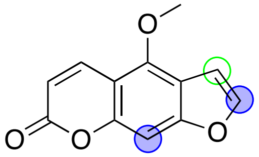
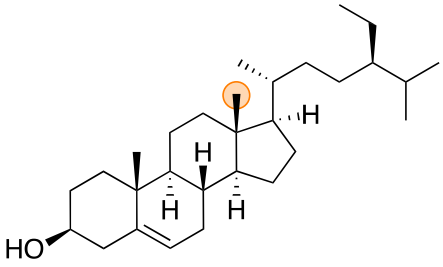
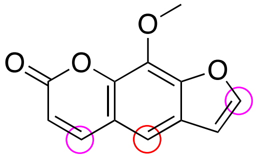
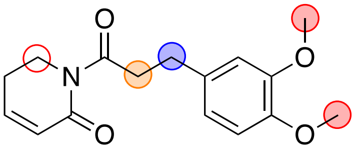
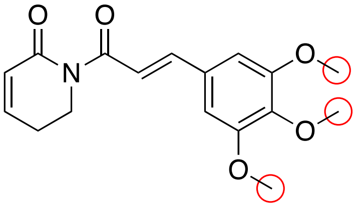
	Digitoxin (0.77) 
	Escin (0.28) 
BROWN (ALK-2) 2.50 2.58 2.62 2.66 2.70 3.02 3.06 3.10 3.14 3.61 3.89 6.57	**Boldine (0.95)** 
	Catalpol (0.25) 
TAN (ALK-3) 1.90 2.74 2.78 2.82 4.17 4.37 7.77	**Brucine (0.97)** 

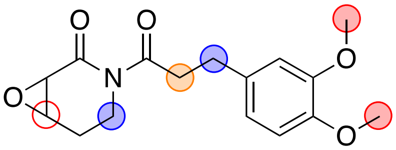
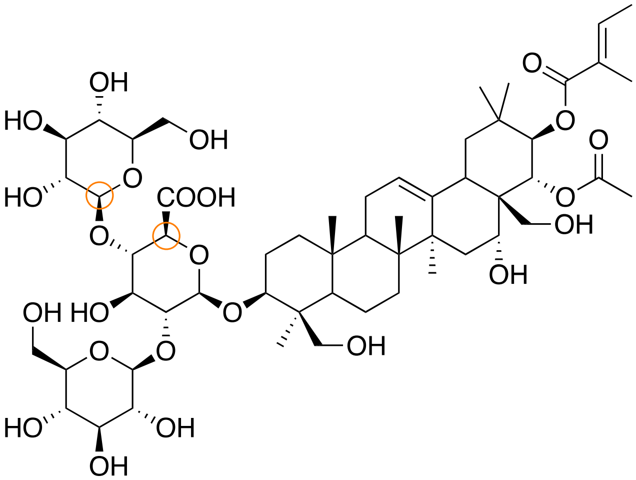
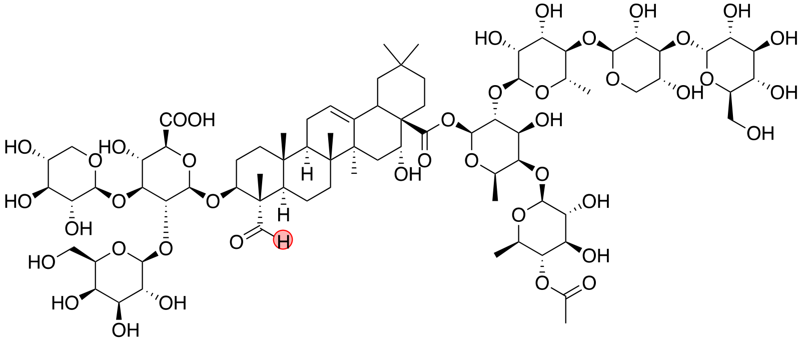
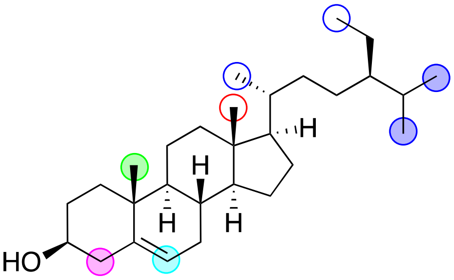
	Caffeine (0.43) 
RED (FLV-1) **1.14 1.18 3.41 3.45** **3.49 3.65 4.53 4.57** **6.45 7.69**	**Daidzein (0.59)** 
	Daidzin (0.5) 
	Rutin (0.89) 
YELLOW (IRG-1) **3.26 3.69 3.73 3.77** **4.21 5.29** 5.57 5.61 5.65 **6.33 7.93** 8.57	**Catalpol (0.8)** 

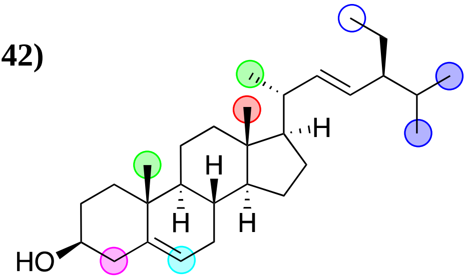
	Catapolside (0.83) 
TURQUOISE **(FRC-2)** **4.29 4.33 6.29** **6.41 7.17 7.21 7.25** **7.53 7.81 7.85 8.01** **8.25** 8.29 8.33	**Bergapten (0.98)** 
	Xanthotoxin (0.84) 
LIGHT YELLOW **(FLV-3)** **7.29 8.17 8.21**	**Daidzin (0.73)** 
	Rutin (0.49) 

GREY 60 (AMD-1) 2.42 2.46 2.94 3.97 7.05	**Alkene Amide (0.8)** 
	Boldine (0.22) 
	Pipleroxide (0.22) 
GREEN YELLOW (AMD-2) 2.54 3.93 6.05 6.93 7.09 7.33 7.65	**Piplartine (0.82)** 
	Pipleroxide (0.4) 
BLUE (PHP-3) 3.85 5.01 5.05 5.09 5.93 5.97 6.01 6.61 6.65 6.69 6.73 6.77	**Eugenol (0.89)** 
	Resveratrol (0.28) 
CYAN (FLV-2) 6.25 6.37 7.41 7.45 8.09 8.13	**Genistein (0.88)** 

	Resveratrol (0.35) 
MAGENTA (PHP-1) 6.17 6.21 6.49 6.53 6.81 6.85 6.97 7.01 7.37	**Genistein (0.35)** 
	Resveratrol (0.91) 
DARK RED (TPN-2) 5.13 5.17 5.21	**Catalpol (0.23)** 
	Catapolside (0.32) 
	Daidzin (0.23) 
	Nerolidol (0.76) 

	Rutin (0.26) 
LIGHT CYAN **(FRC-2)** **0.70 7.13 7.49** **7.73 7.97** 9.40	**Bergapten (0.47)** 
	Sitosterol (0.26) 
	Xanthotoxin (0.38) 
LIGHT GREEN **(AMD-1)** **2.86 2.90 3.18 3.22** **3.81**	**Alkene Amide (0.53)** 
	Piplartine (0.26) 

	Pipleroxide (0.69) 
ROYAL BLUE (STR-1) **4.49** **9.48** 9.56	**Escin (0.52)** 
	Quillaja Saponin (0.88) 
SALMON (STR-3) 0.58 **0.74** **0.86** **1.02** **2.26** **2.30** **5.33**	**Sitosterol (0.76)** 

	Stigmasterol (0.42) 

A.2. *S. cereviseae* inhibition assays

Table A.2.1. Dose-response inhibition data for compound 1.

Concentration (μM)	AUC 1*	AUC 2*	AUC 3*
100.0	0.381	0.367	0.331
50.0	0.877	0.891	0.616
25.0	0.897	0.879	0.905
12.5	0.942	0.953	0.881
3.25	0.978	0.973	0.893
1.56	0.974	0.962	0.879

* AUC: area under growth curve normalized by total area of the curve generated from control sample (YPD + yeast + acetone). Numerals 1, 2 and 3 indicate assay trials.

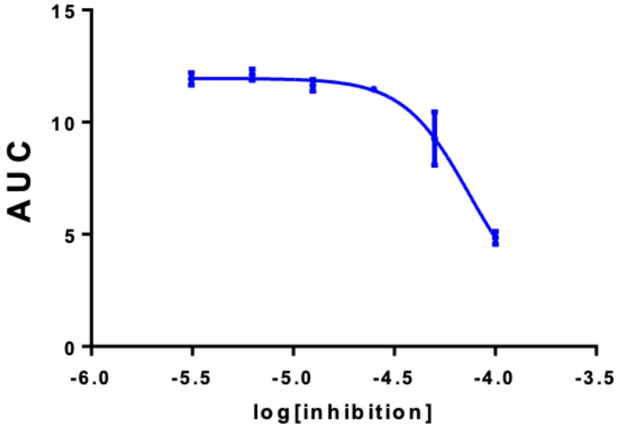


Figure A.2.1. Log(C) vs. inhibition response for compound **1.**

Table A.2.2. Dose-response inhibition data for compound 2.

Concentration (μM)	AUC 1*	AUC 2*	AUC 3*
100.0	0.326	0.203	0.137
50.0	0.211	0.217	0.199
25.0	0.213	0.224	0.199
12.5	0.213	0.227	0.241
3.25	0.552	0.210	0.467
1.56	0.966	0.963	0.826

* AUC: area under growth curve normalized by total area of the curve generated from control sample (YPD + yeast + acetone). Numerals 1, 2 and 3 indicate assay trials.

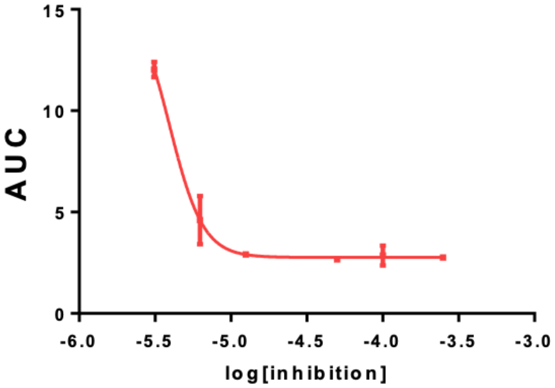


Figure A.2.2. Log(C) vs. inhibition response for compound **2.**

A.3. ^{1}H and ^{13}C NMR spectra

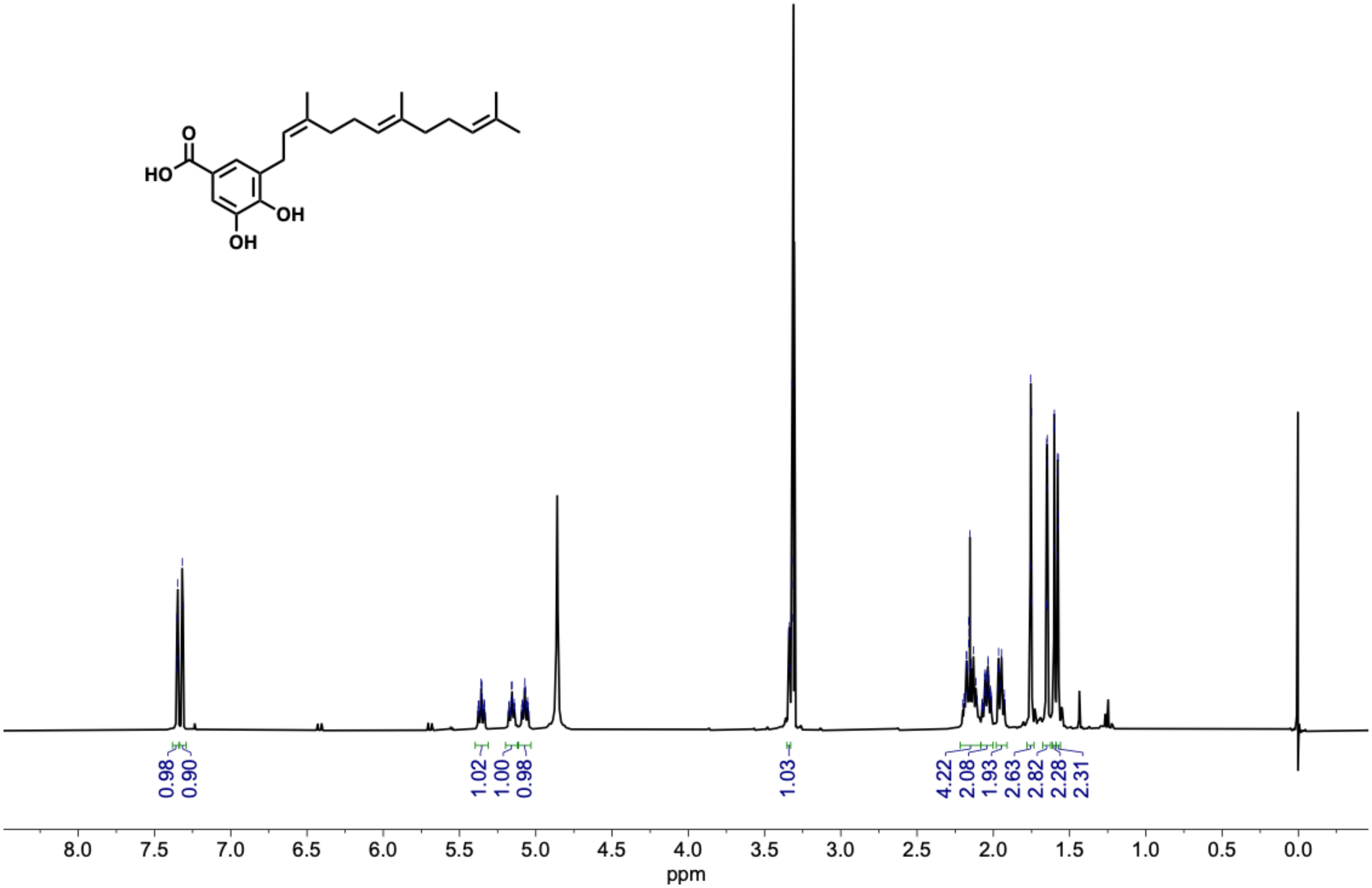


Figure A.3.1. ^{1}H NMR (400 MHz, CD$_3$OD) spectrum of compound **1**.

