Paleobiology, 49(2), 2023, pp. 232–258
DOI: 10.1017/pab.2022.35

PALEOBIOLOGY
A PUBLICATION OF THE Paleontological SOCIETY

Article

Linking host plants to damage types in the fossil record of insect herbivory

Sandra R. Schachat, Jonathan L. Payne, and C. Kevin Boyce*

Abstract.—Studies of insect herbivory on fossilized leaves tend to focus on a few, relatively simple metrics that are agnostic to the distribution of insect damage types among host plants. More complex metrics that link particular damage types to particular host plants have the potential to address additional ecological questions, but such metrics can be biased by sampling incompleteness due to the difficulty of distinguishing the true absence of a particular interaction from the failure to detect it—a challenge that has been raised in the ecological literature. We evaluate a range of methods for characterizing the relationships between damage types and host plants by performing resampling and subsampling exercises on a variety of datasets. We found that the components of beta diversity provide a more valid, reliable, and interpretable method for comparing component communities than do bipartite network metrics and that the rarefaction of interactions represent a valid, reliable, and interpretable method for comparing compound communities. Both beta diversity and rarefaction of interactions avoid the potential pitfalls of multiple comparisons. Finally, we found that the host specificity of individual damage types is challenging to assess. Whereas bipartite network metrics are sufficiently biased by sampling incompleteness to be inappropriate for fossil herbivory data, alternatives exist that are perfectly suitable for fossil datasets with sufficient sample coverage.

Sandra R. Schachat, Jonathan L. Payne, and C. Kevin Boyce. Department of Geological Sciences, Stanford University, Stanford, California 94305, U.S.A. E-mail: schachat@stanford.edu, sschachat@schmidtsciencefellows.org, jlpayne@stanford.edu, chkenboy@stanford.edu

Accepted: 20 October 2022
*Corresponding author.

Introduction

Insect herbivory on fossilized leaves (hereafter "fossil herbivory") has been noted incidentally for more than 100 years (Potonié 1893). However, the systematic collection of herbivory data only came with the advent of the Damage Type system (Wilf and Labandeira 1999), in which each type of insect damage—for example, circular holes below 1 mm in diameter, circular holes between 1 and 5 mm in diameter—is assigned a unique number and is classified into a broader functional feeding group (Labandeira et al. 2007).

Traditionally, quantitative analyses of fossil herbivory have focused on two topics: the richness of damage types in a fossil assemblage or for a particular host plant (Wilf and Labandeira 1999) and the intensity of insect damage as measured by the percentage of leaf area removed by herbivores (Beck and Labandeira 1998). Another layer of biological and analytical complexity can be added by linking particular host plants to particular damage types. On the one hand, quantitative methods in paleontology and ecology have progressed tremendously during the past two decades, making it possible to conduct complex analyses of fossil herbivory data with a single line of code after loading a software package. On the other hand, such analyses require more complete datasets than are typically available in studies of fossil herbivory. Many newly possible analyses also rely upon more assumptions about biological processes and data structure or

estimate more parameters than do traditional analyses. In such cases, the underlying assumptions and their effects can become more difficult to identify and address.

Research Topics That Link Host Plants to Damage Types

Three interrelated research topics link host plants to damage types: host specificity, component communities, and compound communities. Host specificity differentiates among generalist and specialist feeding strategies. A component community is the entire suite of heterotrophs that rely, directly or indirectly, on a plant taxon: its herbivores and their predators, parasitoids, and parasites (Root 1973). A suite of coexisting component communities, that is, those of the different plant species within the same forest, is called a "compound community" (Reice 1974; Whittaker and Levin 1977; Basset 1992; Novotny et al. 2002). All of these topics present challenges when translated to the fossil record.

The host specificity of each fossil insect damage type is typically measured on a scale of 1 to 3 (Labandeira et al. 2007). Generalized damage types, occurring on a range of distantly related plant hosts, have a score of 1. Damage types of intermediate specificity have a score of 2. Specialized damage types, restricted to very closely related plant hosts, have a score of 3. These scores are assigned to damage types that occur on three or more specimens in a fossil assemblage. The assignment of these scores at various fossil assemblages is difficult to replicate, because the boundaries between the scores are not defined quantitatively—these symbols are merely qualitative and contain no quantitative information—but the many datasets that have become available since 1999 can be used for sensitivity analyses to evaluate the validity and reliability of this system.

