Methodological advances for hypothesis-driven ethnobiology

Orou G. Gaoue1,2,3*, Jacob K. Moutouama1, Michael A. Coe4, Matthew O. Bond4, Elizabeth Green1, Nadejda B. Sero1, Bezeng S. Bezeng2 and Kowiyou Yessoufou2

1Department of Ecology and Evolutionary Biology, University of Tennessee Knoxville, 569 Dabney Hall, Knoxville, TN, 37996, U.S.A.
2Department of Geography, Environmental Management and Energy Studies, University of Johannesburg, APK Campus, Auckland Park, Johannesburg, 2006, South Africa
3Faculty of Agronomy, University of Parakou, Parakou, BP 123, Benin
4Department of Botany, University of Hawai'i at Mānoa, 3190 Maile Way, 101, Honolulu, HI, 96822, U.S.A.

ABSTRACT

Ethnobiology as a discipline has evolved increasingly to embrace theory-inspired and hypothesis-driven approaches to study why and how local people choose plants and animals they interact with and use for their livelihood. However, testing complex hypotheses or a network of ethnobiological hypotheses is challenging, particularly for data sets with non-independent observations due to species phylogenetic relatedness or socio-relational links between participants. Further, to account fully for the dynamics of local ecological knowledge, it is important to include the spatially explicit distribution of knowledge, changes in knowledge, and knowledge transmission and use. To promote the use of advanced statistical modelling approaches that address these limitations, we synthesize methodological advances for hypothesis-driven research in ethnobiology while highlighting the need for more figures than tables and more tables than text in ethnobiological literature. We present the ethnobiological motivations for conducting generalized linear mixed-effect modelling, structural equation modelling, phylogenetic generalized least squares, social network analysis, and predictive modelling. For each element of the proposed ethnobiologists quantitative toolbox, we present practical applications along with scripts for a widespread implementation. Because these statistical modelling approaches are rarely taught in most ethnobiological programs but are essential for careers in academia or industry, it is critical to promote workshops and short courses focused on these advanced methods. By embracing these quantitative modelling techniques without sacrificing qualitative approaches which provide essential context, ethnobiology will progress further towards an expansive interaction with other disciplines.

Key words: community phylogenetics, ethnobiology, theory-driven ethnobotany, quantitative ethnobotany, statistical modelling, structural equation modelling, social network analysis, phylogenetic generalized least squares

CONTENTS

I. Introduction ................................................................. 2282
II. Overview of the advanced quantitative methods ........................................... 2283
III. Advanced statistical methods for hypothesis testing in ethnobiology ................. 2284
   (1) Generalized linear mixed-effect models (GLMMs) ................................... 2284
   (2) Structural equation modelling (SEM) ............................................... 2285
   (3) Phylogenetic generalized least squares (PGLS) ..................................... 2287
   (4) Testing for a phylogenetic signal ................................................... 2288
   (5) Social network analysis and exponential random graph modelling (ERGM) ........ 2290
      (a) Multivariate regression using social network connectivity to predict individual knowledge and knowledge distribution of networks ........................................... 2292

* Address for correspondence (Tel: +1 865 974 3046; E-mail: ogaoue@utk.edu)
I. INTRODUCTION

It is becoming increasingly clear that the human dimension of conservation science plays a significant role in solving various traditional and emerging conservation crises (Poe, Norman & Levin, 2014). To be successful, solutions to these conservation problems must be culturally appropriate, judged by how they are adopted and implemented by and with the local people who live with the biodiversity we want to conserve (Winter, Ticktin & Quazi, 2020). Therefore, it is critical that we study, understand, and value local people’s knowledge of their environment and how they manage it. Embracing the contribution that ethnobiology can make in this context is often treated in parallel in the literature with the discipline often misunderstood or even not considered. For example, studies on the human dimension in conservation list numerous integrated or combined disciplines (e.g. environmental anthropology, environmental sociology, human-environment geography, environmental humanities). However, they fail to acknowledge ethnobiology as a key discipline within the domain of social conservation science (Bennett et al., 2017). This suggests that ethnobiology needs further development to engage with the biological and mathematical sciences. To achieve this, ethnobiology must embrace the theory-inspired and hypothesis-driven approach we recently advocated (McClatchey et al., 1999; Gaoue et al., 2017) along with several other authors spanning more than half a century (Ford, 1978; Bennett, 2005; Martin, 2007; Albuquerque, 2009; Albuquerque & Hanazaki, 2009; Hurrell & Albuquerque, 2012; Albuquerque et al., 2019a). Moreover, to improve methodological rigour, it is important that ethnobiology also embraces advanced methodological approaches, particularly new statistical modelling methods without sacrificing the strength of qualitative methods that provide important contexts.

Ethnobiology is the study of the interactions between people, organisms, and culture. Because of its interdisciplinary nature, most ethnobiologists are either trained biologists who are comfortable conducting interviews and asking questions about plants/animals or environmental use, or trained anthropologists who are interested in learning about, identifying and cataloging plant or animal species. This double origin of ethnobiologists is consistent with the hybrid methodological culture of this discipline and has led to some of the methodological challenges that it now faces. Ethnobiologists have been exposed to a diversity of fields and analytical methods of inquiry for several decades. These methods are well established in series of excellent books (e.g. Alexiades, 1996; Cunningham, 2001; Martin, 2007; Albuquerque et al., 2014) and are regularly used in this field. However, most ethnobiological studies tend to include more tables than figures and lack statistical rigour. The seminal papers of Phillips & Gentry (1993a,b) led to several studies seeking to test hypotheses using classic statistical methods such as one-way analysis of a variance, Student t-test and chi-squared. However, most quantitative methods in ethnobiology were limited to metrics such as species importance indices (Phillips & Gentry, 1993a,b; Hoffman & Gallaher, 2007). With the development of new statistical methods such as classification, ordination, and general linear models, tremendous progress has been made in ethnobiology (Moerman, 1979; Voeks & Leony, 2004). Begossi (1996) introduced the well-known ecological methods of diversity rarefaction techniques which gained much support in ethnobotanical studies. As a result, several studies on the diversity of local knowledge of plant use have conducted rarefaction analyses to estimate plant knowledge richness (Colwell, Mao & Chang, 2004). Despite this progress, most quantitative ethnobiological methods are used without a clear conceptual framework, constraining our understanding of the mechanisms behind plant use and selection. Recently Gaoue et al. (2017) proposed a synthesis of 17 testable hypotheses in ethnobotany and thereby built a foundation upon which advanced statistical methods can be applied to enhance ethnobiology as a scientific discipline. There is now a small but increasing number of hypothesis-driven ethnobiological studies (Almeida et al., 2003; Alencar et al., 2010; Soldati & Albuquerque, 2012; Hart et al., 2017; Seyler et al., 2019; Bond & Gaoue, 2020; Coe & Gaoue, 2021; Muleba, Yessoufou & Rampedi, 2021).
Because most Ph.D graduates are more likely to find careers in industry rather than tenure-track academic positions, and the ethnobiologists that do find academic jobs tend to join life sciences or ecology and evolutionary biology departments, it is important that ethnobiological education includes a robust quantitative analysis toolkit (Beans, 2018). This toolkit must go beyond classical statistical analyses to include advanced statistical modelling tools. An understanding of these advanced analytical approaches will make ethnobiologists more competitive in the job market and also allow them to advance our understanding of how people select plants and animals for use, thereby shaping their environment in a non-random way. We postulate that the slow adoption of modern statistical tools in ethnobiology is due to the apparent complexity they present and the need to use syntax-based programming software such as R (R Development Core Team, 2019) to run these analyses. We also postulate that a clear presentation of the ethnobiological applicability of these advanced methods, practical examples using these methods to address ethnobiological research questions and well-commented seed R scripts implementing these methods will improve the likelihood that ethnobiologists embrace these new tools.

Herein, we synthesize methodological advances for hypothesis-driven research in ethnobiology. This includes how to test hypotheses in ethnobiology using (i) generalized linear mixed-effect modelling, (ii) structural equation modelling, (iii) phylogenetic generalized least squares, (iv) community phylogenetic tools, (v) social network analysis, (vi) species distribution modelling and spatial ecological tools, and (vii) predictive modelling. For each quantitative method, we include a brief presentation of the methods and its goals, why/how it can be used to test ethnobiological hypotheses and provide example or seed R scripts showing how it can be applied to real field data (see Codes R1-6 and associated data sets in Gaoue et al., 2021). We selected these quantitative methods for two main reasons. First, these methods are commonly used in the biological sciences, and familiarity with these methods will allow ethnobiologists to communicate their science better to a broader audience. Second, these methods are particularly useful to test ethnobiological hypotheses (Gaoue et al., 2017; Albuquerque et al., 2019b), enabling ethnobiologists to avoid common statistical data-analysis errors.

II. OVERVIEW OF THE ADVANCED QUANTITATIVE METHODS

Gaining a mechanistic understanding of why local people select plants or animals for use is a major question of interest for ethnobiologists (Gaoue et al., 2017; Albuquerque et al., 2019a). In particular, understanding how and why plant/animal traits and human traits synergistically affect local people’s knowledge and use of plants is at the core of ethnobotany inquiries. Several theoretical and empirical studies suggest the likelihood that a given plant species or family is selected and used as medicine depends on the socio-demographic traits of who is using the plant (Gaoue et al., 2017). For example, some plant species are used more by a given gender (Voeks & Leony, 2004) or age group (Albuquerque et al., 2011; Souto & Tickin, 2012) than others. Medicinal and food plants are not selected randomly but according to their chemical traits (Moerman, 1979, 1996), and also indirectly for their physical traits (Ford & Gaoue, 2017). It is important to be able to investigate these features systematically to gain a mechanistic understanding of plant selection by human societies across historical and contemporary scales.