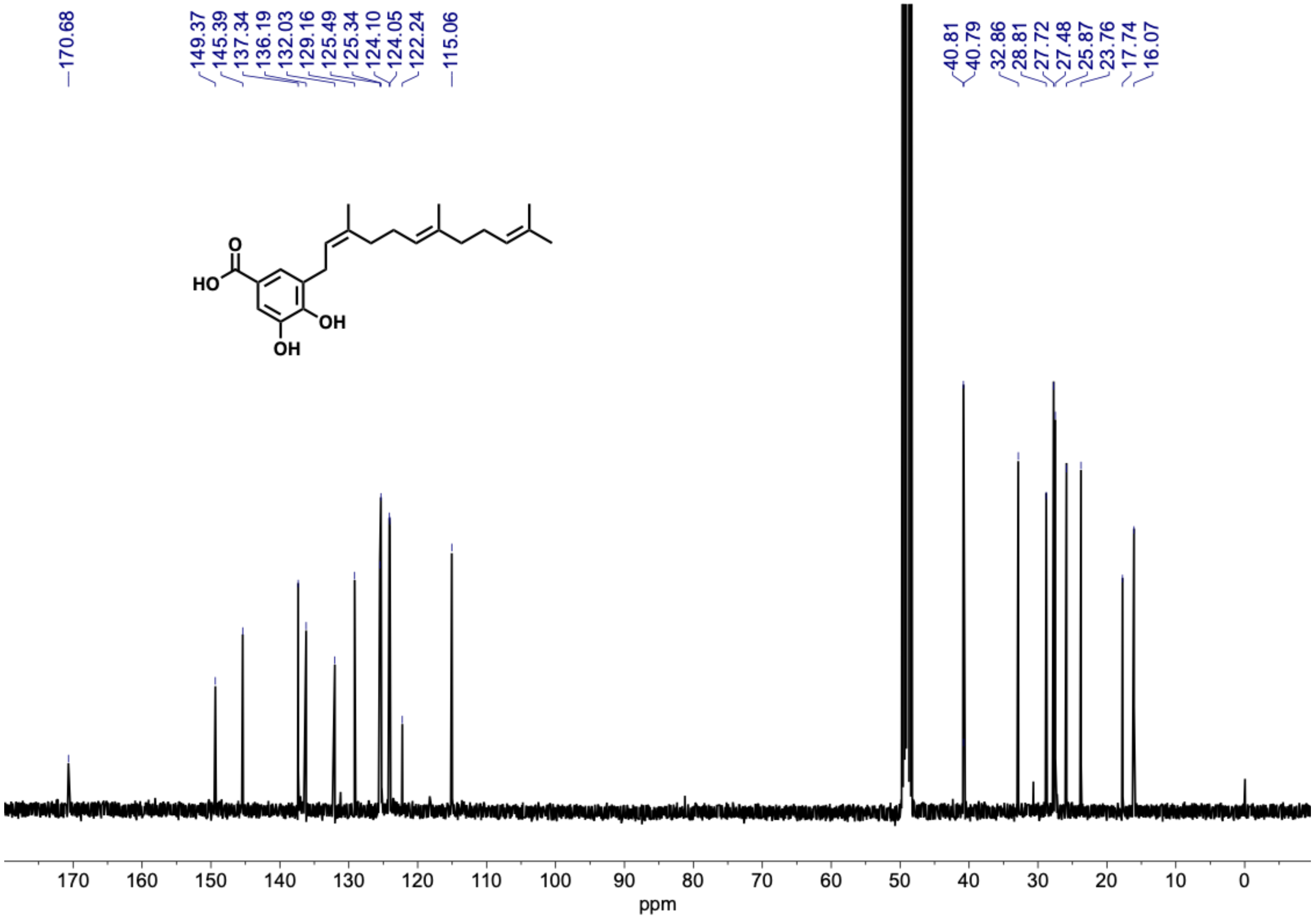


Figure A.3.2. ^{13}C NMR (101 MHz, CD_3OD) spectrum of compound **1**.

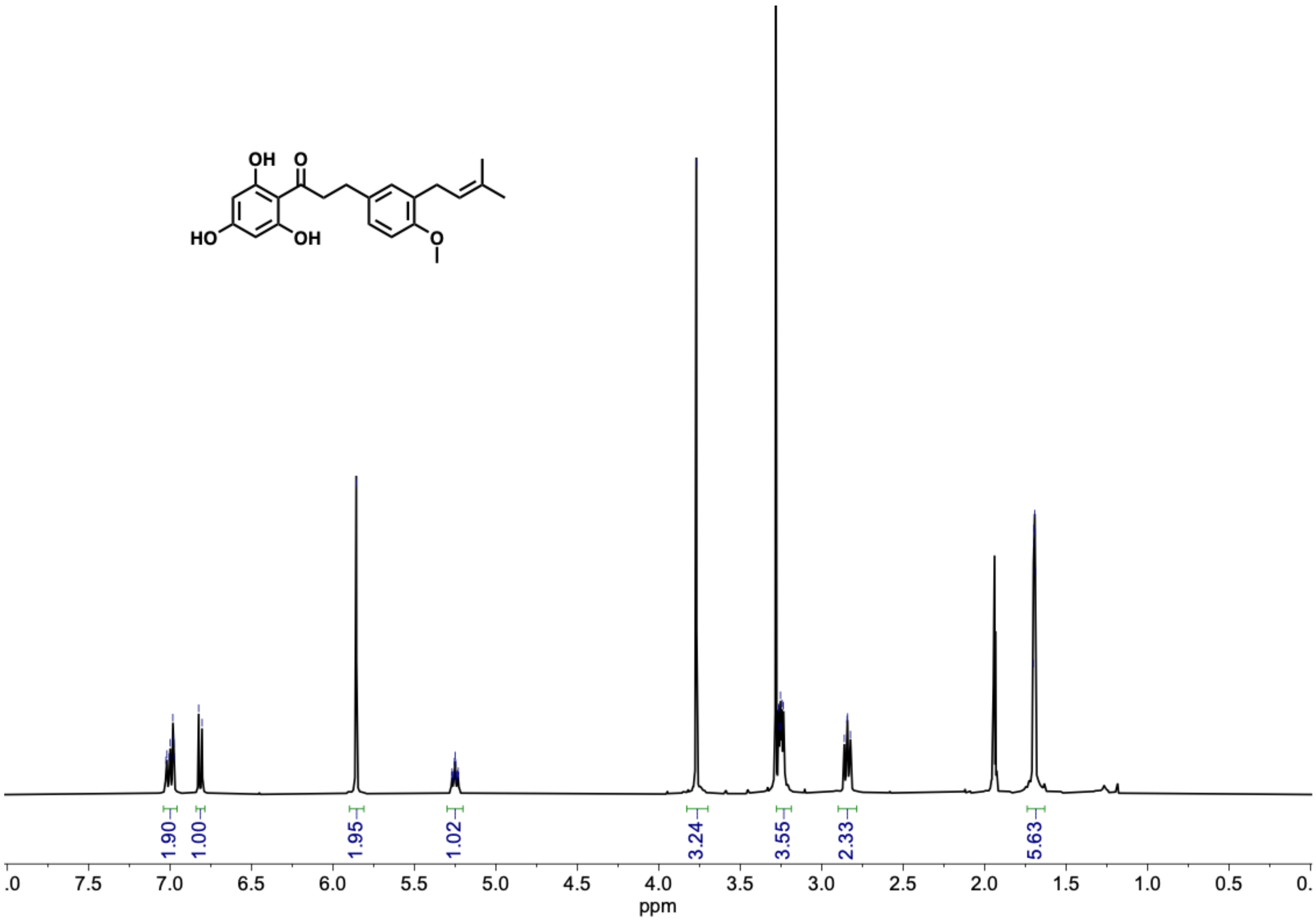


Figure A.3.3. ^{1}H NMR (400 MHz, CD$_3$CN) spectrum of compound **2**.

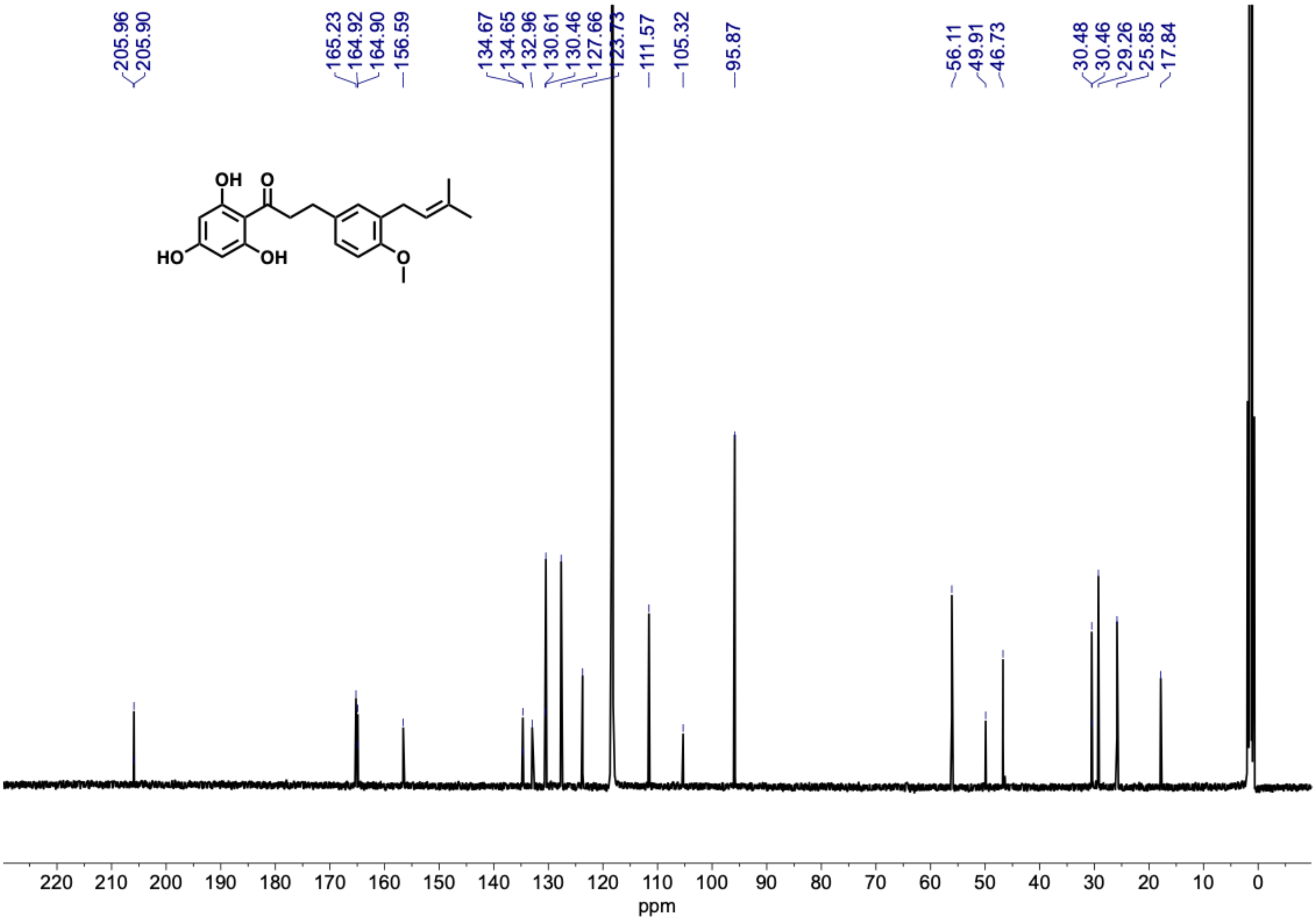


Figure A.3.4. ^{13}C NMR (101 MHz, CD_3CN) spectrum of compound **2**.

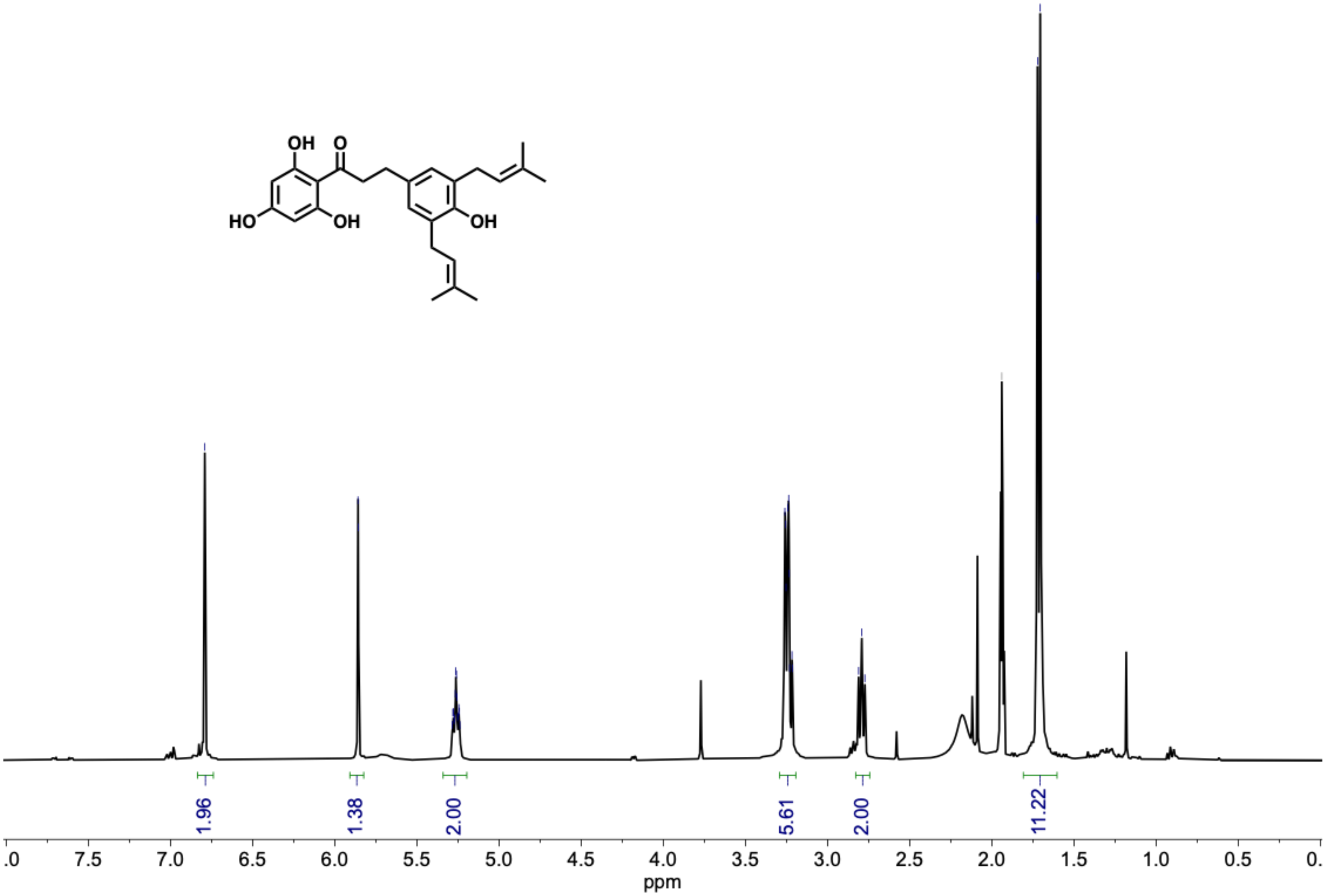


Figure A.3.5. ^{1}H NMR (400 MHz, CD_3CN) spectrum of compound **3**.