For component communities, identification of the secondary consumers associated with the herbivores on a host plant is challenging with fossils (Greenwood 1991; Martínez-Delclòs and Martinell 1993; Smith and Moe-Hoffman 2007). Many of the iconic insect Lagerstätten also contain abundant plant fossils (from oldest to youngest: Carpenter 1997; Wittry 2006; Novokshonov 1997; Ponomareva et al. 1998; Cairncross and Anderson 1995; Anderson 1999; Dobruskina 1995; Shcherbakov 2008; Huang et al. 2016; Ren et al. 2019; Huang 2016; Xiao et al. 2022; Ribeiro et al. 2021; Wappler et al. 2009; MacGinitie 1969; Wilson 1978; Grande 1984; Dayvault et al. 1995; Wappler et al. 2012; Dunne et al. 2014; Douglas and Stockey 1996; Labandeira 2002; Constenius et al. 1989; Greenwalt and Labandeira 2013; Wilde and Frankenhäuser 1998; Lutz et al. 2010; Wappler et al. 2012; Wilson 1978; Meyer 2003; Allen et al. 2020). When the relevant plants and insects do co-occur, it is nearly impossible to link a particular insect taxon (within a given feeding guild) to a particular damage type (within a given functional feeding group). Nonetheless, component communities in the fossil record have been widely discussed using damage types as proxies for herbivore taxa (Labandeira 1998, 2002; D'Rozario et al. 2011; Slater et al. 2012, 2015; Labandeira and Currano 2013; Labandeira et al. 2013, 2016, 2018; Ding et al. 2014, 2015; Schachat et al. 2014, 2015; Feng et al. 2017; Kustatscher et al. 2018; Xu et al. 2018; Correia et al. 2020; Liu et al. 2020). However, here too, there is reason for caution: even the fossil floras that have been most thoroughly sampled for insect herbivory contain various damage types that occur on only one specimen (Wilf et al. 2005, 2006; Prevec et al. 2009; Wappler 2010; Knor et al. 2012; Wappler et al. 2012; Donovan et al. 2014; Adroit et al. 2018; Labandeira et al. 2018; Xu et al. 2018; Deng et al. 2020), indicating that many damage types remain unobserved due to incomplete preservation and sampling. Because we cannot find every damage type from a fossil assemblage, and because we cannot link damage types to the insect taxa in a one-to-one manner, the term "component community" as developed in the context of modern ecology may be somewhat inapplicable. These issues then scale up to consideration of compound communities.

Whereas convincingly complete sampling is hardly inevitable in the modern, it is at least possible. This can be seen in a recent comparison of fossil and modern herbivory (Swain et al. 2022). For one of the modern datasets considered, 1500 m^2 of leaf area was sampled for each plant species (Novotny et al. 2012). For

another modern dataset, between 1500 and 10,500 m^2 of leaf area was sampled for each plant species (Novotny et al. 2005). In contrast, between 0.000497 and 1.62 m^2 of leaf surface area was sampled for each fossil plant taxon included in the study. Relative to the four paleontological studies, the amount of leaf area examined in the two highlighted modern studies is at least 925 times greater, and can be more than 20 million times greater.

Despite these issues, the general concepts drawn from modern ecology that underlie discussions of component communities in the fossil record are nevertheless valid. Ancient plants surely had specialist and generalist herbivores that formed component communities along with their secondary consumers on each plant host species. Thus, these concepts are worthy of consideration, although we must be wary of the fidelity with which those communities might be documented in the fossil record. In particular, bipartite network analysis has recently been applied to fossil herbivory datasets to address questions about host specificity and component communities (Currano et al. 2021; Swain et al. 2022). Bipartite networks may be used to connect taxa at two trophic levels, such as plants and their herbivores or herbivores and their parasitoids. Alternatively, beta diversity (Baselga 2010, 2017; Baselga and Orme 2012) and rarefaction of interactions (Dyer et al. 2010) can be used to examine herbivore specialization and component communities based on the leaf damage record. Calculating the beta diversity of damage types on different host plants is a straightforward way to compare component communities. Rarefying interactions is a straightforward way to quantify the diversity of associations within a compound community. Here, these alternatives are evaluated through sensitivity analyses to determine how much sampling is required for stable results, with the aim of ascertaining whether and how quantitative methods can be used to evaluate host specificity, component communities, and compound communities in studies of fossil herbivory. Bipartite network analysis requires special consideration because of the assumptions it requires of the fossil record and because of the risks associated with the large number of metrics that are generated.