The data sets used to test these hypotheses often include species or participants as observations. To test hypotheses, one must take care not to violate the assumption of independent observations (Kenny & Judd, 1986; Romano & Kromrey, 2009). However, because plant species have evolutionary relationships and human participants can have kinship or social relationships or some form of spatial autocorrelation, it is critical to control for non-independent observations. Phylogenetic generalized least square (PGLS) methods can be used to control for such phylogenetic relatedness between plant species and confirm the importance of phylogeny in conferring medicinal value to plants (Saslis-Lagoudakis et al., 2014; Ernst et al., 2015; Yessoufou, Daru & Muasya, 2015). Considering kinship by using subjects as random effects in generalized mixed-effect or hierarchical models can also be used to account for family or genealogical relatedness among participants.

Ethnobiology as a discipline should embrace the role of socio-relational linkages (Bodin, Crona & Ernstson, 2006; Crona & Bodin, 2006; Bond & Gaoue, 2020) in the creation, transformation and transmission of knowledge (Gaoue et al., 2017). Of particular interest is the role of social influence, social capital, and homophily in shaping the dynamics of local ecological knowledge (Prell, 2012; Kendal et al., 2018; Kluger et al., 2020). Although social relations are at the core of community dynamics, these interactions are often neglected in ethnobotanical literature [but see Hopkins (2011) and Reyes-García et al. (2013)] as strong drivers of knowledge and subsequent practice. It is critically important to control for the impact of such relational effects when testing the effect of other drivers of knowledge such as human demographic traits or species physical and chemical traits. Social network analysis can be used in ethnobiology to understand the drivers of plant or animal selection and use for livelihood, and also to assess knowledge transmission mechanisms (Gallois et al., 2018; Castiñeira Latorre, Canaverò & Arin, 2020).

Separating correlation from causation is one of the most important discussions in biological and social sciences (Shipley, 2016). However, demonstrating causation is not possible using most common statistical methods, which only measure correlation. Finding evidence of causation is particularly challenging when working with multiple predictor variables. Multiple predictors are often rare in ethnobiological
III. ADVANCED STATISTICAL METHODS FOR HYPOTHESIS TESTING IN ETHNOBIOLOGY

The first set of methods tests for significant effects of drivers of ethnobiological knowledge (generalized mixed effect-models and structural equation modelling). The second set addresses limitations in the way these analyses are used by correcting for phylogenetic relatedness or accounting for social and family ties or spatial links (phylogenetic generalized least squares and social network analysis). The third set evaluates how well models generalize to data that were not used to develop the model. This approach, known as predictive modelling, has been used primarily to develop species distribution models, which create species range maps under current and future climate conditions. However, predictive modelling approaches can be integrated with many other kinds of statistical models, including all models discussed in this review.

(1) Generalized linear mixed-effect models (GLMMs)

Ethnobiological motivation – Mixed-effects models or hierarchical or multilevel models are popular statistical methods (Breslow & Clayton, 1993; Pinheiro & Bates, 2000; Dean & Nielson, 2007) widely adopted in ecology and evolution (Bolker et al., 2009; Zuur et al., 2009; Grueber et al., 2011). This class of models receives much interest because of its ability to handle unbalanced clustered, longitudinal, blocked design, non-normal, non-independent, and hierarchical or multilevel data (Pinheiro & Bates, 2000; Bolker et al., 2009). These models can also be used for descriptive, inferential, and predictive modelling (Gelman & Hill, 2007). Despite the popularity of multilevel modelling, its use in ethnobiological studies is very rare (Brandt et al., 2013; Guadagnini & Gravato, 2013; Torres-Avilez, de Medeiros & Albuquerque, 2016). For a theory-inspired and hypothesis-driven ethnobiology (Gaoue et al., 2017), mixed-effects models provide strong statistical inference and limit biases due to pseudo-replications and different forms of non-independence of observations.

Most ethnobiological studies collect data with inherent multilevel structure such as villages, plots, communities, and taxonomic groups. For example, individuals interviewed in villages are generally chosen non-randomly, or within social groups where it is impossible to ensure independence among villages, and also among interviewees. Here, the study design presents a hierarchy with multiple levels from the individual to the neighbourhood, village, community and district. Each level of these structured data may impact the information delivered by an individual. Analysis of this type of data is feasible with multilevel models. Also, different ethnobiological studies are conducted in the same communities but vary temporally (time-series data). These data cannot be considered as independent since they are all from the same communities (pseudo-replication). Such data are, however, interesting to understand the dynamics of ethnobiological knowledge over time. Thus, a statistical control for time becomes necessary in the analysis of such time-series data to reveal the effects of time on ethnobiological knowledge generation dynamics within a community.

Multilevel models, unlike classical analysis of variance, incorporate both fixed and random effects, with variance components that can be partitioned according to the structure of the data (Pinheiro & Bates, 2000). Fixed effects are “factors whose levels are experimentally determined or whose interest lies in the specific effects of each level, such as effects of covariates, differences among treatments and interactions” (Bolker et al., 2009, p. 128). Random effects are “factors whose levels are sampled from a larger population, or whose interest lies in the variation among them rather than the specific effects of each level” (Bolker et al., 2009, p. 128).

Application – One of the major hypotheses in ethnobiology predicts that traditional knowledge will increase with age and in some cases, women will have more local ecological knowledge than men (Gaoue et al., 2017). This is one of the most tested hypotheses in ethnobiology, often indirectly because the primary goal of most studies is to demonstrate rather than partition local knowledge. To test this hypothesis, we can use data on the age, gender and educational level of participants selected randomly in different villages within random regions of the study area. The response variable is binary (knows/does not know) and captures the probability that a participant knows a given medicinal plant. Age, gender, and education level are the fixed terms while the village, region and the participant selected represent random terms. If the intention is to test the same hypothesis in three climatic regions, then a given region becomes a fixed factor. For these two scenarios, we used the package glmmTMB (Brooks et al., 2017) in R to fit multilevel models (Fig. 1) since the response variable is binary and requires a generalized linear mixed-effects model (GLMM) with a binomial error structure (Crawley, 2013). We illustrate this and provide data and seed code to reproduce this analysis as Code R1 in Gaoue et al. (2021).
A direct implication of multilevel data is the need for random effects models. The variables which stand as levels or random effects are frequently left out of data analyses when they should be included as random effects. In such cases, traditional linear models are less adequate than multilevel models which incorporate random effects. Multilevel models integrate different structures for random effects such as individual, crossed and nested structures as well as various types of models accounting for different levels of complexity including a random intercept, random slope, and both random intercept and slope models. For example, a nested random effect could arise from establishing plots as nested in sites (Wall et al., 2019) or from genera nested within a given plant family (Bebber et al., 2007). Another example is related to testing the theory of non-random medicinal plant selection (Moerman, 1979; Ford & Gaoue, 2017; Robles Arias et al., 2020; Muleba et al., 2021). Data used to test this theory include the total number of species per family, nb_med_fam, as the response variable, the total number of species per family, nb_fam_total, as fixed term whereas the family, family_spp is the random term.

Another advantage of multilevel models for ethnobotanical data analysis is the wide range of complex data that can be handled. The dependent or response variables can be measurements, counts, binomial, multinomial, or ordinal. Multilevel models offer more parameters to specify highly complex families such as zero-inflated or one-inflated distributions. Many R packages are available to fit multilevel models including lme4 (Bates et al., 2015), nlme (Pinheiro & Bates, 2000), glmmTMB (Brooks et al., 2017), glmmADMB (Bolker et al., 2012), MCMCglmm (Hadfield, 2010) and brms (Bürkner, 2017).

(2) Structural equation modelling (SEM)

Ethnobiological motivation – Structural equation modelling (SEM) is a statistical method that can help elucidate the mechanisms underlying plant selection (Loken, 2009). SEM fits a network of hypotheses to data and is well known in the social sciences for its ability to handle complex models (Moreira et al., 2016). SEM has three advantages over traditional statistical methods (Grace et al., 2010). First, unlike most of methods that measure correlation, SEM can assess causation (Lefcheck, 2016). Second, it enables the measurement of errors for correlated variables that are inherent to social science. Finally, SEM can deal with latent variables, i.e. non-observable quantities or factors underlying observed variables that are inherent to social science (Tarka, 2018).
Below we detail three steps to analyse the drivers of plant selection for medicinal purposes using SEM and major theories in ethnobiology.