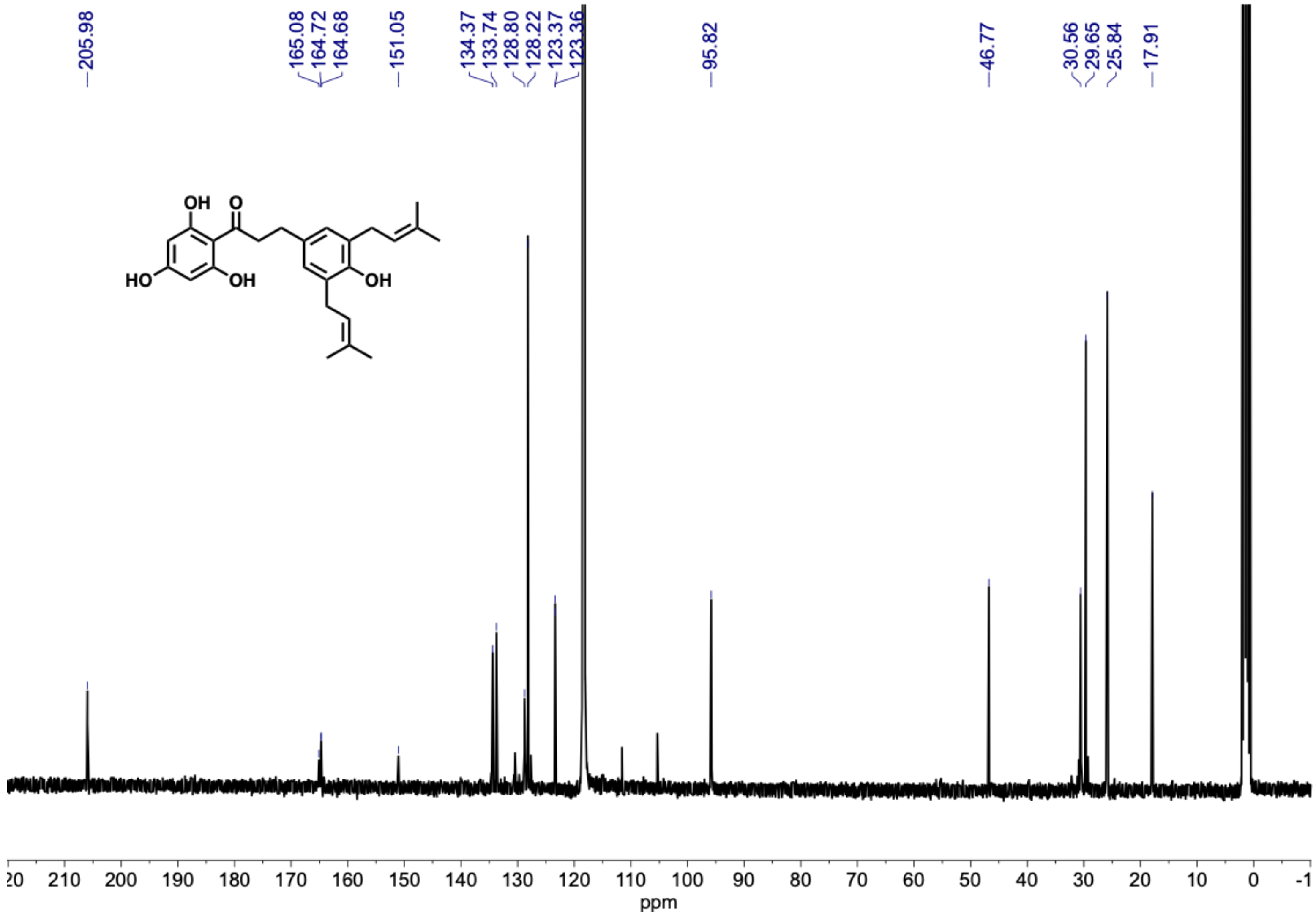


Figure A.3.6. ^{13}C NMR (101 MHz, CD$_3$CN) spectrum of compound **3**.

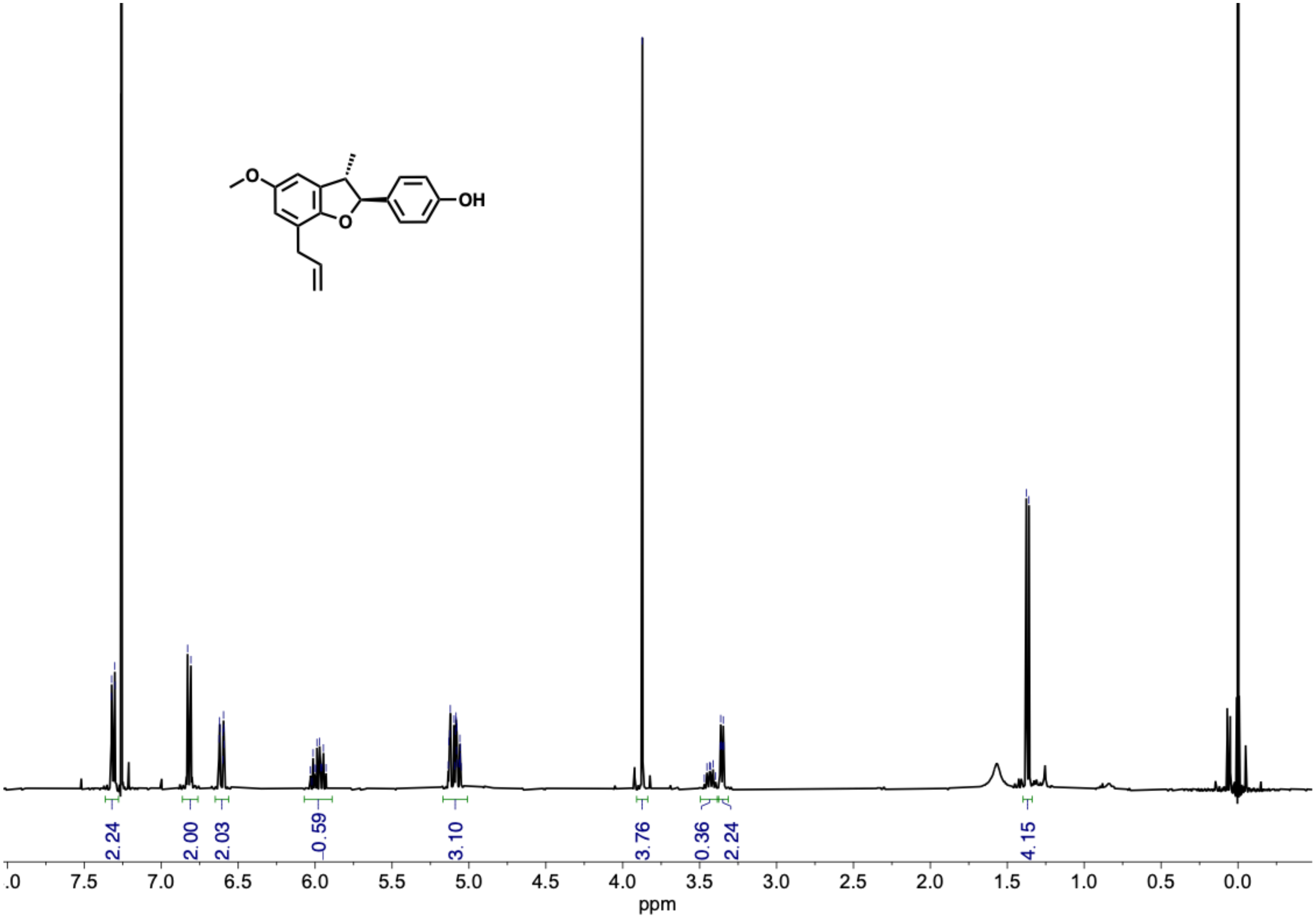


Figure A.3.7. ^{1}H NMR (400 MHz, $CDCl_3$) spectrum of compound **4**.

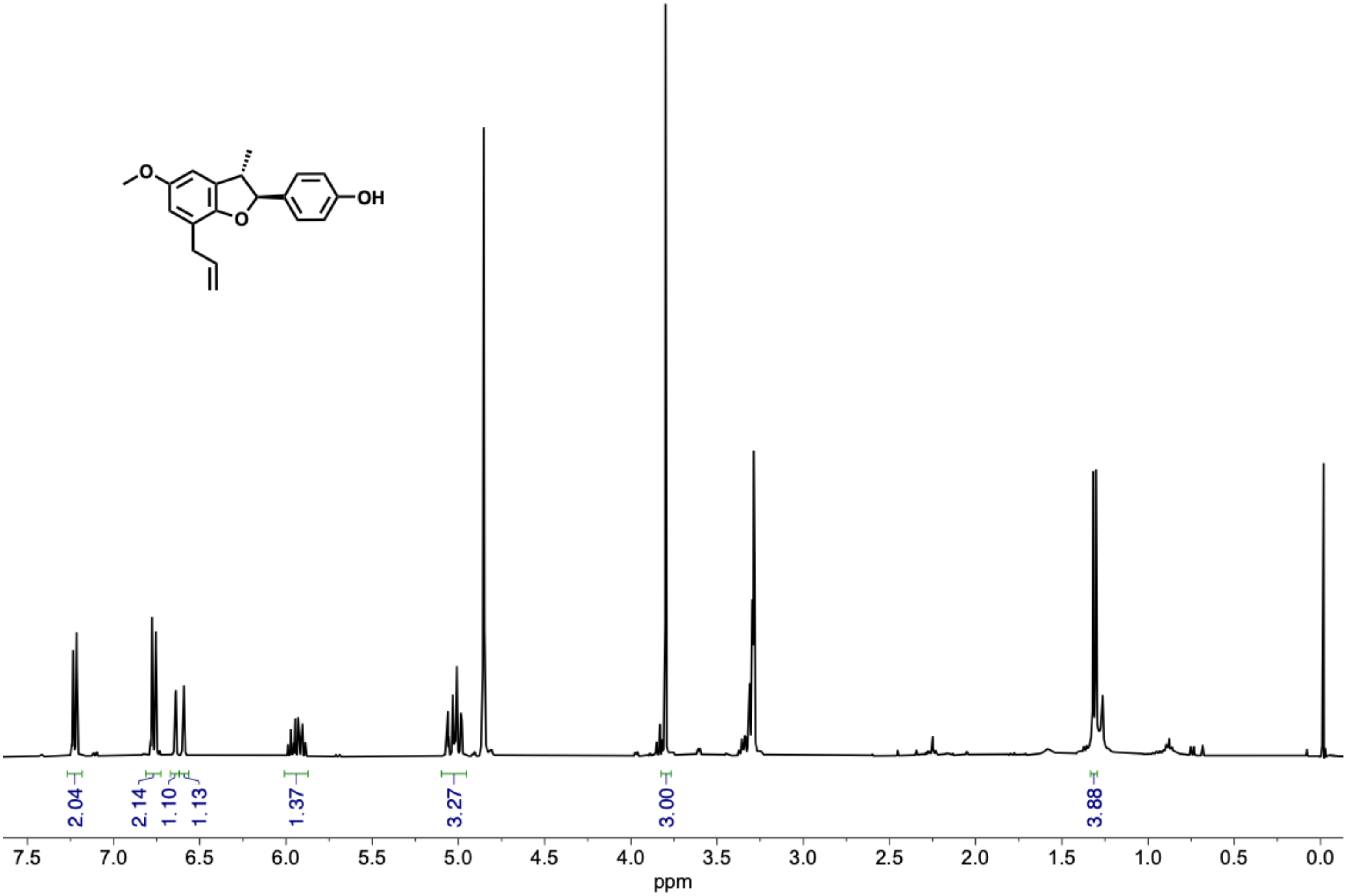


Figure A.3.8. ^{1}H NMR (400 MHz, CD_3OD) spectrum of compound **4**.

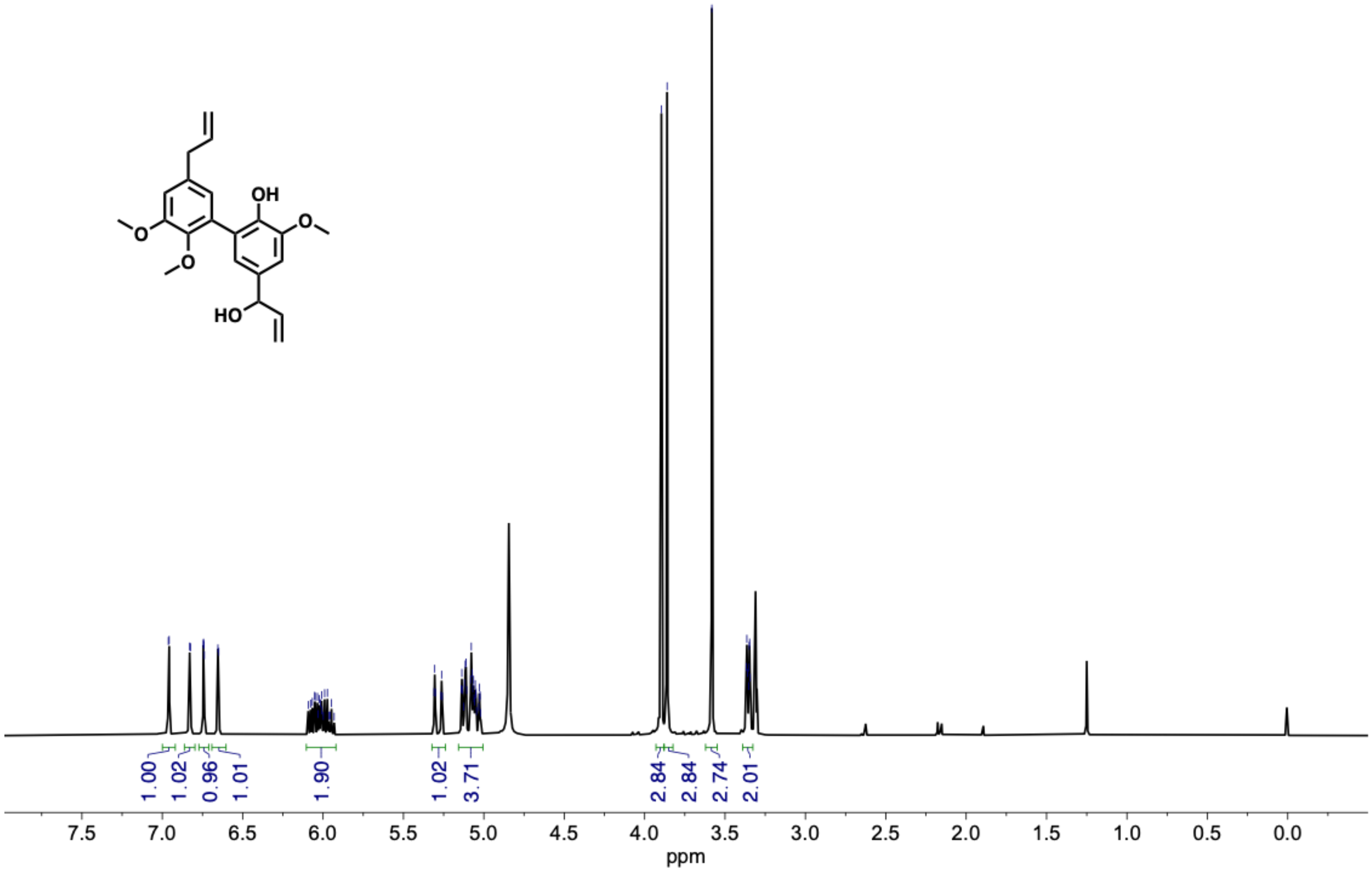


Figure A.3.9. ^1H NMR (400 MHz, CD_3OD) spectrum of compound **5**.