Theoretical Issues with Bipartite Network Analysis

Treating Damage Types as Analogues of Herbivore Taxa.—Methods that link particular host plants to particular damage types often treat damage types as analogues for herbivorous insect taxa. For example, the two recent studies that performed bipartite network analysis on fossil herbivory data (Currano et al. 2021; Swain et al. 2022) used a software package (bipartite; Dormann et al. 2008) intended for modern ecological networks that requires direct substitution of damage types for herbivore taxa—constituting an explicit, specific assumption that has not been substantiated and likely never can be. Only one study has used neontological data to evaluate the correlation between damage types and herbivores (Carvalho et al. 2014). In two tropical forests, the diversities of damage types and insect herbivores were found to be correlated, reaffirming the value of the traditional paleontological metric of damage type diversity. However, no claim was made as to whether the apparent specialization of a damage type reliably indicates whether the damage type was produced by a specialist herbivore.

Simple arithmetic supports the idea that specialized herbivores are responsible for many occurrences of "generalized" damage types: with hundreds of thousands of herbivorous insect species and only a few hundred damage types, no one-to-one correspondence between insect species and damage types is possible. For example, DT012, the most common type at both forests studied by Carvalho et al. (2014), was found on all 12 host plant species examined and was caused by 50 insect species (46 of them specialists) in one locality and 37 insect species (23 of them specialists) in the other. All that complexity is collapsed into a single generalist when fossil damage types are treated as substitutes for actual herbivores.

"Trophic species" are occasionally used as substitutes for consumer taxa in studies of ecological networks. Dunne and colleagues explain that trophic species are generated "by aggregating taxa with the exact same set of predators and prey" (Dunne et al. 2014). In other words, trophic species are "functional groups of all organisms in a web that appear to share

the same set of consumer and resource species" (Memmott et al. 2000). In the context of insect herbivory, a trophic species would be a group of herbivore taxa, however distantly related, that feed on the same host plant taxon in a similar manner—and share the same predators and parasitoids. As can be seen in the preceding discussion, the aggregation into a single unit (a damage type) of specialist herbivore taxa that cause morphologically similar damage on disparate host plant taxa violates the trophic species concept. Moreover, various workers disagree with the use of trophic species (Pringle and Hutchinson 2020).

Sampling Incompleteness.—All sampling of the fossil record is incomplete, but methods that link particular host plants to particular damage types are far more biased by incomplete sampling than are the methods that address the diversity and intensity of insect herbivory. For a tally of the number of insect damage types on two host plant taxa, as an example, the more completely sampled host could be iteratively subsampled down to the amount of surface area or sample coverage available for the less completely sampled host plant (Fig. 1A). Although the subsampling procedure might cause a failure to detect a significant difference that would become apparent with additional sampling, any significant differences observed among the subsampled damage type diversities are likely, although not guaranteed, to reflect true differences. Thus, estimating damage type diversity by subsampling two incompletely sampled host plants is a common and uncontroversial endeavor. We do not know which specific damage types evaded detection, but we do not need to know this in order to estimate the damage type diversities of these two host plants when subsampled to the same surface area or sample coverage.

When it comes to estimating host specificity or comparing component communities, however, the unknowable identities of unobserved damage types are of paramount importance. According to the criteria that have traditionally been used to assign host-specificity scores (Wilf and Labandeira 1999), a damage type must occur on only three specimens in order to receive such a score. The data are taken at face value, and the appearance of a damage type on three leaves is deemed adequate to designate a damage type as specialized, regardless of the possibility that a fourth or fifth observation might occur on a different