Application: building a SEM – First, we define a metamodel describing how plant secondary chemistry and life forms can mediate the effect of resource availability on plant species use values (Fig. 2). This model is based on logical predictions from the resource availability hypothesis and the ecological apparency hypothesis. The ecological apparency hypothesis predicts that plants that are easily found by herbivores (e.g. long-lived species) should invest heavily in quantitative defences (e.g. tannins) that are effective against all herbivores (Feeny, 1976). By contrast, plants that are difficult to locate (short-lived herbaceous plants) should invest smaller amounts of resources in qualitative defences (e.g. alkaloids) that are effective against all but specialist herbivores. Given that qualitative defences are expensive to produce, short-lived species will be less likely to be used for medicinal purposes than long-lived species. The resource availability hypothesis predicts that plant defensive capabilities are mediated by their capacity to replace lost tissue (Coley, Bryant & Chapin, 1983). Therefore, fast-growing plants should invest relatively little in defence and use mobile resources that can be replaced rapidly via senescing tissue. By contrast, slow-growing plants, characteristic of low-resource environments, should invest more in chemical defence because plant tissue is more costly to replace. Therefore, long-lived (slow-growing) species will likely be used for medicinal purposes more than short-lived (fast-growing) species.

Second, we estimate the path coefficients. Path coefficients are the strength of the relationship between variables. For example, we predict that species use value will increase with the concentration of secondary compounds ($a_{22}$) and with increasing complexity of life form ($a_{12}$) captured here using the proxy of plant height (Fig. 2). This can be implemented using various packages including piecewiseSEM (Lefcheck, 2016), lavaan (Wahren, 1966) or brms (Bürkner, 2017) in R (R Development Core Team, 2019). Finally, to test if there is strong support for the resource availability hypothesis versus the ecological apparency hypothesis, ethnobiologists could compare the path coefficients that represent these two models. If $|a_{21}*a_{22}| > |a_{12}|$, then one would conclude that there is stronger support for the resource availability hypothesis than the ecological apparency hypothesis in explaining the variation of species use value (Fig. 2).

Structural equation modelling can also be used to analyse the influence of age and gender on the dynamics of local ecological knowledge. Female gender, increasing age, illiteracy, and decreasing formal education are all positively correlated with medicinal plant knowledge (Voeks & Leon, 2004). Older people are expected to have more knowledge of useful plants (Souto & Ticktin, 2012). Women tend to know more species and their specific uses than men (Souto & Ticktin, 2012). The degree of illiteracy is higher in rural areas (Vandebroek & Balick, 2012; Reyes-Garcia et al., 2013a). From these relationships, ethnobiologists can first build a metamodel (Fig. 3A) and then estimate the path coefficients using R (Fig. 3B). One can then compare the estimated path coefficients to understand which pathway best explains the factors that shape plant knowledge. For example, the meta-model includes the following pathways: plant knowledge is shaped directly by people’s age ($b_{11}$) or gender ($b_{22}$), or indirectly through the mediated effect of literacy which can be affected by age ($b_{41}*b_{22}$) or gender ($b_{42}*b_{22}$) (Fig. 3A). Similarly, urbanization can directly reduce plant knowledge ($b_{12}$) due to limited availability of plants to interact with or via the mediated effect of increased literacy ($b_{21}*b_{22}$) which might be associated with a greater opportunity for schooling that might accompany urbanization (Fig. 3A).

Finally, we used the same data set to illustrate how SEM can be implemented in R (Fig. 4; Code R2 in Gaoue et al., 2021). We used a simplified version of the metamodel linking knowledge, age, gender, and degree of urbanization (Fig. 3A). Here knowledge is binary and represents the ability of a person to know or not know the importance of plants. Similarly, gender (male or female) is binary. The degree of urbanization is the distance between each participant and the closest large city. We used the package piecewiseSEM to fit the model (Fig. 4). Because our endogenous variables are binary, we used generalized linear models with binomial error distribution. The results of the model (Fig. 3B) show that local knowledge was strongly influenced by gender ($\beta = 0.615 \pm 1.078, P < 0.0001$), age ($\beta = 0.413 \pm 0.026, P < 0.0001$) and degree of urbanization ($\beta = 0.1655 \pm 0.0069, P < 0.05$). These results suggest that local knowledge increases with age; the older someone is, the more they know about the importance of plants. Moreover, people in rural areas had a better knowledge of local plants than people in large cities. We infer from the SEM that age was the stronger driver (Fig. 3B) since it has a larger path coefficient than the other two paths (the paths linking...
knowledge directly with urbanization and that linking knowledge with urbanization via a mediation effect of gender).

**3) Phylogenetic generalized least squares (PGLS)**

**Ethnobiological motivation** – All plant species used in traditional ethnomedicine are evolutionarily related. Some plant families will be over- or under-utilized for medicinal purposes due to shared evolutionary traits including the presence or absence of high concentrations of secondary plant compounds (Moerman, 1979; Heinrich & Verpoorte, 2012; Ford & Gaoue, 2017; Souza, Williamson & Hawkins, 2018). Indeed, it has been demonstrated that plant family is a strong predictor of species use values (Phillips & Gentry, 1993a). Although these factors have been acknowledged by ethnobiologists and likely play a significant role in driving plant use and selection among human societies, many analyses of data that include plant taxa have failed to use statistical models that account for species shared evolutionary history. In ethnobiology, general linear models are in widespread use for analyses that consider plant taxa. However, the independence of residuals is a critical assumption of these models (Quinn, Keough & Reid, 2002; Mundry, 2014) that is often overlooked or violated where plant taxa are considered (Revell, 2010), potentially making these approaches misleading because they fail to control for species shared evolutionary history. We caution that PGLS should be used by ethnobiologists where plant taxa are included as observations (Gaoue et al., 2017). Examples include investigations of cultural keystone species (Coe & Gaoue, 2020) and the utilitarian redundancy model (Coe & Gaoue, 2021), which we highlight below.

PGLS provides a mean to fit a linear regression to investigate the effect of a given predictor variable on a response variable while controlling for non-independence between observations due to shared phylogenetic history (Revell, 2010; Mundry, 2014). These models include different assumptions of character evolution (e.g. Brownian motion) and branch-length scaling parameters such as $\lambda$ (Pagel, 1999; Revell, 2010; Mundry, 2014), which are important considerations as they can affect the results, and thereby impact the reliability of PGLS analyses (Revell, 2010; Mundry, 2014) and statistical inference in ethnobiology.

**Application** – The first step in using a PGLS model for ethnobiological data is to create a phylogeny of the medicinal

---

**Fig 3.** (A) Metamodel showing how to test a direct and indirect influence of age, literacy, gender, and degree of urbanization (rural area) on medicinal knowledge. $b_{11}, b_{12}, b_{22}, b_{21}, b_{31}, b_{32}, b_{41}, b_{42}$ are the path coefficients and represent the strength of each relationship between the variables. + indicates a positive correlation between variables; − indicates a negative correlation. (B) Causal diagram presenting hypothesis supported by the data (see Code R2 in Gaoue et al., 2021). Path coefficients were obtained by running the piecewiseSEM model in R for the meta-model (A) as presented in Fig. 4. Dashed lines identify non-significant relationships and the red line highlights a negative relationship.

**Fig 4.** R syntax to implement structural equation modelling (SEM) of the metamodel presented in Fig. 3, to test hypotheses using ethnobiological data. This SEM was implemented using the R package piecewiseSEM which allows the inclusion of endogenous variables with non-normal residuals. In this specific case, we use binomial error for all endogenous variables that are binary.
species cited by participants or a phylogeny of the entire local flora, depending on the ethnobiological question of interest. Traditional approaches to generating a phylogeny requires DNA sequences, although this is time-consuming and such sequences often were unavailable for the plants of interest. However, recent technological advances and wide scale DNA sequencing projects and initiatives (e.g. the Barcode of Life project) have led to the availability of large numbers of DNA sequences in public repositories (e.g. GenBank/EBI), accelerating the assemblage of detailed phylodgensies. Mega-trees of thousands of species have been assembled (e.g. Zanne et al., 2014) that can be pruned to generate phylodgensies suitable for a given geography or research context. The development of R scripts and libraries such as the S. Phylomaker or V. PhylaMaker packages in R (Qian & Jin, 2016; Jin & Qian, 2019; R Development Core Team, 2019) have made this process easier and more accessible, thus allowing researchers to include phylogeny as a factor in ethnobotanical studies (Haris Saslis-Lagoudakis et al., 2014; Yessoufou et al., 2015; Lei et al., 2020). The mega-tree used by these R packages includes all plant families, approximately 10000 genera and 70000 species of vascular plants, enabling researchers to create phylogeny trees relevant to their study sites in which all plant families and the majority of genera are resolved (Jin & Qian, 2019). A phylogeny can also be generated using the online Phylomatic platform (https://phylodiversity.net/phylomatic/).

Coe & Gaoue (2021) provide an example of a phylogenetic tree created for 62 medicinal plant species from 33 families and 57 genera using the S. Phylomaker function (Qian & Jin, 2016) in R version 3.4.3 (R Development Core Team, 2019) (Fig. 5). It was constructed from a comprehensive phylogeny for vascular plants (Jin & Qian, 2019) consisting of 31389 tip labels and 31387 internal nodes. The resulting phylogenetic tree contains 62 tips and 59 internal nodes (Coe & Gaoue, 2021).

After the construction of a phylogenetic tree, PGLS models can be implemented in R using packages such as ape (Paradis et al., 2020). We recommend starting with a saturated model that includes all predictors followed by the development of subsequent nested model candidates created by sequentially removing one of the predictors. Use of the Akaike information criterion (AIC) will allow the selection of the best-fitting models (Crawley, 2013) by comparing models’ ΔAIC (the difference in AIC between each model), and identifying the model with the lowest AIC. In this approach, models with ΔAIC less than 2 are considered as the best models with the greatest explanatory power.