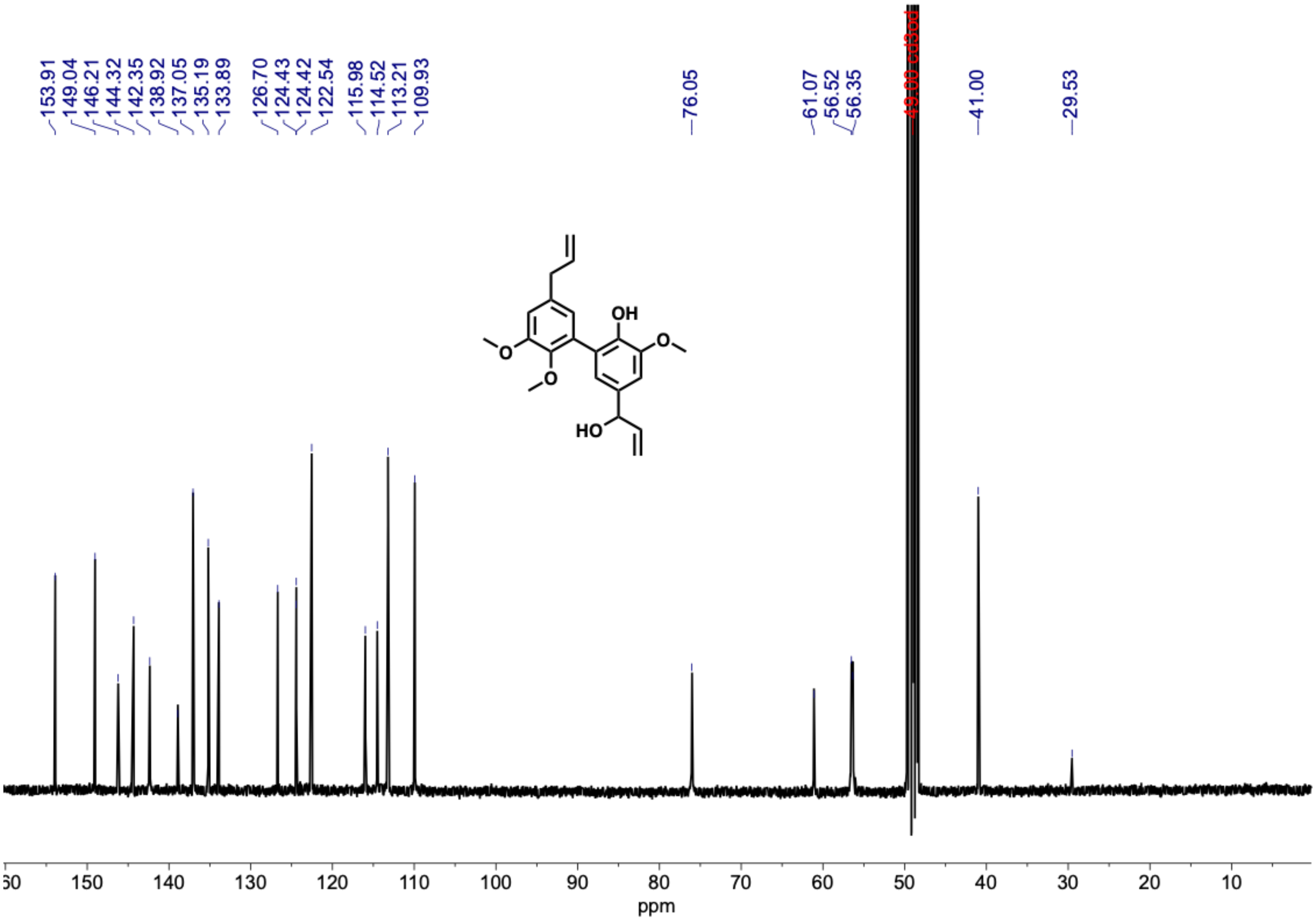


Figure A.3.10. ^{13}C NMR (101 MHz, CD_3OD) spectrum of compound **5**.

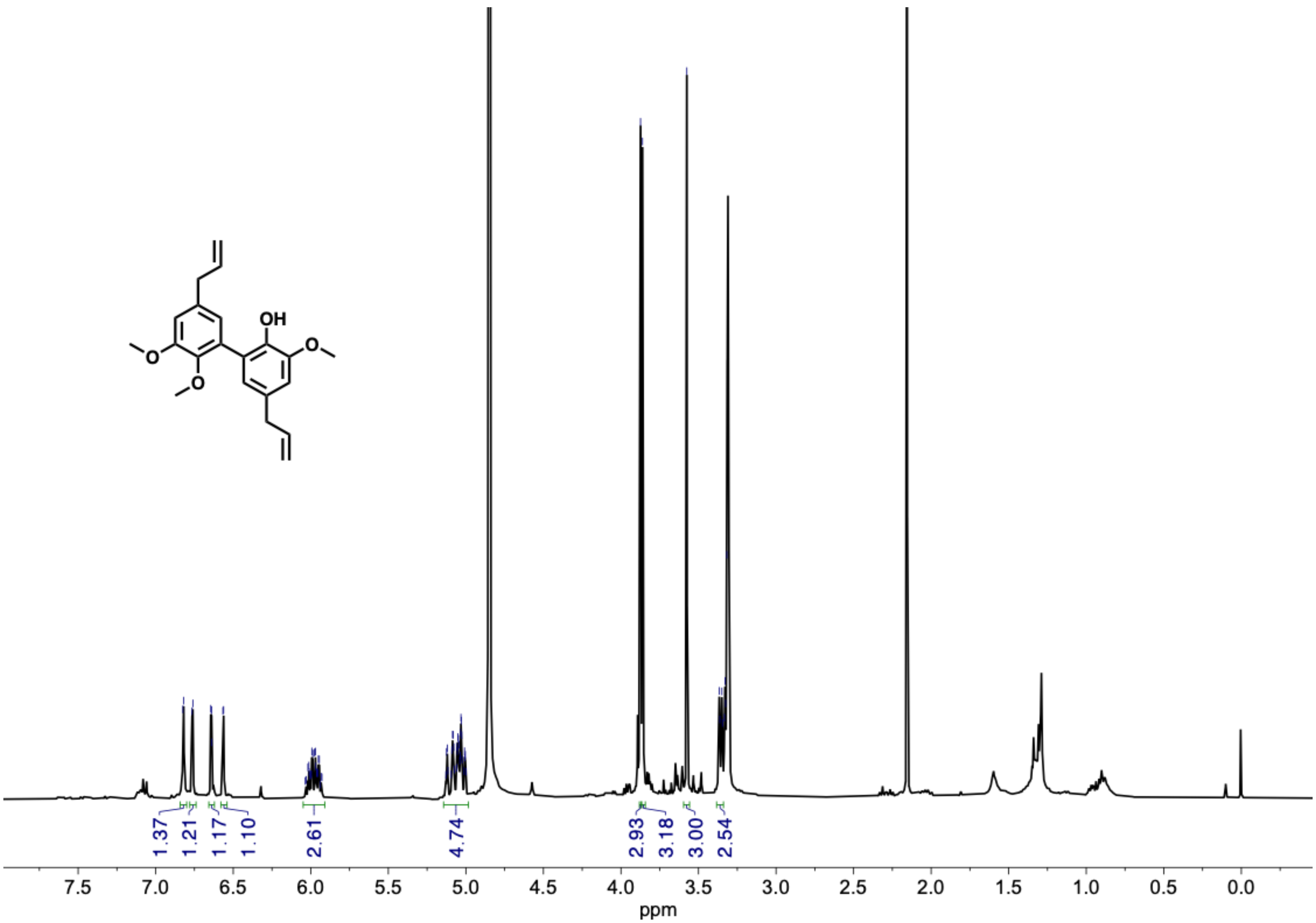


Figure A.3.11. ^{1}H NMR (400 MHz, CD_3OD) spectrum of compound **6**.

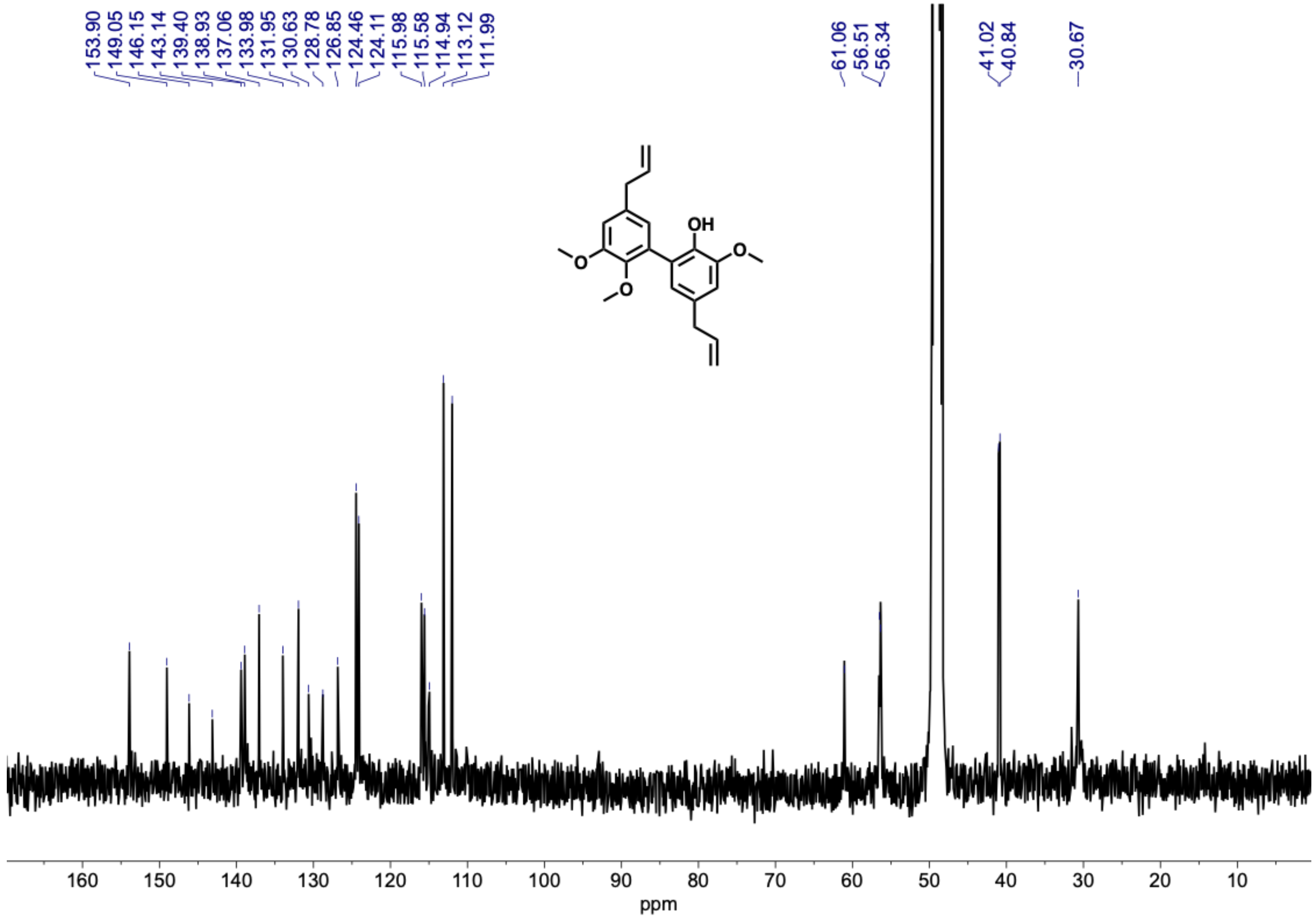


Figure A.3.12. ^{13}C NMR (101 MHz, CD_3OD) spectrum of compound **6**.

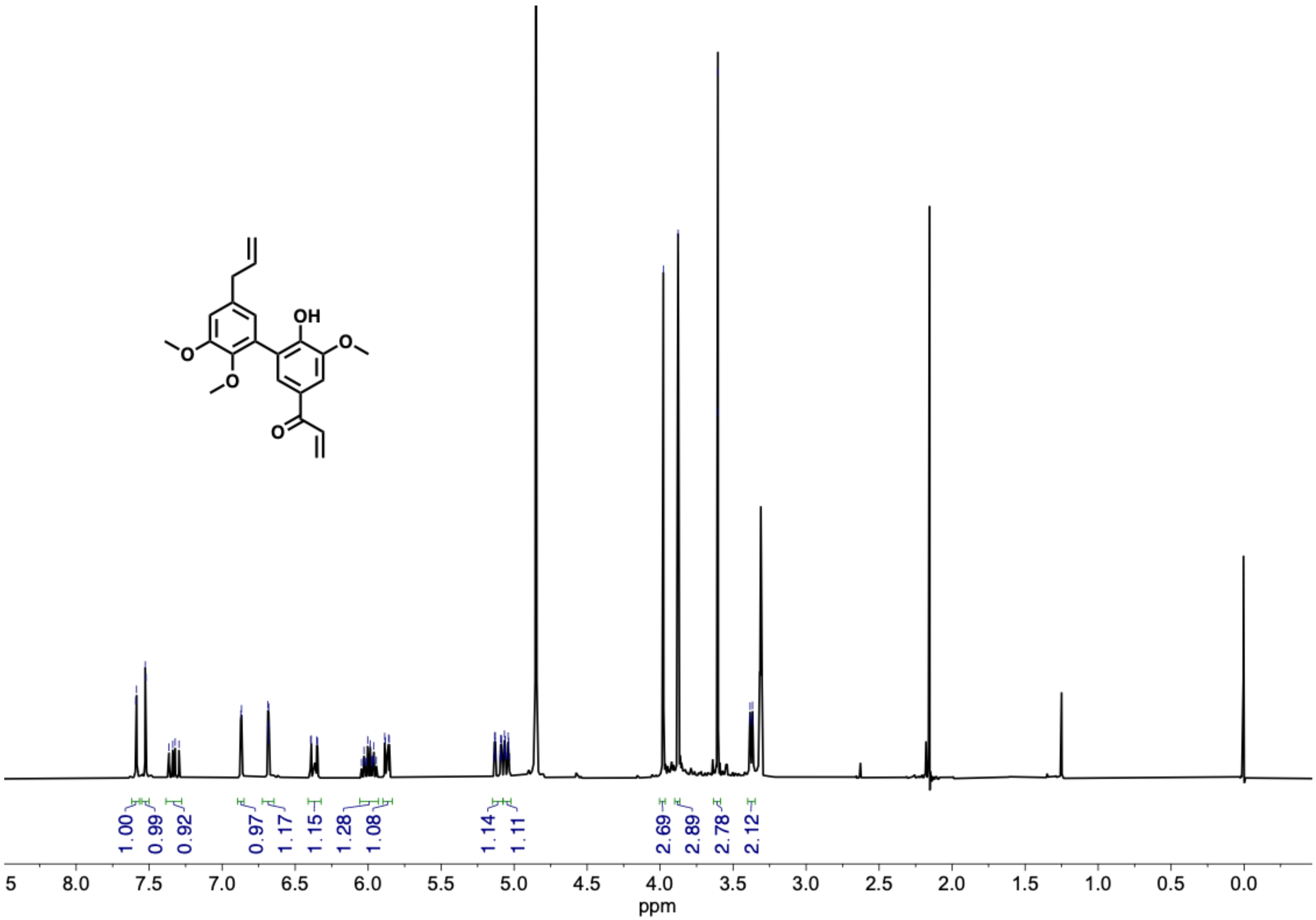


Figure A.3.13. ^{1}H NMR (400 MHz, CD_3OD) spectrum of compound **7**.