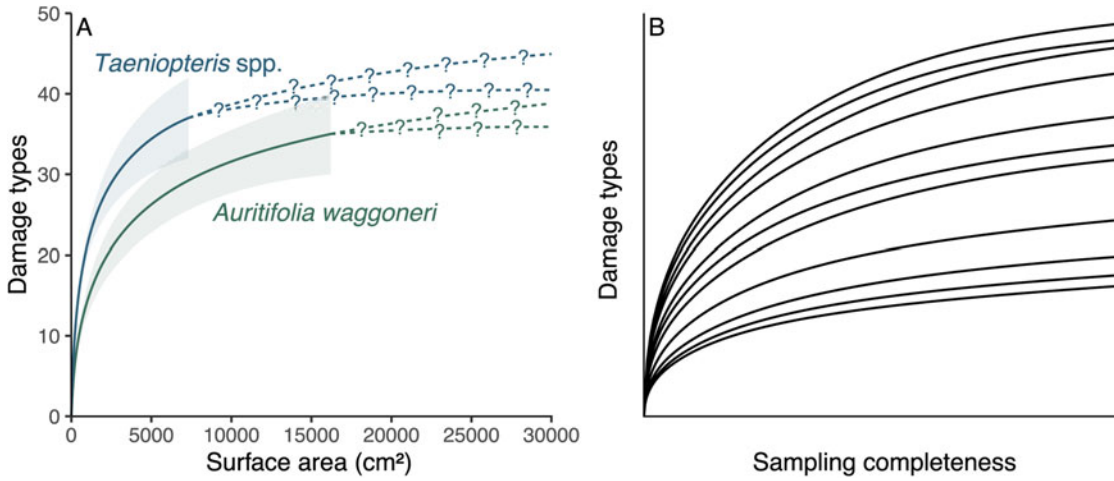

FIGURE 1. A comparison of the sampling completeness that can be expected for studies of fossil herbivory (A) with the sampling completeness needed for methods that link host plants to damage types to be unbiased by sampling completeness (B). A, Rarefaction of damage types on the two dominant host plants at the Colwell Creek Pond assemblage. The solid lines and corresponding 84% confidence intervals represent interpolated damage type diversity, and the dashed lines with question marks represent extrapolated diversity. B, An illustration of the sampling completeness that is needed for bipartite network analysis not to be biased by sampling: the rarefaction curve for each host plant should have sample coverage sensu Chao and Jost (2012) above 0.99. All rarefaction curves shown in this panel have coverage between 0.995 and 0.997.

host and thus change the host-specificity score. The procedures used to compare component communities are incapable of distinguishing a true absence of a damage type on a host plant from the failure to detect a damage type that was present on the host. Differentiating true absences from failures to detect is known to pose tremendous difficulties in both neontological (Blasco-Moreno et al. 2019) and paleontological (Smith et al. 2022) studies.

Attempts to compare host specificity and component communities across different assemblages complicate matters even further. As an example drawn from Permian assemblages of Texas for which damage type data are available for each specimen, the amount of broadleaf area examined from Colwell Creek Pond (Schachat et al. 2014) is approximately 4 times that of Williamson Drive (Xu et al. 2018) and more than 15 times that of Mitchell Creek Flats (Schachat et al. 2015) or South Ash Pasture (Maccracken and Labandeira 2020). There is just no good way to compare host specificity and component communities across these assemblages, because subsampling Williamson Drive and Colwell Creek Pond down to the amount of surface area examined at Mitchell Creek Flats and South Ash Pasture will fundamentally change the relationships among host plants and their damage types. At Colwell Creek Pond, DT014 has been observed on 2 *Auritifolia waggoneri* Chaney, Mamay, DiMichele & Kerp, 2009 specimens and on 20 *Taeniopteris* spp. Brongniart, 1828 specimens. DT247 has been observed on 15 *A. waggoneri* specimens and 2 *Taeniopteris* spp. specimens. If the data from Colwell Creek Pond are subsampled to one-fifteenth of the original amount of surface area, the specificity coding of the damage types that are still observed at this lower level of sampling will fundamentally change: various damage types will appear more specialized than they are, and in many dimensions, the component communities of the two dominant host plants will appear more distinct than they are.

In the words of Blüthgen et al. (2008: p. 3387), "rarely observed species are inevitably regarded as 'specialists,' irrespective of their actual associations, leading to biased estimates of specialization." For rarefied damage type diversity and for the intensity of herbivory, the results generated at lower levels of sampling completeness are simply a less precise, underpowered version of the results generated at higher levels of sampling completeness (Schachat et al. 2018). For component communities, however, the results generated with less sampling are fundamentally changed. Indeed, misleading results when sampling is not exhaustive are exactly what biologists found when they subsampled some of the canonical datasets that have been used to construct bipartite networks (Morris et al. 2014: fig. 3) as part of the large body of work that has emerged to evaluate how incomplete sampling biases bipartite network metrics (Goldwasser and Roughgarden 1997; Vázquez and Aizen 2003; Blüthgen et al. 2006, 2008; Dormann et al. 2009; Dorado et al. 2011; Gibson et al. 2011; Costa et al. 2016; Fründ et al. 2016; Jordano 2016; Kuppler et al. 2017; Maia et al. 2018; Henriksen et al. 2019).