Coe & Gaoue (2021) discuss PGLS model selection and the importance of controlling for species’ shared evolutionary history using ethnobiological data. They investigated the applicability of the utilitarian redundancy model, an ethnobiological theory that provides a framework for understanding the effects of species’ shared evolutionary history, therapeutic redundancy, use values and local preferences on medicinal species use pressure (Gaoue et al., 2017; Coe & Gaoue, 2021). The study analysed ethnobiological data collected from 30 participants during fieldwork (interviews, focus groups and plant collections, participant observations) in the Shipibo-Konibo community of Paoyhan located in the Ucayali river region of the Peruvian Amazon. Of particular interest, Coe & Gaoue (2021) demonstrate how failure to control for shared evolutionary history between medicinal species used in ethnomedicinal contexts in the Ucayali river region of the Peruvian Amazon rainforest may result in misleading conclusions. While previous studies (Ferreira Júnior et al., 2012) of the utilitarian redundancy model demonstrated that preference by local people was a strong predictor of medicinal species use pressure, Coe & Gaoue (2021) demonstrate that when controlling for phylogenetic history in predictive models, the effect of local preference on medicinal species use pressure depends more on therapeutic redundancy. Further, the authors demonstrate that the absence of phylogenetic correction in predictive models led to an erroneous conclusion that local preference alone was a significant predictor of medicinal species use pressure (Table 1). We provide the data (see also Coe & Gaoue, 2021) and seed code to reproduce this analysis for PGLS and phylogenetic tree construction as Code R3 in Gaoue et al. (2021).

(4) Testing for a phylogenetic signal

Etnobiological motivation – Alternative approaches to PGLS include testing for a phylogenetic signal in a given ethnobiological variable. This test can reveal whether closely related species share similar values for a variable due to their shared evolutionary history. For example, one may ask if closely related species tend to have similar use value in a given community or village. This approach has been used to uncover cross-cultural convergence of medicinal plant use as well as the phylogenetic underpinning of plant therapeutic value (Saslis-Lagoudakis et al., 2012; Haris Saslis-Lagoudakis et al., 2014; Yessoufou et al., 2015; Teixidor-Toneu, Jordan & Hawkins, 2018; Lei et al., 2020). Where there is a phylogenetic signal in use value, phylogenetic information could be used to predict the use value of a species for which we lack such information or to group species in different use value categories. Thus, a species belonging to a ‘hot node’ of high use-value species on a phylogeny has a high probability of also having a high use value (Fig. 6).

The ethnobiological variables at hand dictate the options of analyses to use. For example, when dealing with a continuous variable (e.g. use value, redundancy, use pressure, etc.), testing for a phylogenetic signal can be done with the phylogen function of the R package phytools (Revell, 2012). Ethnobiologists can estimate Blomberg’s K value or Pagel’s λ, which range from 0 (no phylogenetic signal) to 1 (strong phylogenetic signal) (Pagel, 1999; Blomberg, Garland & Ives, 2003). For Pagel’s λ, the log likelihood ratio can be used to examine if λ is significantly greater than zero. To test for significance of the K test, values of the ethnobiological variables at hand can be reshuffled multiple times across the phylogeny to calculate whether the observed K value departs significantly or not from
random expectations. Both λ and K tests can be implemented using the phylosig function in the R package phylotools.

However, for binary variables, e.g. medicinal status (medicinal versus non medicinal), it is more appropriate to use the D statistic (Fritz & Purvis, 2010). This can be calculated using the R function phylo.d implemented in the package caper (Orme et al., 2014). A value of $D = 0$ is indicative of a phylogenetic signal as expected under a Brownian model; $D < 0$ indicates a strong signal (in the scenario shown in Fig. 6, medicinal status is likely to be highly conserved and thus to have a strong signal); and $D \geq 1$ indicates overdispersion (absence of a signal). The significance of the D value is estimated by comparison of the observed D with the distributions of scaled D from simulations.

The net relatedness index (NRI) and net taxon index (NTI) metrics from community phylogenetics (Webb et al., 2002)

Fig 5. Phylogeny of plant species cited for medicinal use in Coe & Gaoue (2021).
can also be used to test for a phylogenetic signal in a binary variable. Both NRI and NTI are used for sets of species, for example to investigate whether a set of medicinal plants in a community or village is phylogenetically more closely related than predicted at random (i.e. are medicinal plants a random subset of the total flora of a village?). One may also be interested in testing whether a set of plants, for example those with red sap, is more related than expected by chance. These metrics are very similar, with the only difference being that NRI assesses the phylogenetic signal deeper in the phylogeny while NTI estimates the signal towards the tips of that NRI assesses the phylogenetic signal deeper in the phylogeny.

Table 1. Comparison of results from phylogenetic generalized least-squared (PGLS) models conducted in R version 3.4.3 using the ape package version 5.0 (Paradis & Schliep, 2019) and a generalized linear model (GLM). Both models were used to test the effects of use values, species use preference and species therapeutic redundancy on the use pressure of medicinal plants from data in Coe & Gaoue (2021). The PGLS model controls for shared evolutionary history between the medicinal plant species cited by participants, whereas the GLM model includes no correction for phylogeny. Significant predictors are in bold.

|                      | Estimate | Std error | t value | P      | AIC   |
|----------------------|----------|-----------|---------|--------|-------|
| GLM                  |          |           |         |        |       |
| Intercept            | 9.569    | 3.415     | 2.802   | 0.007  | 550.84|
| Preference           | 9.761    | 5.785     | 3.620   | 0.0006 | 558.45|
| PGLS                 |          |           |         |        |       |
| Intercept            | -0.569   | 16.414    | -0.035  | 0.972  | 558.45|
| Preference           | 9.863    | 5.319     | 1.854   | 0.069  |       |
| Redundancy           | 13.254   | 3.287     | 4.032   | 0.0002 |       |
| Preference × redundancy | -9.267   | 3.565     | -2.599  | 0.012  |       |

Fig. 6. A hypothetical phylogeny illustrating phylogenetic signal in medicinal status and plant use value (UV). Clade A is particularly rich in medicinal plants compared to clade B, suggesting that there is a phylogenetic signal in medicinal status. Since medicinal status is a binary variable, the D statistic of Fritz & Purvis (2010) is suitable to test this signal. Similarly, clad A tends to harbour plants with higher UV than clad B. For this continuous variable, a signal test such as Pagel’s λ or Blomberg’s K is more appropriate.

(5) Social network analysis and exponential random graph modelling (ERGM)

Ethnobiological motivation – Social network theory states that the flow of information through social networks is determined by
the configuration of human relationships comprising the network (Granovetter, 1973). In the context of ethnobiology, social network theory hypothesizes that a person’s knowledge is influenced by their location in their social network, the connectivity of that network, and the knowledge of other people in that network (Gaoue et al., 2017). Social networks can be analysed by representing people as nodes connected by ties or links which represent sharing of knowledge or biological materials (Barnes et al., 2017; Kluger et al., 2020). Social network analysis has been applied in ethnobiology to test two major groups of predictions. The first is based on the hypothesis that connections between nodes in a network are associated with how much knowledge each person has and how evenly distributed knowledge is across the network (Janssen et al., 2006). The second is based on the hypothesis that network structural patterns are the result of sociocultural

**Fig 7.** R syntax used to test for a phylogenetic signal in ethnobiological variables from Coe & Gaoue (2021). This syntax tests for a phylogenetic signal in plant species use value (UV), therapeutic redundancy and use pressure using Blomberg’s $K$ test, Pagel’s $\lambda$, net relatedness index (NRI) and net taxon index (NTI).
processes which affect the formation of knowledge-sharing relationships (Lusher, Koskinen & Robins, 2013).

(a) Multivariate regression using social network connectivity to predict individual knowledge and knowledge distribution of networks

At the level of individuals, nodes with greater connectivity (e.g. degree centrality) are predicted to have more knowledge (e.g. the number of plant uses known) because well-connected nodes have more nodes they can learn from. Several node-level predictions have been supported in previous ethnobiological studies (Hopkins, 2011; Calvet-Mir et al., 2012; Reyes-García et al., 2013b). However, we are unaware of any existing ethnobiological studies that test these predictions at the network level. An example network-level prediction is that networks with greater connectivity (e.g. higher density, lower diameter) have less variation in knowledge (e.g. the standard error of the number of plant uses known by each person in the network) because connectivity increases the spread of information. One potential reason for the lack of network-level tests is the effort required to collect a data set with many networks. While this is possible in a single study (e.g. Reyes-García et al., 2013b), meta-analysis may be a simpler option.

Connectivity metrics can be calculated at node level (e.g. degree, broker, and betweenness centrality) or network level (e.g. density, centralization, clustering, modularity, and diameter). Multivariate regression techniques (such as GLMM) are ideal statistical tests for these predictions, since they allow multiple network metrics to be included, as well as potentially confounding variables. Confounding variables at node level include age, gender, education, and ethnicity (Hopkins, 2011; Calvet-Mir et al., 2012; Reyes-García et al., 2013b; Díaz-Reviriego et al., 2016), while variables such as biomedical access, distance from metropolitan areas, and plant availability may confound network-level analyses (Vandebroek et al., 2004; Cámara-Leret et al., 2014).