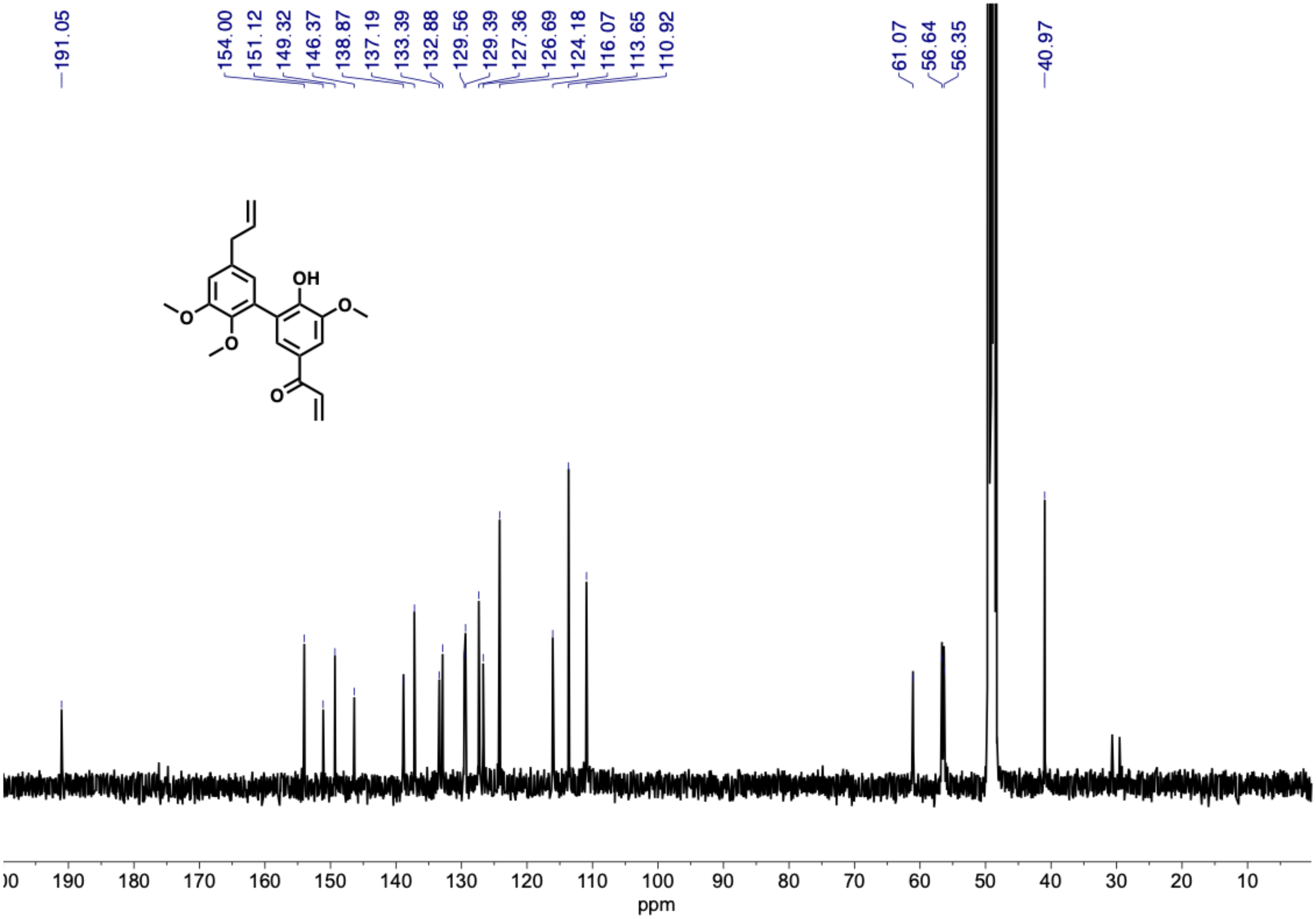


Figure A.3.14. ^{13}C NMR (101 MHz, CD_3OD) spectrum of compound **7**.

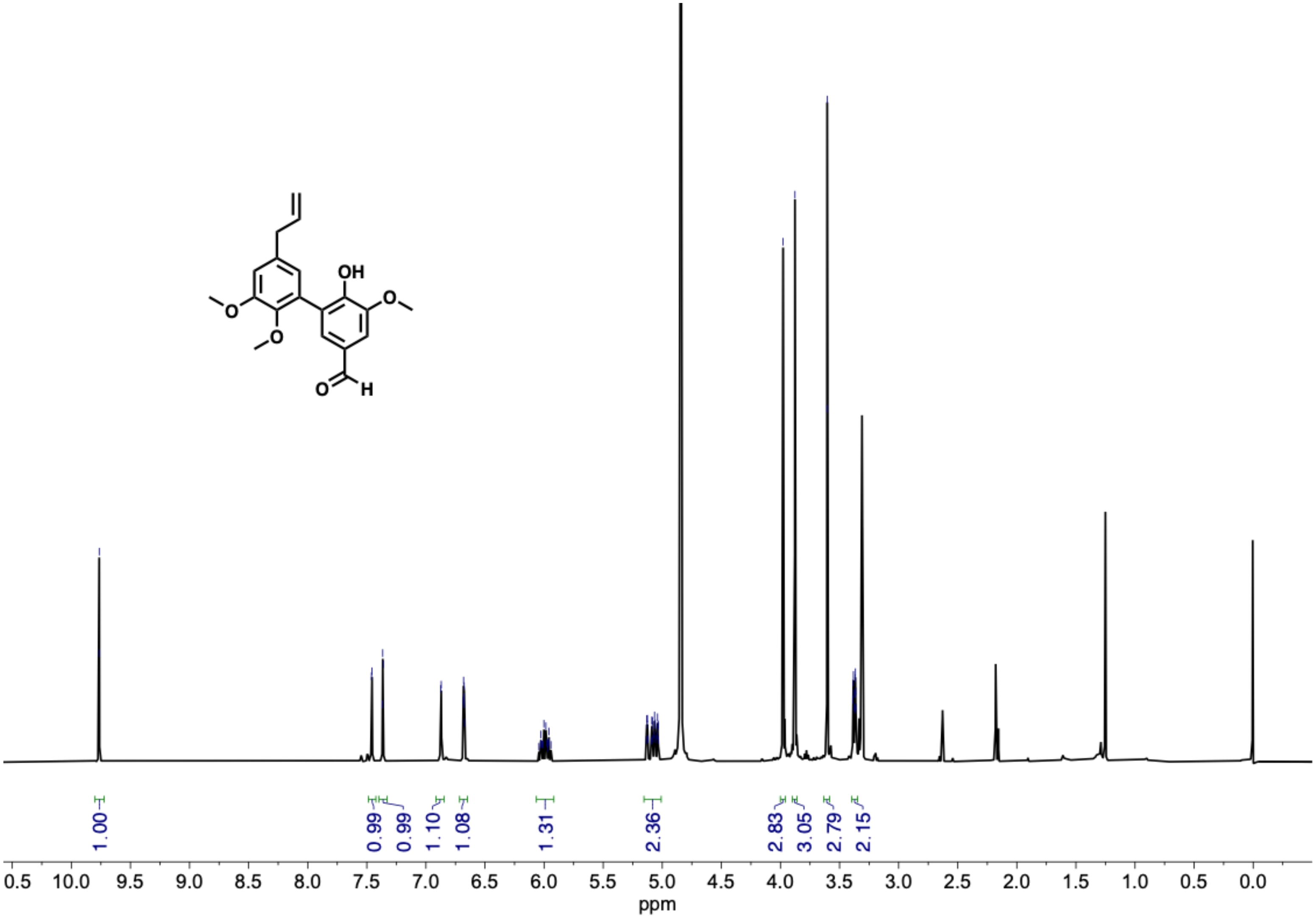


Figure A.3.15. ^{1}H NMR (400 MHz, CD_3OD) spectrum of compound **8**.

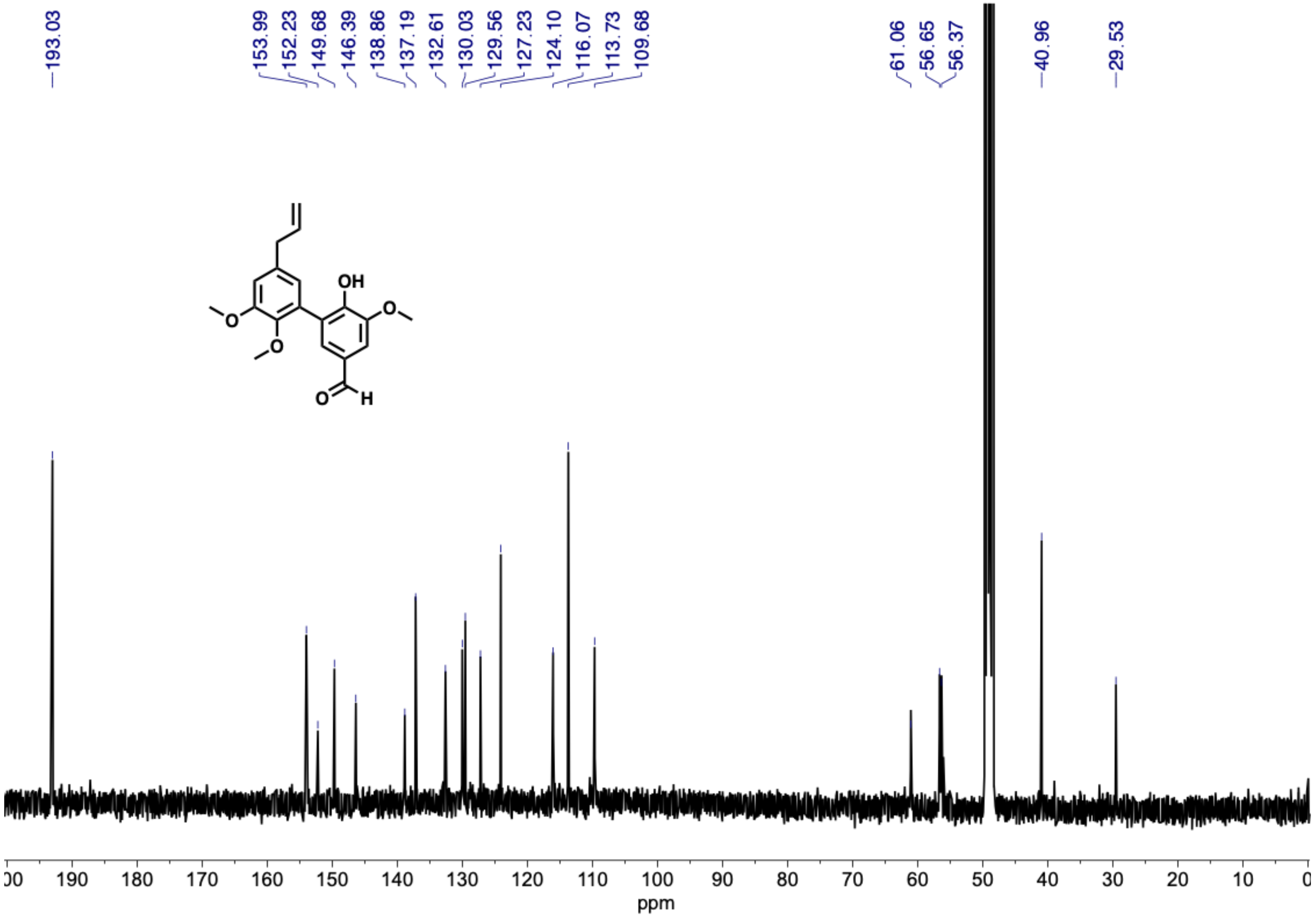


Figure A.3.16. ^{13}C NMR (101 MHz, CD_3OD) spectrum of compound **8**.