A related pitfall of bipartite network analysis that looms large in the neontological literature may well be insurmountable for studies of fossil herbivory: sampling evenness. Before the construction of bipartite networks, the sampling of fossil leaves for insect damage types should be not only complete at the level of the assemblage but should be similarly complete across all host plants within the assemblage—that is, sampling of all host plants under consideration should be even (Gibson et al. 2011; Doré et al. 2021). In studies of modern communities, sampling evenness can be achieved in various ways, for example, equal amounts of time being dedicated to hand-collecting of insects and equal numbers of beating samples collected for each of 10 tree species (Basset et al. 1996) and equal amounts of surface area sampled for each plant species (Novotny et al. 2012). However, exhaustive sampling of all host plant taxa under consideration is a near impossibility for studies of fossil herbivory (Fig. 2). Most species in a given community are rare (Diserud and Engen 2000), and many if not most studies of fossil herbivory have examined fewer than 1000 leaves due to a combination of small numbers of specimens preserved in the fossil record and limited time that investigators can invest in each study. Therefore, in studies of fossil herbivory, most

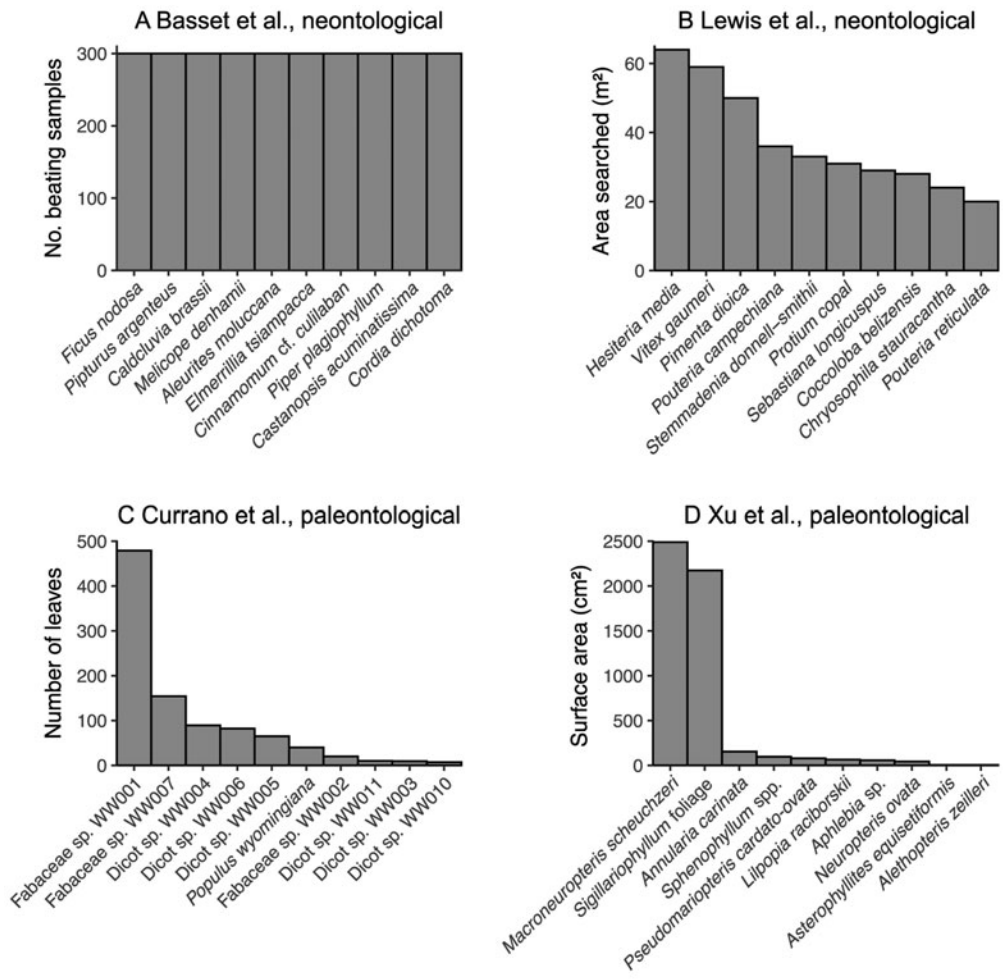

FIGURE 2. The sampling evenness for host plants in neontological (A, B) and paleontological (C, D) datasets that can be used to link host plants to herbivores or damage types. A, Basset et al. (1996); this maximally even sampling is representative of various other neontological studies of plant–insect networks (Novotny et al. 2002, 2004, 2012; Lundgren and Olesen 2005; Olesen et al. 2008; Pinheiro et al. 2008; Gibson et al. 2011; Grass et al. 2013; Trøjelsgaard et al. 2015; Oleques et al. 2019; Zemenick et al. 2021). B, Lewis et al. (2002). C, Currano et al. (2008). D, Xu et al. (2018).

plant hosts are represented by a maximum of a few hundred leaves.

Combining the concepts of sampling completeness and evenness, Morris et al. (2014) recommended constructing bipartite networks for datasets in which all rarefaction curves—in this case, damage type diversity curves for all host plants—asymptote (Fig. 1B). Various neontological food web studies have affirmed and followed this recommendation (e.g., Smith-Ramírez et al. 2005; Burkle and Irwin 2009; Mokam et al. 2014; Kemp and Ellis 2017; Peguero et al. 2017; Arceo-Gómez et al. 2018; Bennett et al. 2018; Maia et al. 2018). However, this is not nearly as easily achieved with paleontological data as with neontological data. Whereas one might question whether it is possible for a rarefaction curve to truly asymptote, the concept of "sample coverage" sensu Chao and Jost (2012) provides a measure of the slope of a rarefaction curve: when the curve has reached an asymptote, its slope equals 0 and coverage equals 1. For our purposes, sample coverage above 0.99 will be considered complete. If a paleontological dataset with 10 or more host plants that have coverage