(b) ERGMs using social learning processes to predict network configuration

This second group of predictions tests mechanistic explanations for how nodes become connected and why some networks have more connectivity than others. Although there are many mechanisms of social learning, examples include homophily (learning from similar people, such as those of the same age, same family, or same gender), prestige (learning from people that are perceived as successful or knowledgeable), and rarity (learning from people with unusual information) (Kendal et al., 2018). Predictions from causal mechanisms can involve three types of variables: configuration variables, node connectivity variables, and node pair variables. For each type of variable, sample predictions based on the mechanisms of homophily, prestige, and rarity are provided in Table 2. Configuration variables test whether or not the configuration is more or less common in the observed network than in a randomly generated network with the same number of nodes and edges. Node connectivity variables are predicted to correlate with the number of edges corresponding to each node. Node pair variables are predicted to correlate with whether there is a link between each possible pair of nodes in the network.

Exponential random graph models (ERGMs) are the primary methods of testing these predictions, because a single ERGM can include configuration variables, node connectivity variables, and node pair variables, which controls for the confounding effects of each variable type (Lusher et al., 2013). However, several studies have used alternative methods to test predictions using only node pair variables, with interesting results (Salali et al., 2016; de Brito et al., 2019; Santoro, Chaves & Albuquerque, 2020). Several predictions involving all three types of variables have been supported in previous ethnobiological studies (Henrich & Broesch, 2011; Labeyrie et al., 2015; Thomas & Caillon, 2016; Bond & Gaoue, 2020).

(c) Future of social network analysis

A weakness of most existing ethnobiological social network analyses is that they fail to explore how networks and learning
processes evolve over time (but see Violon, Thomas & Garine, 2016; Santoro et al., 2020). Although diachronic methods such as temporal ERGM exist, they have yet to be applied in ethnobiological studies (Kluger et al., 2020). Another emerging technique is network analysis for linked social and ecological systems (Barnes et al., 2017; Kluger et al., 2020). Although bipartite socio-ecological networks have provided information about the resilience of plant-use networks (Cámara-Leret, Fortuna & Bascompte, 2019; Castiñeira Latorre et al., 2020), these approaches have yet to be used to test predictions about integrated socio-ecological networks of ethnobiological knowledge of useful plants.

### Application: constructing a network, calculating connectivity metrics, and running an ERGM

The first step in social network analysis is to establish predictions based on social learning mechanisms and the sociocultural context of the study (for examples see Table 2). The second step is to define criteria for who will be included in the network; this can be challenging because social network boundaries are not always clear (Laumann, Marsden & Prensky, 1991). The third is to collect data from at least 90% of the network on who they share information or biological material with (Kolaczyk & Krivitsky, 2015), and additional data required to test each prediction (e.g. knowledge.
demographics, family relationships). The fourth step is the creation of an adjacency matrix or edgelist csv file and a node-attribute csv file (Hoffman, 2020). These can be loaded into R using the sna (Butts, 2008b), network (Butts, 2008a), or igraph (Csardi & Nepusz, 2006) packages to create a network and add attributes to nodes in network. Finally, connectivity metrics are calculated using the sna (Butts, 2008b) or igraph (Csardi & Nepusz, 2006) packages. ERGMs are commonly
fitted in the \texttt{arm} R package, then evaluated using goodness-of-fit (Hunter et al., 2008) and degeneracy (Goodreau et al., 2008). Like GLMM, SEM, and PGLS, ERGMs can be compared using the Akaike information criterion (AIC) (Crawley, 2013) for each model. To guide use of social network analysis and ERGM for ethnobiological data we provide data and seed code (see Fig. 9) as Code R3 in Gaoue et al. (2021). For more information, the reader is referred to online tutorials (e.g. Morris et al., 2019; Hoffman, 2020).

(6) Species distribution modelling (SDM)

Ethnobiological motivation – The resource availability hypothesis in ethnobiology predicts that plants are used medicinally because they are locally accessible or have a higher abundance (Voeks, 2004; Lucena, Lima Araújo & Albuquerque, 2007; Hart et al., 2017). One of the most accessible spatial tools that could be used to test this hypothesis is the species distribution model (SDM). While SDMs do not quantify abundance, they can be used to estimate a maximum likelihood surface of species occurrences using environmental data and presence or absence of an organism. For example, SDMs allow for correlation analysis to be performed using environmental variables and species occurrence data. In this context, these data can be used to identify areas of climatic suitability or unsuitability for a species currently or projected into the future under different climate change scenarios. Quantifications of environmental niches from SDMs have multiple uses in ecology and conservation, including predicting distances between patches of metapopulations, species assemblage modelling to estimate diversity and community composition (Guisan & Thuiller, 2005), and hypothesis testing (Anderson, Peterson & Gómez-Laverde, 2002). Combined SDMs can be used not only to locate new populations or suitable habitats for potential relocations of culturally important plants, but also as a flexible and adaptable tool to test ethnobotanical hypotheses such as the resource availability hypothesis or ecological apparentness hypothesis in novel ways (Dai et al., 2017; Bond et al., 2019).

The temporal component of SDMs can be used to address drivers of change in biodiversity of ecological communities in an ethnobiological context. For example, a framework has been proposed to amalgamate species occurrences, environmental and geographic space, and time from heterogeneous data sources to infer essential biodiversity structure and function (Jetz et al., 2019). This conceptual framework could be used to support or refute hypotheses regarding the impacts of plant selection by human societies. In particular, the cultural keystone species theory suggests that there are essential plant species that, among other elements, are unique and irreplaceable to human communities (Garibaldi & Turner, 2004). It is challenging to discriminate between the relative values of different species to a community, but this framework can help simplify this process as well as reveal quantitative insights about the roles people play in ecological systems.

Optimal foraging theory, rooted in ecology, predicts there is a trade-off between plant value and effort spent foraging or processing related (time and distance) (Sih & Christensen, 2001). One of the greatest limitations to testing optimal foraging theory in ethnobiology is the lack of geographic diversity in existing spatial analyses (Gaoue et al., 2017). With an increase in remote-sensing data availability and quality, coupled with new spatial analysis methodologies, the potential for testing across different regional and cultural contexts is expanding. For example, when interviewing community members to estimate distances travelled to harvest plants, distance becomes a subjective measure that varies with plant, accessibility, abundance, and movement path. While distances travelled to harvest plants can be measured in the field, researchers can model travel data using Lévy distributions (Rhee et al., 2011) or custom-made movement models linking environmental variables to paths (McClimontock et al., 2012). These models can then be tested for informative quantities such as straightness (McCarthy et al., 2010), first-passage time (Fauchald & Tveraa, 2003), or more complex fractal dimensions of movement times (Namazi, 2017). These quantitative techniques could be useful and versatile tools for testing optimal foraging theory in different spatiotemporal and cultural contexts.

It is also common for traditional ecological knowledge (Berkes, Colding & Folke, 2000) to encompass correlations between plant locations and plant use associated with environmental or abiotic variables. For example, an ethnobiological survey of indigenous communities in Nepal found an increase in species richness of medicinal plants gathered and used at higher altitudes (Kunwar & Bussmann, 2008). There also are ethnobiological hypotheses linked to plant phytochemistry, such as the diversification hypothesis which predicts that local communities introduce species to increase the diversity of medicines available to treat various ailments (Alencar et al., 2010). Incorporating a spatial component into such questions allows researchers to investigate them using SDM: for example, how does the medicinal quality of plant species vary across space? A potential tool that could be used to address this question is genomic spatial modelling, which quantifies population-level biodiversity across landscapes (Fitzpatrick & Keller, 2015). Entire genomes of plant species have been sequenced and analysed for potentially adaptive single nucleotide polymorphisms (SNPs) linked to environmental variables. For SNPs linked to plant phytochemistry, spatially or environmentally correlated medicinal plant quality as a function of genomic similarity can be projected and tested.

Conceptualizing and testing how, why, and which resource values of communities are distributed in geographic and environmental space can also be facilitated through adaptation of a suite of models by inVEST (Sharp et al., 2018). If medicinal plants are contextually designated as provisional ecosystem services, questions regarding the balance between population supply, human value, and human consumption can be addressed. InVEST could be particularly useful for accounting for the complexity of
interaction dynamics and economics in ethnobiological systems, such as spatial variation of medicinal plant value as a function of culturally important sites or other community-specific cultural metrics.

(7) Predictive modelling

Ethnobiological motivation – One of the greatest challenges across biological and cultural science is identifying generalizable principles of socio-ecological systems and understanding their predictive power (Linguist et al., 2016; Hofman, Sharma & Watts, 2017; Magliocta et al., 2018; Forey & Linguist, 2020). While these insights can be gathered through reviews or meta-analyses, they can also be generated using predictive models (Kuhn & Johnson, 2013). Predictive models predict future or unseen events by understanding the patterns in existing data. By contrast, descriptive models explain the nature and cause of patterns in existing data.