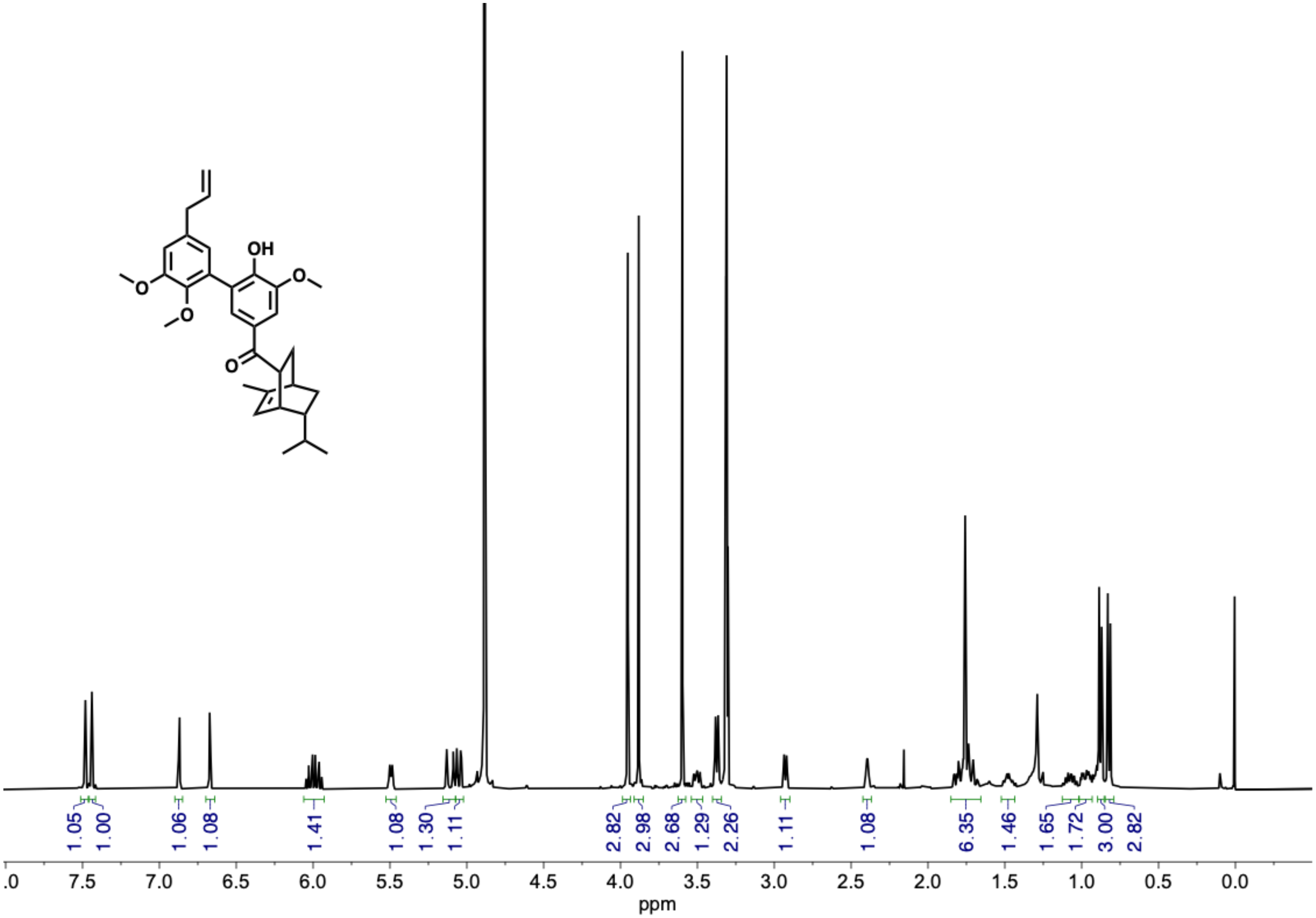


Figure A.3.17. ^{1}H NMR (400 MHz, CD_3OD) spectrum of compound **9**.

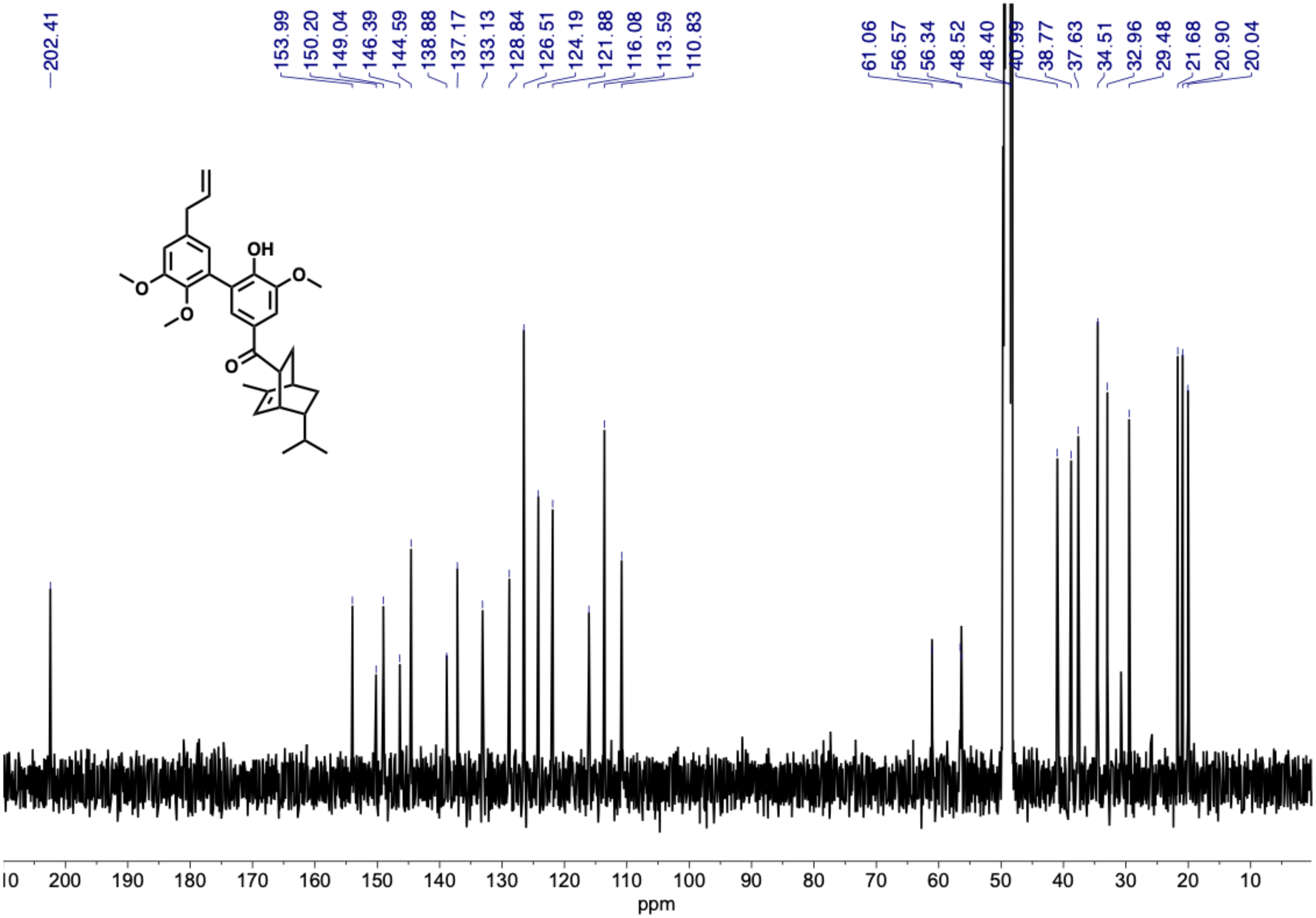


Figure A.3.18. ^{13}C NMR (101 MHz, CD$_3$OD) spectrum of compound **9**.

A.4. R script for ^{1}H NMR data treatment, network analysis and subsequent statistical treatment of modules