above 0.99 eventually becomes available, it can be used to evaluate whether slightly lower amounts of coverage continue to yield reliable results. The Appendix lists examples of host plants that have been censused for fossil herbivory for which sample coverage of damage types is above 0.99.

The requirement that all rarefaction curves reach an asymptote is unrealistic for essentially the entirety of the fossil record of insect herbivory as it is currently sampled. At the Willershausen assemblage (Adroit et al. 2018), for which more than 7000 angiosperm leaves were examined, coverage for the 10 most abundant host plants ranges from 0.90 to 0.99. However, at Castle Rock (Wilf et al. 2006), another of the few assemblages with more than 2000 angiosperm leaves examined for which damage type data are available for each specimen, coverage of the top 10 host plants is much lower, with some taxa preserving no damage at all and the highest coverage only reaching 0.72. At the Bílina–DSH assemblage (Knor et al. 2012), also with more than 2000 angiosperm leaves examined, coverage of the top 10 host plants ranges from 0.59 to 0.90. Therefore, low sample coverage of damage types for individual host plants is clearly not due to lack of investigator effort; this is a characteristic of some of the best-sampled assemblages. Rather, low sample coverage of damage types for individual host plants is a near-inevitability given the vastly uneven frequencies of both host plants and damage types in fossil assemblages. Even the less common host plants must be represented by enough specimens for their individual damage diversity rarefaction curves to asymptote.

HARKing.—A "reproducibility crisis" in science (O'Boyle et al. 2017; Fraser et al. 2018; Hutson 2018; Parker et al. 2019; Bissonette 2021; Nelson et al. 2021; O'Dea et al. 2021) has reinforced the need for caution surrounding practices such as multiple comparisons and hypothesizing after the results are known (HARKing). When the popular R package bipartite is used with its default settings to study plant–herbivore interactions, the *networklevel* function calculates 47 bipartite network metrics and the *grouplevel* function calculates 30 metrics: 15 for each host plant taxon and 15 for each herbivore taxon (Dormann et al. 2008)—77 metrics despite few studies addressing 77 distinct questions. Such a multitude of metrics raises the risk of spurious correlations whereby a small minority of metrics support preconceived notions by chance.