Although the techniques presented herein are mostly descriptive models (except for SDMs), any model can all be adapted to produce predictive insights. Key to predictive models is train/test splitting, which randomly splits data into two sets using predefined percentages (Kuhn & Johnson, 2013). The larger data set (the training set) is used to train the model, while the smaller data set (the testing set) evaluates how well the model generalizes to predict new data. To reduce possible bias in the testing data, the train/test split is typically performed in multiple rounds so that each part of the data is used to test the model in at least one round. The evaluation metrics from each round then are combined (e.g. averaged); this robust evaluation method is called cross validation (Kuhn & Johnson, 2013).

Comparing model performance on the training and testing sets ensures that predictive models avoid two problems encountered by all statistical models: overfitting and the inclusion of non-causal variables (Lever, Krzywinski & Altman, 2016). Predictive models with evaluation statistics that are much better for training data than testing data indicate overfitting, i.e. a model that matches the training data so precisely that it fails to capture the overall data trend in a generalizable way (Quinn et al., 2002). Overfitted predictive models can be adjusted and re-run; the model with the best test data evaluation statistics representing the best (most generalizable) model. Predictive models can reduce overfitting in three ways: early stopping reduces the number of model iterations, regularization changes the number and effect size of predictor variables, and ensembling aggregates models with different levels of complexity to identify the overall data trend (Kuhn & Johnson, 2013).

The advantages of predictive modelling have led to breakthroughs in ecology, psychology, linguistics, molecular biology, and genetics (Jung, 2016; Weinstein et al., 2016, 2020; Korotcov et al., 2017; Demirci et al., 2018; Jäger, 2018). However, we are unaware of any ethnobiological applications of predictive modelling, except for SDM. One possible reason for the lack of its widespread use in ethnobiology is that large sample sizes are required to ensure model accuracy and increase statistical power when using a train/test split – a minimum of several hundred samples, but ideally several thousand samples (Varoquaux, 2018) are required for robust predictive models. A second potential reason may be that data scientists who are the primary users and developers of predictive modelling software packages primarily use Python instead of R, which tends to be used by academic scientists (Ozgur et al., 2017).

Application – Although predictive models are not currently used in ethnobiology (aside from SDM), they have many potential applications. The most common predictive models are regression models (both linear and non-linear) and decision tree regression (Kuhn & Johnson, 2013), for which train/test splits and cross validation are easily completed in R using the caret package (Kuhn, 2008). Predictive models can also be applied to data with phylogenetic (e.g. PGLS), grouping (e.g. GLMM), or spatial (e.g. SDM) structures, although the train/test split must be carefully planned to account for each dependence structure (Roberts et al., 2017). Predictive models can also be integrated with causal inference models (e.g. SEM) (Jacobucci, Grimm & McArthur, 2016; Kyriazos, 2018) or social network analysis (e.g. ERGM) (Shojaie, 2013; Pinter & Eisenstein, 2020). To illustrate the use of train/test splitting and cross validation for ethnobiological data we provide data and a seed code (see Fig. 10) as Code R6 in Gaoue et al. (2021). For more information, readers are referred to online tutorials (e.g. Dalpiaz, 2020).

IV. DISCUSSION

(1) Synthesis of new methodological approaches to hypothesis-driven ethnobiology

Herein we present a set of advanced statistical modelling approaches and quantitative analysis tools that can be used in hypothesis-driven ethnobiology. Although theories and hypotheses in ethnobiology can be tested in various ways, some of these methods provide more detail or statistical power than others. For example, although GLMMs can control for some forms of non-independence of observations, they are not able to test the power of indirect or mediating effects to the extent that is possible with SEM. For example, mixed-effect models can test the hypothesis that local knowledge or use values increase with participant age and with years of residence in a region but are different for men and women. However, because age and residence time can be strongly correlated, the latter might need to be removed from the model; even if residence time is included, mixed-effect models will not be able to demonstrate a causal link between age and residence time. By contrast, using SEM provides causal evidence for how the effect of age on local knowledge or use values is mediated by residence time, and thus that age may not be relevant if it is unrelated to residence time and possibly to years of experience interacting with the local environment.
The use of SEM thus is powerful and should be embraced in ethnobiology to disentangle the effects of various drivers of knowledge but also to provide a mechanistic understanding of how these drivers influence plant knowledge and how that translates into the differential use of plants and ecological systems. Simple SEM using the lavaan library (Wahren, 1966) does not account for non-independence in observations that mixed-effect models effectively address. However more advanced piecewise SEM (Lefcheck, 2016) includes random effects and temporal or spatial correlation.
structures, all of which are likely to yield more informative results.

A unique advantage of SEM is the possibility to test and compare the strength of alternative pathways for the same global model. For example, the resource availability and ecological apparency hypotheses offer alternative explanations of why some plant lineages may produce a greater quality of medicinal products than others. The ecological apparency hypothesis proposes that more apparent species such as trees would produce a higher quality of medicinal products and consequently local people will disproportionately select woody species for medicinal purposes. The resource availability hypothesis also suggests that woody species are more likely to be selected but here the growth rate is the causal mechanism. SEM could be used to compare these pathways to identify which one has the greatest support.

The choice of which model to use requires balancing the study purpose and data availability. For example, although GLMMs can control for non-independent or hierarchical data (such as person and village or plant genus and family), they cannot test hypotheses about variables that are used as random effects. Testing hypotheses involving plant evolutionary relationships or human social relationships requires using community phylogenetic and social network analysis techniques, respectively. However, if genetic or social network data are not available, analysis will be limited to GLMMs.

SDM and other spatial analysis tools are rarely used in ethnobiology (but see Bond et al., 2019). However, SDM can play an important role in spatially quantifying resource knowledge and use, a step that is critical for priority planning and strategic resources management. In addition, to test the resource availability hypothesis or the ecological apparency hypothesis, estimating plant availability or apparency is a labour-intensive process using classical ecological surveying methods such as plot-based or transect-based density or abundance measures. The use of SDM can allow the rapid generation of estimates allowing for a more general test of such hypotheses (Dai et al., 2017).

(2) A call to embrace quantitative modelling without sacrificing qualitative approaches which establish context

The definition of ethnobiology as the study of interactions between people, culture, and organisms has perhaps promoted a culture of simple documentation of these interactions. Several authors have critiqued this approach, emphasizing that although documenting interactions is valuable, understanding the processes and mechanisms of plant-human interactions will be essential to finding solutions for the Anthropocene crises (Bennett, 2005; Albuquerque & Hanazaki, 2009; Gaoue et al., 2017). We agree with this view and suggest that while qualitative approaches are an important part of ethnobiology, they must not be considered the primary focus of ethnobiological research. This evolution of ethnobiology is similar to the development of other disciplines, such as community ecology (Lawton, 1999; Turchin, 2001; Berryman, 2003). Early studies in ethnobiology focused on documenting patterns of plant/animal–people interactions. Similarly, early studies in community ecology primarily focused on describing patterns of variation in vegetation to provide support for the organismic concept of communities (Clements, 1936) or the individualist continuum concept (Gleason, 1926). Current studies of community ecology instead focus on processes driving these patterns and understanding what factors promote species coexistence and therefore biological diversity (Vellend, 2010).

Ethnobiology must follow a similar progression with an increasing focus on studying not just the patterns but also the processes underpinning plant/animal–people interactions. This can be achieved by developing clear hypotheses; defining such theoretical arguments is vital to avoiding similar pitfalls to that of community ecology which resulted in Lawton (1999) suggesting that the discipline lacked a unifying conceptual framework. Currently, most ethnobotanical studies are collections of case-specific results without guiding principles. As a result, they fail to provide appropriate visualization in figures, focusing instead on text and tables. Embracing theory-inspired and hypothesis-driven approaches will advance ethnobiology, increasing its status and recognition among other disciplines (McClatchey et al., 1999; Gaoue et al., 2017; Gonçalves-Souza et al., 2019). Testing hypotheses using advanced statistical modelling techniques will also lead to an increasing use of figures to illustrate the results. Ethnobiology must embrace these quantitative tools without sacrificing the use of qualitative methods of inquiry which provide an invaluable context to quantitative results. Consistent with the Nagoya Protocol (Talaat, 2013; Teran, 2016), it is important to reiterate calls from other authors (Bussmann, 2019; Golan et al., 2019) for ethnobiologists to follow ethical protocols, obtain institutional review board (IRB) approval and undergo appropriate training prior to data collection. Under these principles, ethnobiological data collection must ensure that prior consent is obtained before interviews and should guarantee that the locations of culturally sensitive knowledge and culturally important species are protected from public access and that participant names are anonymized to ensure confidentiality and the protection of indigenous rights during ethnobiological data collection (Vandebroek, 2017; McCune, 2018; Mednicewi, 2018).

(3) Training in quantitative skills for ethnobiologists

Most statistical modelling techniques presented here are rarely included in biometry or biostatistics graduate courses taught in natural science departments (Touchon & McCoy, 2016). In addition, many ethnobiology programs have closed, suggesting a long-term decline of researchers in this field and fewer new graduate students. It thus may be challenging for students to gain access to necessary statistical courses and we advocate a creative approach to bringing this knowledge into the mainstream of ethnobiology.
Students can search for these courses in specialized departments at their universities, and training workshops focused on teaching these skills should be encouraged by ethnobiological professional societies. As a result of Covid-19, many workshops and courses are now available online, increasing access to all (Apostol, 2020; Sikora et al., 2020; Vandebroek et al., 2020) and there are many free tutorials and code snippets available online for R and Python. Learning these techniques may require creativity but their implementation has the potential to change the course of ethnobiology.