Required packages: WGCNA, ggplot2, ggdendro, egg, tidyr, ggpubr, factoextra, Morpho, HDMD

```
###====================ImportingData=======================###

#Read data for ".txt" files
MyData<-read.table("Adults_Data.txt", header = TRUE, row.names = 1,check.names = F) #change file name

#Read data for ".csv" files
MyData<-read.csv("CABsample.csv",sep=",", header = TRUE, row.names = 1,check.names = F) #change file name

###========================Spectral Editing=========================###

#If the data was generated using the Script version of MNova and it's already normalized and cleared of empty columns,
#skip directly to the network analysis section (rows must contain samples and columns the binned ppm in NMR or a peak/feature in MS).

#-----------------------------------------------------------------------------------------------------------
#NRM_diagn calculates an accumulated sum of spectra, and then iteratively calculates the mean of sums per chemical shift after removing one peak at a time.
#This method is a good diagnostic for intense solvent peaks that did not get totally removed during the spectrum treatment step.
#The generated plot shows the average area per bin for each removed peak, a red line indicates the threshold factor for the acceptable deviation from the mean (2x standard).
#In addition to the visual diagnostic, it also returns a list of ppm values that violate the threshold (standard = 2X mean average sum).
#-----------------------------------------------------------------------------------------------------------

NMR_diagn<-function(nmr,thresh=2){

  nmrSum<-data.frame(sum=apply(nmr,1,sum, na.rm=T)) #creates a sum of spectra across each ppm
  for (i in 1:nrow(nmrSum)){nmrSum$meanWO[i]=mean(nmrSum$sum[-i])} #for each removed ppm [i], calculates mean area per ppm without [i] (if its area is too big, it will lead to a much smaller value)
  thresh=mean(nmrSum$meanWO)-thresh*sd(nmrSum$meanWO) #sets the deviation from the mean sum of areas that should be considered an outlier
plot(ggplot(nmrSum,aes(x=as.numeric(row.names(nmrSum)),y=meanWO))+theme(axis.t
itle.x = element_text(size=12,margin=margin(0.3,0,0,0,"cm")),
    axis.text.x =
element_text(size=10,color="black",margin=margin(0.1,0,0,0,"cm")),axis.title.y =
element_text(size=12,margin=margin(0,0.3,0,0,"cm")),
    axis.text.y =
element_text(size=10,color="black",margin=margin(0,0.1,0,0,"cm")),plot.margin=margi
n(0.3,0.3,0.3,0.3,"cm"))+scale_x_reverse(
    name="ppm",expand=c(0,0),breaks =
seq((max(as.numeric(row.names(nmrSum)))%/%0.25+1)*0.25,by=-
0.25,(min(as.numeric(row.names(nmrSum)))%/%0.25
    )*0.25))+scale_y_continuous(name="Mean of sums without
outlier")+geom_hline(yintercept=thresh,linetype="dashed",color="red")+geom_line())

  ggsave("NMRoutlier.pdf",width=12,height = 8,units = "in")

  NMRoutlier<-row.names(nmrSum[which(nmrSum$meanWO<=thresh),]) #returns the
list of outliers
}

#-----------------------------------------------------------------------------------------------------------
#NMR_edit works for an NMR dataset that is generated without the MNova scripts (this
option results in fewer alignment issues)
#It normalizes each spectra by total sum (set as 100), then removes all columns
containing only values under the acceptable signal threshold (0.0001 is the standard)
#The resulting table is in the correct format for network analysis (ppm as columns)
#-----------------------------------------------------------------------------------------------------------

NMR_edit<-function(nmr,minCut=0.0001,outlier=NULL){

  if(is.null(outlier)==FALSE) {nmr< NMR[which(row.names(nmr)!=outlier),]} #removes
outlier peaks (from NMR_diagn)
  nmr<-t(nmr[which(complete.cases(nmr)),]*100)/apply(nmr,2,sum, na.rm=T)
#"which[complete.cases()]" removes variables containing NA-only values across the
dataset (solvent cut areas), then the data is transposed to have variables across the
coluumns. "apply" gathers the sum for each original column (samples), which is used to
normalize the spectra
  nmr[which(nmr<minCut)]<-0 #reassigns any bin area under the signal threshold as
zeroes
  nmr< NMR[,which(apply(nmr,2,sum)>0)] #removes zero-only variables (retains only
columns which sum>0)
}
```

```
###================Network Analysis===================###
#---------------------------------------------------------------------------------------------------------
#Net_power creates a plot to determine your beta-threshold value, this may be all over
the place but I start with the threshold at the beginning of the plateau#
#Red numbers indicate the SFTM fit values (primary axis), while the blue ones show the
corresponding mean node connectivity (secondary Y)
#The red line indicates the ideal threshold of 0.8 correlation with a Scale-Free Topology
model
#---------------------------------------------------------------------------------------------------------

Net_power<-function(dat,cutfit=0.80){

  sft = pickSoftThreshold(dat, powerVector = c(c(1:30))) #from WGCNA; calculates the
soft threshold parameters from an array of power values
  yfactor=(max(sft$fitIndices[,"mean.k."])%/%10+1)*10 #parameter for adjusting the
secondary Y axis

  plot(ggplot(sft$fitIndices,aes(x=Power))+geom_text(aes(y=-
sign(slope)*SFT.R.sq,label=Power),size=4,color="red")+geom_text(aes(y=mean.k./yfact
or,
    label=Power),size=4,color="blue")+scale_y_continuous(name=expression('Scale-Free
Topology Model Fit - signed R'^2),breaks=seq(0,1,0.2),
    sec.axis = sec_axis(~.*yfactor,name="Mean Node
Connectivity",breaks=seq(0,yfactor,yfactor/5)),limits=c(0,1))+geom_hline(yintercept=cut
fit,linetype="dashed",
    color="red")+labs(x=expression(paste("Power (", beta," )")),title="Network
Topology")+theme(plot.title = element_text(hjust=0.5,
    face="bold",size=16),axis.title = element_text(size=12),axis.text =
element_text(color="black", size=10), axis.title.y.left = element_text(
    margin=margin(0,0.2,0,0,"cm")), axis.title.y.right =
element_text(margin=margin(0,0,0,0.3,"cm"),size=12),plot.margin=margin(0.3,0.3,0.3,0.
3,"cm")))

  ggsave("NetTopo.pdf",width=12,height = 8,units = "in")
}

#---------------------------------------------------------------------------------------------------------
#Net_build creates and return the network. It also saves the TOM parameters for
graphical construction of the network on Cytoscape.
#To be entered: NMR data array, power, module-merging cut parameter ([1- degree of
correlation], 0.25 is the standard) and min module size (3 is standard).
#---------------------------------------------------------------------------------------------------------

Net_build<-function(dat,pwr,cutpar=0.25,
modSize=3,save=F,dataSave="MyData",cytThresh=0.02){
  #from WGCNA
  MyNet<-blockwiseModules(dat, power =pwr,TOMType = "unsigned", minModuleSize
=modSize,reassignThreshold = 0, mergeCutHeight = cutpar,
    numericLabels = TRUE, pamRespectsDendro = FALSE,saveTOMs =
TRUE,saveTOMFileBase = paste(dataSave,"TOM.csv",sep = "-"))

  if (save==T){
    #prep data for module composition table
    MEnodes<-data.frame("Module"=labels2colors(MyNet$colors))
    row.names(MEnodes)<-sprintf("%.2f",as.numeric(names(MyNet$colors)))

    #prep data for ME table
    MEs<-ME_name2color(MyNet,F)

    # gives a file of modules and eigenvalues across samples
    write.csv(MEs,paste(dataSave,"MEs.csv",sep = "-"))

    # gives a file of module-chemical shift (node) affiliation
    write.csv(MEnodes,paste(dataSave,"MEnodes.csv",sep = "-"))

    #prep data for network visualization in Cytoscape
    options(stringsAsFactors = FALSE)
    TOM<-TOMsimilarityFromExpr(dat, power=pwr) #recalculates TOM
    modules<-colnames(MEs[,-ncol(MEs)]) #all modules but grey are included
    probes<-colnames(dat) #extracts all ppm values
    selModules<-is.finite(match(MEnodes$Module, modules))
    modProbes<-probes[selModules] #excludes ppms from grey
    modTOM<-TOM[selModules, selModules] #filters TOM from grey nodes
    dimnames(modTOM)<-list(modProbes, modProbes)

    #export the edge and nodes file for cytoscape you may need to play with the threshold
here a bit to make it reflect the modules
    cyt = exportNetworkToCytoscape(modTOM,
                                   edgeFile = paste(dataSave,"Cyto_edges.txt", sep=""),
                                   nodeFile = paste(dataSave,"Cyto_nodes", ".txt", sep=""),
                                   weighted = TRUE,
                                   threshold = cytThresh,
                                   nodeNames = modProbes,
                                   altNodeNames = modules,
                                   nodeAttr = MEnodes$Module[selModules])
  }
  return(MyNet)
}
```

```
#------------------------------------------------------------------------------------------------------
#ME_name2color changes column names in the MEs table of entered network from
numbers to the corresponding module colors, while also removing grey
#------------------------------------------------------------------------------------------------------

ME_name2color<-function(net,rmGray=F){
  if (rmGray){
    MEs<-net$MEs[,!names(net$MEs)=="ME0"]
  }else{
    MEs<-net$MEs
  }
  names(MEs)<-labels2colors(as.numeric(substring(names(MEs),3)))
  return(MEs)
}


###=================Network Visualization====================###

#------------------------------------------------------------------------------------------------------
#Net_modules gives a list of the modules and the number of nodes in them as a graphical
representation
#------------------------------------------------------------------------------------------------------

Net_modules<-function(MyNet){

  mergedColors = labels2colors(MyNet$colors) #extracts module identity as colors
  moduleTable<-aggregate(mergedColors,by=list(mergedColors),FUN=length)
#consolidates and adds up nodes per module
  moduleTable<-cbind(moduleTable,Perc=100*moduleTable$x/sum(moduleTable[,2]))
#calculates % nodes in module from total

plot(ggplot(moduleTable,aes(y=reorder(Group.1,Perc),x=Perc,fill=Group.1))+geom_bar(
stat="identity")+labs(x="% of nodes",
    y="Modules",title="Network
Composition")+scale_fill_manual(values=moduleTable[order(moduleTable$Group.1),]$
Group.1
    )+theme(legend.position = "none",plot.title =
element_text(hjust=0.5,face="bold",size=16, margin=margin(0,0,0.3,0,"cm")),
    axis.title = element_text(size=12),axis.text = element_text(color="black",size=10),
axis.title.y = element_text(margin=margin(
    0,0.3,0,0,"cm")),axis.title.x =
element_text(margin=margin(0.3,0,0,0,"cm")),plot.margin=margin(0.3,0.3,0.3,0.3,"cm")

)+scale_y_discrete(labels=toupper(moduleTable[order(moduleTable$Perc),]$Group.1))+
geom_text(aes(label=x), hjust=-0.3,size=4
    ) + scale_x_continuous(minor_breaks = seq(1, 25,
1),limits=c(0,(max(moduleTable$Perc)%/%5+1)*5)))

  ggsave("ModuleComp.pdf",width=12,height = 8,units = "in")
}

#----------------------------------------------------------------------------------------------------------
#Cluster_dendro recreates the Node dendrogram with legend bars of module affiliation
(by colors).
#Module affiliation is shown for Modules generated by a Static cut (standard par is 0.99),
and by Dynamic cut (Par determined in Net_build)
#----------------------------------------------------------------------------------------------------------

Cluster_dendro<-function(MyNet, statCut=0.99){

  statiCut=as.character(cutreeStaticColor(MyNet$dendrograms[[1]],cutHeight =
statCut,minSize = 3)) #merges modules by the Static cut method (hard tree cut)

  mergedColors =
data.frame(cbind(labels2colors(MyNet$colors),X=names(MyNet$colors),statColors=stati
Cut,origiColors=labels2colors(
    MyNet$unmergedColors),Y=1),stringsAsFactors = F) #gathers module color data for
legend bar graph (unmerged and two merging methods)
  mergedColors<-mergedColors[MyNet$dendrograms[[1]]$order,] #sorts colors by the
node dendrogram
  mergedColors$order<-c(1:nrow(mergedColors)) #sets new order var for plotting
purposes
  mergedColors$Y<-as.integer(mergedColors$Y)

  #the next plots will show the module affiliation by color
  p3<-
ggplot(mergedColors,aes(x=reorder(X,order),y=Y,fill=V1))+geom_bar(stat="identity",wi
dth=1,fill=mergedColors$V1)+scale_fill_manual(
    values=mergedColors$V1)+ylab("Dynamic \nCut")+theme(legend.position =
"none",plot.title = element_text(hjust=0.5,face="bold",
    size=16),axis.title.x = element_blank(),axis.text = element_blank(),axis.title.y =
element_text(angle=0,vjust=0.5,size=12,margin=margin(0,-0.3,0,0,"cm")),
    axis.ticks = element_blank(),panel.border = element_rect(colour = "black", fill=NA,
size=0.5))+scale_y_continuous(expand=c(0,0),
    breaks = c(0,1),limits = c(0,1))+theme(plot.margin=margin(0,0.3,-0.1,0.3,"cm"))

  p2<-
ggplot(mergedColors,aes(x=reorder(X,order),y=Y,fill=statColors))+geom_bar(stat="iden
tity",width=1,fill=mergedColors$statColors)+scale_fill_manual(
    values=mergedColors$statColors)+ylab("Static \nCut")+theme(legend.position =
"none",plot.title = element_text(hjust=0.5,face="bold",
    size=16),axis.title.x = element_blank(),axis.text = element_blank(),axis.title.y =
element_text(angle=0,vjust=0.5,size=12,margin=margin(0,-0.3,0,0,"cm")),
    axis.ticks = element_blank(),panel.border = element_rect(colour = "black", fill=NA,
size=0.5))+scale_y_continuous(expand=c(0,0),
    breaks = c(0,1),limits = c(0,1))+theme(plot.margin=margin(0,0.3,0.3,0.3,"cm"))

  dendro2graph<-dendro_data(MyNet$dendrograms[[1]]) #extracts dendro data for
graphic manipulation

  for(i in 1:length(dendro2graph$segments$yend)){
    if(dendro2graph$segments$yend[i]==0){
      if((dendro2graph$segments$y[i]-0.05)>0) dendro2graph$segments$yend[i]<-
dendro2graph$segments$y[i]-0.05
    }
  } #this will shorten the leaves from the dendrogram (aesthetically more organized)

  gridLines=NULL

  for(i in 2:length(mergedColors[,1])) if(mergedColors[i,1]!=mergedColors[i-1,1])
gridLines<-c(gridLines,mergedColors[i,6]-0.5) #generates the gridlines in the
dendrogram that align with the Dynamic cut modules

  p4<-ggplot(segment(dendro2graph))+geom_vline(xintercept =
gridLines,lwd=0.3,color="white")+geom_segment(aes(x=x,y=y,
    xend=xend,yend=yend))+labs(title = "Cluster Dendrogram",
    y="Height")+theme(plot.title =
element_text(hjust=0.5,face="bold",size=16,margin=margin(0,0,0.3,0,"cm")),axis.title.x
= element_blank(),
    axis.title.y = element_text(size=12, color="black",angle =
90,margin=margin(0,0.3,0,0,"cm")),axis.text.y = element_text(size=10,
    color="black"),axis.text.x = element_blank(),axis.ticks.x =
element_blank())+scale_y_continuous(expand=c(0.01,0),
    limits =
c((min(dendro2graph$segments$yend)%/%0.1)*0.1,1))+scale_x_discrete(breaks=NULL)
+theme(plot.margin=margin(0.3,0.3,0,0.3,"cm"
    ))+geom_hline(yintercept=statCut,linetype="dashed",color="red")

  plot(ggarrange(p4,p3,p2,nrow = 3, ncol = 1, heights = c(8,1,1), align = "v")) #shows all
plots together

  ggsave("ClusterDendro.pdf",width=12,height = 8,units = "in")
}
```

```
#-------------------------------------------------------------------------------------------------------------
#Module_dendro creates a Module dendrogram with the corresponding colors.
#Pearson correlation and average method used as parameters
#-------------------------------------------------------------------------------------------------------------

Module_dendro<-function(net){

  #extract MEs info from network
  MEs<-ME_name2color(net,T)

  scaledME<-as.matrix(scale(t(MEs)))
  modDendro<-
as.dendrogram(hclust(d=get_dist(x=scaledME,method="pearson"),method = "average"))
#module dendrogram
  modOrder<-order.dendrogram(modDendro)

  #extracts data for plotting circles with module colors

  ddata<-dendro_data(modDendro)
  colors<-as.character(ddata$labels$label)
  print(colors)

 plot(ggdendrogram(data=modDendro,rotate=F)+theme_gray()+theme(panel.grid.minor.x
= element_blank(),panel.grid.major.x = element_blank(),axis.ticks =
element_blank(),axis.text.y = element_text(size=10,
    color="black", angle=0,margin=margin(0,0,0,0,"cm"),hjust = 1),axis.title.x =
element_blank(),axis.text.x = element_text(size=12,
    color="black", angle=90,margin=margin(0,0,0,0,"cm"),hjust = 1,vjust=0.5),axis.title.y
= element_text(size=12,
    color="black",margin=margin(0.3,0.3,0.3,0.3,"cm")),plot.margin = margin(0.3,0.3,
    0.3,0.3,"cm"),plot.title = element_blank())+labs(ylab("Distance"))+labs(y="Distance")
+scale_y_continuous(expand = c(0.01,0.015))+
    geom_point(data = ddata$labels,aes(x = x,y = y,fill=label),size = 8,shape = 21,
    show.legend = F)+scale_fill_manual(values=colors))

  ggsave("MEdendro.pdf",width=12,height = 8,units = "in")
}

#-------------------------------------------------------------------------------------------------------------
#ModSample_heatmap plots a heatmap of module eigenvalues for the sample set
#-------------------------------------------------------------------------------------------------------------