For bipartite network studies, calculating a single bipartite network metric per study has been recommended to avoid "metric hacking," that is, the "nonmutually exclusive use of multiple network metrics that are correlated by variables held in common (e.g., number of host plant taxa or sampling completeness) and the inflation of type I error rates as a result of indiscriminate selection of network metrics, comparisons or hypotheses after analyses have been conducted" (Webber et al. 2020: p. 110). This warning echoes concerns raised more than a decade earlier: "Network analyses of mutualistic or antagonistic interactions between species are very popular, but their biological interpretations are often unclear and incautious" (Blüthgen 2010: p. 187). The unclear meanings of bipartite network metrics raise the specter of the "file drawer" problem, in which results that are inconclusive, negative, or do not fit with the authors' agenda are not reported (Fraser et al. 2018). The complexity of bipartite networks makes their analysis subject to these risks in a way that traditional metrics of herbivore damage diversity and intensity are not. A priori decisions about which metrics are most relevant to a given ecological question may address this concern, but Webber et al. (2020) note that the appropriate metric is often unclear for any particular scenario.

Methods

Bipartite networks and several alternative methods were evaluated using existing data, with a focus on the Willershausen assemblage (Adroit et al. 2018) as the angiosperm-dominated assemblage with a complete, publicly available dataset that has the highest number of leaves examined. Of the assemblages previously examined in the context of bipartite networks (Currano et al. 2021), Willershausen is emphasized as a conservative test, because it is among the few assemblages most likely to have sufficient sampling completeness to

quantify host specificity, component communities, and compound communities.

All analyses were performed with R v. 4.1.1 (R Development Core Team 2021). Color schemes were generated with the packages colorbrewer (Neuwirth and Brewer 2014) and scico (Pedersen and Crameri 2020). For all subsampling routines, we first subsampled each dataset to approximately half of its original size and to progressively smaller sizes, using round numbers when possible for the sake of readability.

Evaluating Bipartite Network Analysis

Sensitivity of Bipartite Network Metrics to Sampling Completeness.—The 28 network-level metrics previously used in fossil herbivory studies (Currano et al. 2021: table S1; Swain et al. 2021, 2022: table 1, appendix S4) that are calculated with the *networklevel* function in the bipartite package (Dormann et al. 2008) were calculated for the Willershausen assemblage, using subsampling and resampling procedures as sensitivity analyses. Leaves that were not identified to the genus level were removed from the dataset. Each subsampling and resampling routine was iterated 1000 times.

In the first set of routines ("complete"), the entire cleaned Willershausen dataset was analyzed, resampled to the original number of leaves in the cleaned dataset (7333), and subsampled to 3500, 1000, 500, and 300 leaves. Following previous methods (Swain et al. 2022), all host plant taxa represented by fewer than five specimens were removed after the data were resampled or subsampled but before any analyses were performed.

To mirror neontological datasets (Basset et al. 1996; Lewis et al. 2002) that were recently compared with fossil herbivory data (Swain et al. 2022), we employed a second set of routines ("top 10") that involved only the 10 host plant taxa at Willershausen with the highest numbers of leaves, ranging from the 948 leaves of *Zelkova ungeri* Unger, 1843 (Kotlaba 1963) down to the 164 leaves of *Betula maximowicziana* Regel, 1868. This top 10 dataset of 3602 leaves was resampled to the original number of leaves and subsampled to 1800, 1000, 500, and 300 leaves.

For the sake of comparison, we calculated damage type diversity with coverage-based rarefaction (Chao and Jost 2012) for each resampled and subsampled dataset, using the *iNEXT* function in the R package iNEXT (Hsieh et al. 2016). We rarefied damage type diversity to the three sample coverage thresholds discussed by Schachat et al. (2022): 0.7, 0.8, and 0.9.

Bipartite Network Metrics and the Potential for HARKing.—To evaluate the possibility of "multiple network metrics that are correlated by variables held in common"—the collinearity among metrics noted as a major pitfall of bipartite network analysis (Webber et al. 2020: p. 110)—the same 28 network-level metrics discussed earlier were calculated for a series of fossil assemblages deposited shortly before, during, and after the Paleocene/Eocene thermal maximum (PETM) and the early Eocene climatic optimum (EECO) in the Bighorn Basin and Wind River Basin. Network metrics were calculated after subsampling the data from each assemblage to 300 leaves, following the procedure of Currano et al. (2021). If a subsample is larger than 50% of the original dataset, the number of possible unique samples decreases. Therefore, subsampling to 300 leaves and generating accurate confidence intervals requires a sample size of at least 600 leaves. The 10 relevant assemblages with 600 or more leaves are Skeleton Coast and Lur'd Leaves from the Bighorn Basin (Wilf et al. 2006); Dead Platypus, Daiye Spa, Hubble Bubble, the South Fork of Elk Creek, PN, and Fifteenmile Creek from the Bighorn Basin (Currano et al. 2008, 2010); and the Wind River Interior and Wind River Edge assemblages from the Wind River Basin (Currano et al. 2019).