The role of mentoring and advising will be critically important to the future of ethnobiological research, both to ensure relevance and the use of approaches grounded in tradition. We urge advisors to encourage and support their students to seek and participate in courses and workshops that help build their quantitative skills. This will also benefit students in terms of their employability (Beans, 2018).

(4) Challenges for implementing hypothesis-driven methods in ethnobiology

In this review, we have presented examples of how methodological advances can be applied to test hypothesis-driven predictions in ethnobiology. Moving the ethnobiology field forward from simply describing plants, animals and their uses to a more in-depth hypothesis-testing discipline requires a paradigm shift amongst ethnobiologists to understand better how people interact with their environment. Testing these theories comes with additional challenges, particularly with data availability (Kindsvater et al., 2018) and the decline in taxonomic expertise (Terlizzi et al., 2003; Costello, May & Stork, 2013). Ethnobiological data were historically scattered across different repositories, research institutions, publications, reports, and personal research databases, making wider access challenging. Recent progress in the creation of centralized online databases such as Évé (Souza & Hawkins, 2020) will facilitate the use of large-scale data sets and hypothesis-driven meta-analyses. Additionally, new avenues should be explored to collect data, including emerging technologies such as machine learning, remote sensing, and citizen science.

Most ethnobiological studies include large lists of plant or animal species. Accurate taxonomic identification will be directly affected by the decline in numbers of trained taxonomists. Clearly, inaccurate taxonomy will lead to erroneous conclusions even with the best statistical model. The declining numbers of trained taxonomists is primarily a result of the lack of priority given to the study of taxonomy and systematics in many university curricula, due in part to insufficient funding, and the increasing limitations of national and international legislation, which hinders taxonomic research explorations (Gaston & May, 1992). We highlight the importance of mentoring and training of taxonomists, especially from less developed countries, and of infrastructural development in taxonomy in developed countries with large collection of specimens (e.g. Royal Botanic Gardens, Kew, Natural History Museum, Missouri Botanical Garden, etc.) that still have practicing taxonomists.

V. CONCLUSIONS

(1) Ethnobiology is the study of people’s interactions with the organisms in their environment. This includes both plants (ethnobotany) and animals (ethnozoology) but also local people’s understanding of the ecological patterns and processes in their environment (ethnecology) and how they develop traditional management strategies (Berkes et al., 2000; Gadgil et al., 2001) involving collective actions that arise from self-organized patterns (Ostrom, 2000) or in response to independent outside forces (Ekin, 1988, 2002).

(2) The discipline has long been challenged to develop clear lines of investigation and theoretical foundations (Albuquerque & Hanazaki, 2009). This will be necessary for the advancement of this discipline and for the creation of educational and professional opportunities which are essential to its survival (Gaoue et al., 2017). Ethnobiology has evolved from simple descriptions of plant/animal use and knowledge to investigating how and why local people select the plants that they use. Recent developments include suggestions for theoretical investigations (Gaoue et al., 2017; Albuquerque et al., 2019a, 2019b) and empirical testing (Lucena et al., 2012; Nascimento et al., 2015; Ford & Gaoue, 2017; Hart et al., 2017; Seyler et al., 2019; Bond & Gaoue, 2020; Coe & Gaoue, 2020; Robles Arias et al., 2020; Muleba et al., 2021).

(3) In this review, we synthesize the advanced statistical methods with which ethnobiologists can test theories and hypotheses, and propose that ethnobiologists embrace these advanced methods which are currently rarely used. Structural equation modelling, social network analysis and phylogenetics methods all provide opportunities to investigate new research questions. This will represent a step towards reconciling ethnobiology with cultural evolution, mathematical sociology, and community ecology.

ACKNOWLEDGEMENTS, AUTHOR CONContributions AND DATA AND CODE AVAILABILITY

Acknowledgements: This paper was inspired by discussions with students who have taken the “Theories and Methods in Ethnobiology” course taught by the lead author at the University of Tennessee Knoxville, University of Hawai’i at Mānoa and various universities in Africa. O.G.G. was supported by funds from the University of Tennessee Knoxville. We thank two anonymous reviewers and
VI. REFERENCES

Alquéreque, U. P. (2009). Quantitative ethnobotany or quantification in ethnobotany? Ethnobotany Research and Applications, 7, 1–3.

Alquéreque, U. P., Cruz da Cunha, L. V. F., de Lucena, R. F. P. & Alves, R. R. N. (2014). Methods and Techniques in Ethnobotany and Ethnography. Springer, New York. https://doi.org/10.1007/978-1-4614-4836-7.

Alquéreque, U. P. & Hanazaki, N. (2009). Five problems in current ethnobotanical research—and some suggestions for strengthening them. Human Ecology 37, 653–661. https://doi.org/10.1007/s10745-009-9259-9.

Alquéreque, U. P., de Mesquita, P. M., Ferreira Júnior, W. S., da Silva, T. C., da Silva, R. R. V. & Gonçalves-Souza, T. (2015a). Social-ecological theory of maximization: basic concepts and two initial models. Biological Theory 14, 73–85. https://doi.org/10.1007/s13752-019-00316-8.

Alquéreque, U. P., Nascimento, A. L. B., Soldati, G. T., Freitas, I. S., Cardoso, J. L. A., Herrell, J. A., Hanazaki, N., de Mesquita, P. M., da Silva, R. R. V., Lidwinski, R. H., Ferreira Júnior, W. S. & Revés-García, V. (2016b). Ten important questions/issues for ethnobotanical research. Acta Botanica Brasilica 33, 376–383. http://dx.doi.org/10.1590/0102-33062016ab3303.

Alquéreque, U. P., Soldati, G. T., Serati, S. S., Ramos, M. A. de, Sai, J. C. & de Souza, L. C. (2011). The use of plants in the medical system of the Falá-0 people (NE Brazil): A perspective on age and gender. Journal of Ethnopharmacology 133, 866–873. https://doi.org/10.1016/j.jethph.2010.11.021.

Alexandre, N. L., de Sousa Araújo, T. A., Morism, E. L. C., Alquéreque, U. P., Leal Alencar, N., de Sousa Araújo, T. A., de Morism, E. L. C. & de Alquéreque, U. P. (2010). The inclusion and selection of medicinal plants in traditional pharmacopoeias-evidence in support of the diversification hypothesis. Economic Botany 64, 68–79. https://doi.org/10.1007/s12231-009-9104-5.

Alexanders, M. N. (1996). Selected Guidelines for Ethnobotanical Research: A Field Manual. The New York Botanical Garden, Bronx.

Aleixo, C. F. C. B. de Lima, E. Silva, T. C., de Morism, E. L. C., Maia, M. B. D. S. & Alquéreque, U. P. (2005). Life strategy and chemical composition as predictors of the selection of medicinal plants from the caatinga (Northeast Brazil). Journal of Arid Environments 62(1), 127–142. https://doi.org/10.1016/j.jaridenv.2004.09.020.

Anderson, R. P., Peterson, A. T. & Gómez-Laverde, M. (2002). Using niche-based GIS modeling to test geographic predictions of competitive exclusion and competitive release in Southern American potato mice. Oikos 98, 5–36. https://doi.org/10.1034/j.1600-0706.2002.011-090116.x.

Apostoli, A.-C. (2020). Students’ perceptions about the impact of COVID-19 on learning process. A sociological approach. Technum Social Sciences Journal 9(1), 495–500.

Barnes, M. L., Bodin, Ö., Guerreiro, A. M., McAllister, R. J. J., Alyea, R. R. M. & Robbins, G. (2017). The social structural foundations of adaptation and transformation in social and ecological systems. Ecology and Society 22, art16. https://doi.org/10.5751/ES-07969-220416.

Butts, C. T. (2008a). Network: a package for managing relational data in R. Journal of Statistical Software 24, 1–36.

Butts, C. T. (2008b). Social network analysis with sna. Journal of Statistical Software 24, 1–51.

Clayton, D. G. (2017). Brms: an R package for Bayesian multilevel models using Stan. Journal of Statistical Software 74, 1–23.

Crome, O. G. & Gaoue, O. G. (2020). Making friends in the network and the structural patterns of interaction. Ecological Research 35, 29–42.

Castiñeira Latorre, E., Canavero, A. & Arim, M. (2020). Ethnobotanical knowledge complexity in a conservation area of northern Uruguay: interlocutors-medical plant network and the structural patterns of interaction. Economic Botany 74, 195–206.

Clements, F. E. (1936). Nature and structure of the climax. The Journal of Ecology 24, 252–284.

Coe, M. A., & Gaoue, O. G. (2020). Most cultural importance indices do not predict species’ cultural keystones status. Human Ecology 48(6), 721–732. https://doi.org/10.1007/s10745-020-00192-y.

Coe, M. A. & Gaoue, O. G. (2021). Phylogeography explains why less therapeutically redundant plant species are not necessarily facing greater use pressure. People and Nature 1–12. https://doi.org/10.1002/pnan.10216.

Colley, P. D. P., Bryant, J. J. P. & Chapin, F. S. (1985). Resource availability and plant antiherbivore defense. Science 230, 895–899.