ModSample_heatmap<-function(net,textSize=12,method="pearson",orient="h"){

  #extract MEs info from network
  MEs<-ME_name2color(net,T)
  MEs$samples<-row.names(MEs)

  scaledME<-as.matrix(scale(MEs[1:(ncol(MEs)-1)])) #scaled variables for HCA without
grey
  sampleDendro<-
as.dendrogram(hclust(d=get_dist(x=scaledME,method=method),method = "complete"))
#sample dendrogram
  sampleOrder<-order.dendrogram(sampleDendro)
  scaledME<-as.matrix(scale(t(MEs[1:(ncol(MEs)-1)])))
  modDendro<-as.dendrogram(hclust(d=get_dist(x=scaledME,method=method),method =
"average")) #module dendrogram
  modOrder<-order.dendrogram(modDendro)

  #extract the sample dendrogram to be displayed in the heatmap and the order of
clustered samples
  longMEs<-pivot_longer(data=MEs,cols = -
c(samples),names_to="module",values_to="ME") #trasforms to long format (necessary
for heatmap)
  longMEs$samples<-
factor(x=longMEs$samples,levels=MEs$samples[sampleOrder],ordered=T) #reorder
samples according to clusters
  longMEs$module<-factor(x=longMEs$module,levels=colnames(MEs[1:(ncol(MEs)-
1)])[modOrder],ordered=T) #reorder modules according to clusters

  #extract the names of ordered clustered modules
  colBar<-data.frame(factor(names(MEs[,-ncol(MEs)][modOrder]),levels=names(MEs[,-
ncol(MEs)][modOrder])),Y=1)  #extracts module colors for labeling the x-axis on the
heatmap
  names(colBar)[1]<-"module"

  if(orient=="h"){
    p1<-ggplot(longMEs,aes(x=module,y=samples,fill=ME)) + geom_tile() +
scale_fill_gradient2(name="Eigenvalue",low="dodgerblue",high = "brown1",
      guide = guide_colorbar(frame.colour = "black",ticks = TRUE,nbin = 10,label.position
= "bottom",
      barwidth = 13,barheight = 1.3,direction = "horizontal",ticks.colour = "black")) +
theme(axis.title = element_blank(),axis.ticks.x = element_blank(),axis.text.x =
element_blank(),
      axis.text.y.right = element_text(margin = margin(0,0,0,0.1,"cm"),size=textSize,
color="black"),panel.border = element_rect(color="black",size=0.5,fill=NA),
      plot.margin = margin(0.3,0,0,0.3,"cm"),legend.position =
"top",legend.title=element_text(size=12,margin = margin(0,0.3,0,0),vjust=0.7),
legend.text = element_text(size=10),legend.box.margin = margin(0,0,-0.3,0,"cm")) +
scale_x_discrete(expand = c(0,0))+scale_y_discrete(expand = c(0,0),
      guide = guide_axis(check.overlap = T),position = "right")

    p2<-ggdendrogram(data=sampleDendro,rotate=T)+theme(legend.text =
element_text(size=12,color="black"),legend.margin = margin(0,0,0.3,-1),
      axis.text.x = element_blank(),plot.margin =
margin(1.9+(14.7/(nrow(MEs)*2))*(2.1/1.6),0.3,-
0.22+(14.7/(nrow(MEs)*2))*(2.1/1.6),0,"cm"),
      plot.title = element_text())+scale_x_discrete(expand=c(0.005,0.005))

    p3<-
ggplot(colBar,aes(x=module,y=Y,fill=module))+geom_bar(stat="identity",width=1,fill=c
olBar$module)+theme(axis.title=element_blank(),axis.ticks = element_blank(),axis.text =
element_blank(),panel.border =
element_rect(color="black",size=0.5,fill=NA),plot.margin = margin(-
0.1,0.3,0.3,0.3,"cm"))+scale_y_continuous(expand = c(0,0),position =
"right")+scale_x_discrete(expand = c(0,0))+labs(y="Modules")

    p4<-ggplot() + annotate("text",x=1,y=1,hjust=0.8,vjust=-0.5,size=5, label =
"",fontface=2) +theme_void()


plot(ggpubr::ggarrange(egg::ggarrange(p1,p3,nrow=2,heights=c(16,1)),egg::ggarrange(p
2,p4,nrow=2,heights=c(16,1)),ncol=2,widths=c(8,2)))

    ggsave("MEsample.pdf",width=12,height = 8,units = "in")

  }else{

    p1<-ggplot(longMEs,aes(y=module,x=samples,fill=ME)) + geom_tile() +
scale_fill_gradient2(name="Eigenvalue",low="dodgerblue",high = "brown1",
      guide = guide_colorbar(frame.colour = "black",ticks = TRUE,nbin = 10,label.position
= "bottom",
      barwidth = 13,barheight = 1.3,direction = "horizontal",ticks.colour = "black")) +
theme(axis.title = element_blank(),axis.ticks.y = element_blank(),axis.text.y =
element_blank(),
      axis.text.x.top = element_text(margin = margin(0,0,0.1,0,"cm"),size=textSize,
color="black", angle=90,hjust=0),panel.border =
element_rect(color="black",size=0.5,fill=NA),
      plot.margin = margin(0,0,0.3,0.3,"cm"),legend.position =
"bottom",legend.title=element_text(size=12,margin = margin(0,0.3,0,0),vjust=0.7),
legend.text = element_text(size=10),
      legend.box.margin = margin(-0.3,0,0,0,"cm")) + scale_y_discrete(expand =
c(0,0))+scale_x_discrete(expand = c(0,0),
      guide = guide_axis(check.overlap = T),position = "top")
    p2<-ggdendrogram(data=sampleDendro,rotate=F)+theme(
      axis.text.y = element_blank(),plot.margin =
margin(0.3,0.10+(27.4/(nrow(MEs)*2))*(2.1/1.6),-
0.3,0.15+(27.4/(nrow(MEs)*2))*(2.1/1.6),"cm"),
      plot.title = element_text())+scale_x_discrete(expand=c(0.005,0.005))

    p3<-
ggplot(colBar,aes(x=module,y=Y,fill=module))+geom_bar(stat="identity",width=1,fill=c
olBar$module)+theme(axis.title=element_blank(),axis.ticks = element_blank(),
      axis.text = element_blank(),panel.border =
element_rect(color="black",size=0.5,fill=NA),plot.margin = margin(0,0.3,0,-
0.1,"cm"))+scale_y_continuous(expand = c(0,0),
      position = "right")+scale_x_discrete(expand =
c(0,0))+labs(y="Modules")+coord_flip()

    p4<-ggplot() + annotate("text",x=1,y=1,hjust=1.5,vjust=0,size=5, label =
"",fontface=2,angle=90) +theme_void()

    plot(ggpubr::ggarrange(egg::ggarrange(p2,p4,ncol=2,widths
=c(40,1)),egg::ggarrange(p1,p3,ncol=2,widths =c(40,1)),nrow=2,heights=c(2,8)))

    ggsave("MEsample.pdf",width=12,height = 8,units = "in")

  }
}

#----------------------------------------------------------------------------------------------------------
#ModTrait_heatmap will plot a heatmap of correlation between module eigenvalues and
the provided traits
#Applicable to numeric traits; for categorical data, use ModTraitC_heatmap
#----------------------------------------------------------------------------------------------------------

ModTrait_heatmap<-function(net,traits,textSize=12,valueSize=3.5){

  MEs<-net$MEs
  names(MEs)<-labels2colors(as.numeric(substring(names(MEs),3)))
  traits<-traits[match(row.names(MEs),row.names(traits)),]
  modCor<-corAndPvalue(MEs[,1:(ncol(MEs)-1)],traits)

  scaledCorr<-as.matrix(scale(t(modCor$cor))) #scaled traits for HCA
  traitDendro<-as.dendrogram(hclust(d=dist(x=scaledCorr),method = "average"))
  traitOrder<-order.dendrogram(traitDendro)
  scaledCorr<-as.matrix(scale(modCor$cor))
  modDendro<-as.dendrogram(hclust(d=dist(x=scaledCorr),method = "average"))
  modOrder<-order.dendrogram(modDendro)
  longCorr<-pivot_longer(data=modCor$cor,cols = -
mod,names_to="trait",values_to="corr")
  longCorr$mod<-factor(x=longCorr$mod,levels=row.names(modCor$cor),ordered=T)
  longCorr$trait<-
factor(x=longCorr$trait,levels=colnames(modCor$cor)[traitOrder],ordered=T) #reorder
samples according to clusters
  longCorr$pVal<-pivot_longer(data=modCor$p,cols = everything(),names_to =
"trait",values_to = "pVal")$pVal

  colBar<-data.frame(factor(row.names(modCor$cor),row.names(modCor$cor)),Y=1)
#extracts module colors for labeling the x-axis on the heatmap
  names(colBar)[1]<-"module"

  p1<-ggplot(longCorr,aes(x=mod,y=trait,fill=corr)) + geom_tile() +
scale_fill_gradient2(name="Pearson corr.",low="dodgerblue",
    high = "brown1",guide = guide_colorbar(frame.colour = "black",ticks = TRUE,nbin =
10,label.position = "bottom",
    barwidth = 13,barheight = 1.3,direction = "horizontal",ticks.colour = "black")) +
theme(axis.title = element_blank(),axis.ticks.x = element_blank(),
    axis.text.x = element_blank(),axis.text.y.left = element_text(margin =
margin(0,0.1,0,0,"cm"),size=textSize,vjust=0.5,color="black"),
    panel.border = element_rect(color="black",size=0.5,fill=NA),plot.margin =
margin(0.3,0.3,0,0.3,"cm"),legend.position = "top",
    legend.title=element_text(size=textSize,margin = margin(0,0.3,0,0),vjust=0.7),
legend.text = element_text(size=0.85*textSize),legend.box.margin = margin(0,0,-
0.3,0,"cm")) + scale_x_discrete(
    expand = c(0,0))+scale_y_discrete(expand = c(0,0),guide = guide_axis(check.overlap
= T),position = "left"
    )+geom_text(aes(label=paste(signif(corr, 2), "\n(", signif(pVal, 1), ")", sep =
"")),size=valueSize,lineheight=0.8,vjust=0.5)
  p2<-
ggplot(colBar,aes(x=module,y=Y,fill=module))+geom_bar(stat="identity",width=1,fill=c
olBar$module)+theme(axis.title.x=element_blank(
    ),axis.title.y = element_text(angle=0,vjust=0.5, size=textSize, face="bold"),axis.ticks =
element_blank(),axis.text = element_blank(),panel.border =
element_rect(color="black",size=0.5,fill=NA),plot.margin = margin(-
0.1,0.3,0.3,0.3,"cm"))+scale_y_continuous(expand = c(0,0))+scale_x_discrete(expand =
c(0,0))+labs(y="Modules")

  egg::ggarrange(p1,p2,nrow=2,heights=c(16,1))
}
```

```
#-----------------------------------------------------------------------------------------------------
#ModTraitC_heatmap will plot a heatmap of mean or median eigenvalues per grouping
defined under traits
#Group distances are calculated as group Mahalanobis distance (MD) and groups are
ordered according to the dendrogram
#Applicable to categorical data
#-----------------------------------------------------------------------------------------------------

ModTraitC_heatmap<-
function(net,traits,subGroup=NULL,subGcolors=NULL,textSize=12,valueSize=3.5,mod
e="mean"){

  #calculates ME-trait associations
  MEs<-ME_name2color(net,T)
  MDist<-MD_dist(MEs,traits)
  meanMEs<-as.data.frame(MDist$means)
  colnames(meanMEs)<-colnames(MEs)
  meanMEs$group<-row.names(meanMEs)

  #generates MEdendro for ordering, based on original data
  scaledME<-as.matrix(scale(t(MEs)))
  modDendro<-
as.dendrogram(hclust(d=get_dist(x=scaledME,method="pearson"),method = "average"))
#module dendrogram
  modOrder<-order.dendrogram(modDendro)
  scaledTraits<-as.matrix(scale(meanMEs[,-ncol(meanMEs)]))
  dist<-MD_dist(MEs,traits)$distance
  row.names(dist)<-meanMEs$group
  trtDendro<-as.dendrogram(hclust(d=as.dist(dist),method="average"))
  trtOrder<-order.dendrogram(trtDendro)

  longMEs<-pivot_longer(data=as.data.frame(meanMEs),cols = -
group,names_to="Module",values_to="mean")
  longMEs$Module<-
factor(x=longMEs$Module,levels=colnames(MEs[modOrder]),ordered=T)
  longMEs$trait<-
factor(x=longMEs$group,levels=meanMEs$group[trtOrder],ordered=T) #reorder
samples according to clusters

  colBar<-
data.frame(factor(colnames(MEs)[modOrder],colnames(MEs)[modOrder]),Y=1)
#extracts module colors for labeling the x-axis on the heatmap
  names(colBar)[1]<-"module"
  p1<-ggplot(longMEs,aes(x=Module,y=trait,fill=mean)) + geom_tile() +
scale_fill_gradient2(name="Mean eigenvalue",low="dodgerblue",
   high = "brown1",guide = guide_colorbar(frame.colour = "black",ticks = TRUE,nbin =
10,label.position = "bottom",
   barwidth = 13,barheight = 1.3,direction = "horizontal",ticks.colour = "black")) +
theme(axis.title = element_blank(),axis.ticks.x = element_blank(),
   axis.text.x = element_blank(),axis.text.y.right = element_text(margin =
margin(0,0,0,0.1,"cm"),size=textSize,vjust=0.5,color="black"),
   panel.border = element_rect(color="black",size=0.5,fill=NA),plot.margin =
margin(0.3,0,0,0.3,"cm"),legend.position = "top",
   legend.title=element_text(size=textSize,margin = margin(0,0.3,0,0),vjust=0.7),
legend.text = element_text(size=0.85*textSize),legend.box.margin = margin(0,0,-
0.3,0,"cm")) + scale_x_discrete(
   expand = c(0,0))+scale_y_discrete(expand = c(0,0),guide = guide_axis(check.overlap =
T),position = "right"
   )+geom_text(aes(label=round(mean,
digits=3)),size=valueSize,lineheight=0.8,vjust=0.5)

  p2<-
ggplot(colBar,aes(x=module,y=Y,fill=module))+geom_bar(stat="identity",width=1,fill=c
olBar$module)+theme(axis.title.x=element_blank(
   ),axis.title.y = element_blank(),axis.ticks = element_blank(),axis.text =
element_blank(),panel.border =
element_rect(color="black",size=0.5,fill=NA),plot.margin = margin(-
0.1,0.3,0.3,0.3,"cm"))+scale_y_continuous(expand = c(0,0))+scale_x_discrete(expand =
c(0,0))

  if(is.null(subGroup)){

    p3<-ggdendrogram(data=trtDendro,rotate=T)+theme(legend.text =
element_text(size=12,color="black"),legend.margin = margin(0,0,0.3,-1),
     axis.text.x = element_blank(),plot.margin =
margin(1.9+(14.7/(nrow(meanMEs)*2))*(2.1/1.6),0.3,-
0.22+(14.7/(nrow(meanMEs)*2))*(2.1/1.6),0,"cm"),
     plot.title = element_text())+scale_x_discrete(expand=c(0.005,0.005))

  }else{

    leafs<-dendro_data(trtDendro)$labels
    groupBar<-
data.frame(x=leafs$x,color=subGcolors[match(subGroup[match(leafs$label,traits)],uniqu
e(subGroup))])

    p3<-ggdendrogram(data=trtDendro,rotate=T)+theme(legend.text =
element_text(size=12,color="black"),legend.margin = margin(0,0,0.3,-1),
     axis.text.x = element_blank(),plot.margin = margin(1.9,0.3,-0.22,0,"cm"),
     plot.title = element_text())+scale_x_discrete(expand=c(0.005,0.005))+
     annotate("rect",xmin=groupBar$x-0.5,xmax =
groupBar$x+0.5,ymin=0,ymax=max(dendro_data(trtDendro)$segments$yend),fill=group
Bar$color,alpha=0.2)

  }
  p4<-ggplot() + annotate("text",x=1,y=1,hjust=0.8,vjust=-0.5,size=5, label =
"",fontface=2) +theme_void()


plot(ggpubr::ggarrange(egg::ggarrange(p1,p2,nrow=2,heights=c(16,1)),egg::ggarrange(p
3,p4,nrow=2,heights=c(16,1)),ncol=2,widths=c(8,2)))

  ggsave("Module-Trait.pdf",width=14,height = 9,units = "in")

}


###================Similarity Analysis===================###
#----------------------------------------------------------------------------------------------------
#Sim_method evaluates the best similarity method for calculating sample distance
#Method's effectiveness is rated based on the ratio between out-group distance and in-
group distance, according to entered categorical variable
#----------------------------------------------------------------------------------------------------

Sim_method<-function(MEs, groups){
  disMethod<-c("euclidean", "maximum", "manhattan", "canberra", "binary",
"minkowski", "pearson", "spearman","kendall")
  meanDist<-
data.frame(method=disMethod,mean_intraD=NA,mean_interD=NA,ratio=NA,pval=NA)
  MEscaled<-scale(MEs)
  traits<-data.frame(sample=row.names(MEs),group=groups)

  for(k in 1:length(disMethod)){

    distMatrix<-as.matrix(get_dist(MEscaled,disMethod[k]))
    distMatrix[upper.tri(distMatrix,T)]<-NA
    MEdist<-as.data.frame(distMatrix)
    MEdist$sample1<-row.names(MEscaled)
    longDist<-na.omit(pivot_longer(data=MEdist,cols=-
sample1,names_to="sample2",values_to="dist"))
    longDist$group1<-traits$group[match(longDist$sample1,traits$sample)]
    longDist$group2<-traits$group[match(longDist$sample2,traits$sample)]
    meanDist$mean_intraD[k]<-
mean(longDist$dist[which(longDist$group1==longDist$group2)])
    meanDist$mean_interD[k]<-
mean(longDist$dist[which(longDist$group1!=longDist$group2)])
    meanDist$ratio[k]<-meanDist$mean_interD[k]/meanDist$mean_intraD[k]
    meanDist$pval[k]<-
formatC(t.test(longDist$dist[which(longDist$group1==longDist$group2)],longDist$dist[
which(longDist$group1!=longDist$group2)],)$p.value,format="e",digits = 2)
  }

  meanDist
}

#-----------------------------------------------------------------------------------------------------------
#Group_dist calculates out-group and in-group distances, according to entered categorical
variable and distance method
#-----------------------------------------------------------------------------------------------------------

Group_dist<-function(MEs,disMethod,groups){

  MEscaled<-scale(MEs)
  traits<-data.frame(sample=row.names(MEs),group=groups)
  distMatrix<-as.matrix(get_dist(MEscaled,disMethod))
  distMatrix[upper.tri(distMatrix,T)]<-NA
  MEdist<-as.data.frame(distMatrix)
  MEdist$sample1<-row.names(MEscaled)
  longDist<-na.omit(pivot_longer(data=MEdist,cols=-
sample1,names_to="sample2",values_to="dist"))
  longDist$group1<-traits$group[match(longDist$sample1,traits$sample)]
  longDist$group2<-traits$group[match(longDist$sample2,traits$sample)]

  meanDist<-
data.frame(group=unique(groups),mean_intraD=NA,sd_intraD=NA,mean_interD=NA,sd
_interD=NA)

  for(k in 1:length(meanDist$group)){

    meanDist$mean_intraD[k]<-
mean(longDist$dist[which(longDist$group1==meanDist$group[k] &
longDist$group2==meanDist$group[k])])
    meanDist$mean_interD[k]<-
mean(longDist$dist[which(longDist$group1==meanDist$group[k] |
longDist$group2==meanDist$group[k])])
```

```
    meanDist$sd_intraD[k]<-
sd(longDist$dist[which(longDist$group1==meanDist$group[k] &
longDist$group2==meanDist$group[k])])
    meanDist$sd_interD[k]<-
sd(longDist$dist[which(longDist$group1==meanDist$group[k] |
longDist$group2==meanDist$group[k])])
  }

  meanDist
}

#----------------------------------------------------------------------------------------------------
#MD_dist returns the group Mahalanobis distance, according to entered categorical
variable
#-----------------------------------------------------------------------------------------------------

MD_dist<-function(data,group){

  covar<-covW(data,group) #make sure to enter category column
  MD<-pairwise.mahalanobis(data,grouping=group,cov=covar,digits=3) #calculates the
distance.

  return(MD)
}
```