Cornwell, R. K., Mao, C. X. & Chang, J. (2008). Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology 89, 2771–2772.
Kluger, L., Kenny, D. A., McClatchey, W. C., Kolaczyk, E., Lever, J. (2016). Interpretation of GW-degree estimates in ERGMs (poster).

Lusher, D., Koskinen, J., & Robins, G. (eds) (2013).

Martin, G. J., Loken, E., Kendal, R. L., Lefcheck, J. S. (1999). Are there general laws in ecology?

Biological Reviews 2302 Orou G. Gaoue et al.

Kembel, S. W., Lucena, R. F. P., Ackerly, D. D., Dell’Angelo, J., Liao, Y., Verburg, P. H., Avis, A. C., Fonseca, B. G., Mateus, J. A. (2012). The ecological apparency hypothesis and the importance of useful plants in rural communities from northeastern Brazil: an empirical test in a tropical coastal environment. Journal of Ethnobiology Letters, 1, 347–350.

Paradies, J. C. & Bates, D. M. (2000). Mix-Effects Models in S and S-PLUS. Springer-Verlag, New York.

Pinter, Y., & Eisenstein, J. (2020). Predicting semantic relations using global graph properties. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2019, 1741–1751.

Poé, M. R., Norman, K. C. & Levin, P. S. (2014). Cultural dimensions of socioecological systems: key connections and guiding principles for conservation in coastal environments. Conservation Letters 7, 166–175.

Prell, G. (2012). Social Network Analysis: History, Theory, and Methodology. Sage, London.

Quinn, G. P., Keough, J. M. & Reid, J. R. (2002). Experimental design and data analysis for biologists. In Austrail Ecological Fourth Edition. Cambridge University Press, Cambridge.

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

Revell, L. J. (2010). Phylogenetic signal and linear regression on species data. Methods in Ecology and Evolution 1, 319–329.

Revell, L. J. (2012). Phytools: an R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3, 217–223.

Reyes-García, V., Guzmé, M., Luiz, A. C., Panéque-Gálvez, J., Maciá, M. J., Orta-Martínez, M., Pino, J., & Rubio-Campos, X. (2013a). Evidence of traditional knowledge-loss among a contemporary indigenous society. Evolution and Human Behavior 34, 249–257.
Reyes-García, V., Molina, J. L., Calvet-Mir, L., Aichnuto-Mata, L., Lastra, J. R., Ontíllera, R., Pareda, M., Parado-de-Santayana, M., Ricart, M., Vallés, J. & Garnatjé, T. (2013b). “Terrauna guadal:” germplasm exchange networks and agroecological knowledge among home gardeners in the Rurban Permit, Spain. *Etnoflora and Etnodiversión*, 9, 53.

Rhee, I., Shin, M., Hong, S., Lee, K., Kim, S. & Chong, S. (2011). On the Levvy walk nature of human mobility. *IEEE/ACM Transactions on Networking*, 19, 630–643.

Roberts, D. R., Bain, V., Ciuti, S., Boys, M. S., Elith, J., Guillemard-Abotts, C., Haeneynten, S., Lahoz-Monfort, J. J., Schorrr, R., Thullier, W., Warton, D. I., Wintle, B. A., Hartig, F. & Dormann, C. F. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylegetic structure. *Ecography*, 40, 913–929.

Robles Arias, D. M., Cevallos, D., Gaoue, O. G., Fadiman, M. G. & Hinde, T. (2020). Non-random medicinal plants selection in the Kialwe community of the Ecuadorian Amazon. *Journal of Ethnobiology and Ethnomedicine*, 16, 112220.

Romano, J. L. & Kromrey, J. D. (2009). What are the consequences if the assumption of independent observations is violated in reliaability generalization meta-analysis studies? *Educational and Psychological Measurement*, 69, 404–420.

Salali, G. D., Chaudhary, N., Thompson, J., Grace, O. M., van der Burgt, X. M., Duble, M., Page, A. E., Smith, D., Lewis, J., Mace, R., Vincius, L. & Migliano, A. B. (2016). Knowledge-sharing networks in hunter-gatherers and the evolution of cumulative culture. *Current Biology*, 26, 2516–2521.

Santoro, F. R., Souto, T., Soldati, G. T. & Roberts, D. R. (2013). Conceptual synthesis in community ecology. *Heliyon*, 6, e04109. https://doi.org/10.1016/j.heliyon.2020.e04109.

Santos, F. R., Souza, E. D. N. F., Sih, A. & Robles Arias, D. M. (2014). The evolution of traditional knowledge: environment shapes medicinal use and why does it fail? *Biological Reviews*, 90, 206–241.

Saslis-Lagoudakis, C. H. & Reyes-García, V. (2016). Knowledge-sharing networks in hunter-gatherers and the evolution of cumulative culture. *Heliyon*, 2, 754–761. https://doi.org/10.1016/j.heliyon.2016.06.003.

Thomas, M. & Callon, S. (2016). Effect of farmers social status and plant biocultural value on seed circulation networks in Vanuatu. *Ethnology and Society*, 21, 133–156.

Ticktin, T., Bevilacqua, S., Francisco, S. & Boero, F. (2009). Taxonomic sufficiency and the increasing insufficiency of taxonomic expertise. *Marine Pollution Bulletin*, 65(56), 556–561. https://doi.org/10.1016/j.marpolbul.2009.02.011.

Turchin, P. (2001). Does population ecology have general laws? *Oikos*, 94, 17–26.

Vanderbeek, I. (2017). Ethical aspects of working with local communities and their biological resources. In *Pharmacognosy*, pp. 645–651 Elisver, Amsterdam.

Vanderbeek, I. & Baleck, M. J. (2012). Globalization and loss of plant knowledge; challenging the paradigm. *PLoS ONE*, 7(3), e37643. https://doi.org/10.1371/journal.pone.0037643.

Vanderbeek, I., Pieren, A., Steep, J. R., Hanazaki, N., Lado, A., Alves, R. R. N., Picking, D., Delgoda, R., Maroij, A. V., van Angel, D., Quaye, C. L., Paniagua-Zambrana, N. Y., Busmans, R. W., Ophondne, G., Arbas, A. M., et al. (2019). Reshaping the future of ethnomedicine research after the COVID-19 pandemic. *Nature Plants*, 6, 723–730.

Vanderbeek, I., van Damme, P., van Puyvelde, L., Arrazola, S. & de Kimpe, N. (2014). A comparison of traditional healers’ medicinal plant knowledge in the Bolivian Andes and Amazon. *South Science and Medicine*, 59, 837–849.

Varela, O. G. (2010). Conceptual synthesis in community ecology. *The Quarterly Review of Biology*, 85, 183–206.

Voeks, R. A. & Garine, E. (2016). Good year, bad year: changing strategies, changing networks? A two-year study on seed acquisition in northern Cameroon. *Ecology and Society*, 21(2), 34. http://dx.doi.org/10.5751/ES-08376-210234.

Voeks, R. A. (2004). Disturbance pharmacopoeias: medicine and myth from the humid tropics. *Annals of the Association of American Geographers*, 94, 686–698.

Voeks, R. A. & León, A. (2004). Forgetting the forest: assessing medicinal plant erosion in eastern Brazil. *Economic Botany*, 58, 94–106.

Wahren, J. (1996). Quantitative aspects of blood flow and oxygen uptake in the human forearm during rhythmic exercise. *Acta Physiologica Scandinavica, Supplement*, 547, 1–113.

Wald, J., Köy, K., Kose, N., Okan, T., Aksoy, E. B., Darb, A. & Allerd, S. (2019). The role of traditional livelihood practices and local ethnobotanical knowledge in mitigating chestnut disease and pest severity in Turkey. *Forests*, 10, 571.

Weber, C. O., Ackery, D. D., McPeek, M. A. & Donoghue, M. J. F. (2002). Phylogenies and community ecology. *Annual Review of Ecology and Systematics*, 33, 475–503.

Weinstein, A. L., El-Derwich, H., Chalbert, S. & Fuentes, M. (2016). Fitting Human Decision Making Models using Python. *Proceedings of the 15th Python in Science Conference*, 1–6.

Weinstein, B. G., Marconi, S., Aubry-Kentz, M., Vincent, G., Sentovino, H. & White, E. P. (2020). DeepForest: A python package for RGB deep learning tree crown delineation. *Methods in Ecology and Evolution*, 11, 1743–1751.

Winter, K. B., Ticktin, T. & Quazi, S. A. (2020). Biocultural restoration in Hawai'i also achieves core conservation goals. *Ecology and Society*, 25, art26.

Yessoufou, K., Daré, B. H. & Muanya, A. M. (2015). Phylogenetic reconstruction of commonly used medicinal plants in South Africa. *Molecular Ecology Resources*, 15, 405–413.

Zanne, A. E., Tank, D. C., Cornwell, W. K., Eastman, J. M., Smith, S. A., FitzJohn, R. G., McGlinn, D. J., O’Meara, B. C., Moles, A. T., Reich, P. B., Royer, D. L., Solis, D. E., Stevens, P. F., Westoby, M., Wright, I. J., et al. (2014). Three keys to the radiation of angiosperms into freezing environments. *Nature*, 506, 89–92.

Zuur, A. F., Ieno, E. N., Walker, N., Saveliev, A. A. & Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer, New York.

Received 19 November 2020; revised 16 May 2021; accepted 18 May 2021; published online 30 May 2021.