Evaluating Alternatives to Bipartite Network Analysis

Beta Diversity.—We evaluated the validity and reliability of measures of abundance gradients (analogous to nestedness: when the damage types observed on one host plant are a subset of the damage types observed on another host plant) and balanced variation in abundance (hereafter "balanced variation"; analogous to turnover: when non-overlapping suites of damage types are observed on different host plants). These are the two components of beta diversity that explicitly account for

differences in abundance (Baselga 2017). Our first analysis of beta diversity focuses on the two host plants represented by the highest numbers of leaves at Willershausen: *Z. ungeri* and *Fagus sylvatica* Linnaeus, 1753. We used each subsampled and resampled dataset generated from the complete Willershausen dataset. Our second analysis of beta diversity focuses on *Auritifolia waggoneri* and *Taeniopteris* spp., the two most abundant host plants at Colwell Creek Pond (Schachat et al. 2014). These two host plants were analyzed at five levels of sampling. They were jointly resampled to the original amount of surface area they comprise in the Colwell Creek Pond dataset (23,527.89 cm^2) and were subsampled to a total of 11,750, 8000, 4000, and 2000 cm^2. Our third analysis of beta diversity focuses on *Macroneuropteris scheuchzeri* (Hoffmann, 1827) Cleal, Shute & Zodrow, 1990 and foliage assigned to *Sigillariophyllum* Grand'Eury, 1877, the two most abundant host plants at Williamson Drive (Xu et al. 2018). These were jointly resampled to the original number of leaves they comprise in the Williamson Drive dataset (1524) and were subsampled to a total of 750, 600, 450, and 300 leaves. Although surface area measurements were taken for Williamson Drive (Xu et al. 2018), we subsampled these data by number of leaves, because the surface area measurements for individual specimens are not available. Each subsampling routine was iterated 1000 times.

Abundance gradients and balanced variation were calculated for each subsampled and resampled dataset using the *beta.pair.abund* function in the R package betapart (Baselga and Orme 2012). We used the *coverage* function in the R package entropart with the "Chao" estimator (Marcon and Hérault 2015) to calculate sample coverage for each of the two plant hosts in each subsampling and resampling routine.

Host Specificity.—The sensitivity of host-specificity scores to sampling completeness was evaluated with the complete and top 10 resampling and subsampling routines for the Willershausen dataset. For each set of sampling routines, we recorded the number of host plant taxa on which we observed a randomly selected damage type within the 99^{th}, 74^{th}, and 49^{th} percentiles of prevalence (Table 1).

We performed a separate sampling procedure to address the impact of absolute and relative surface area on estimates of host specificity. For this procedure we used the data from Colwell Creek Pond (Schachat et al. 2014), because this assemblage contains a large amount of surface area examined, and because surface area measurements are available for each individual specimen along with damage type data. We sampled specimens belonging to *A. waggoneri*, *Taeniopteris* spp., *Evolsonia texana* Mamay, 1989, and *Supaia thinnfeldioides* White, 1929, with replacement, to a series of 51 equally spaced surface area thresholds from 500 cm^2 to 25,500 cm^2. The smallest of these is approximately 2% of the total surface area, and the largest of these is approximately 100% of the total surface area. We resampled the data to each threshold 10,000 times, for a total of 510,000 iterations. For each iteration, we noted whether DT032 and DT120—which are distributed across all four of these host plant taxa—were restricted to only one host plant, thus falsely appearing to be specialized. If so, we noted the number of specimens on which the damage type had been observed.

TABLE 1. The percentiles of leaves on which damage types were observed at the Willershausen assemblage.

	Complete			Top 10		
	99^{th} percentile	74^{th} percentile	49^{th} percentile	99^{th} percentile	74^{th} percentile	49^{th} percentile
Number of leaves	721	16	6	381	22	4
Damage types	DT003	DT033, DT145	DT010, DT021, DT052, DT081, DT142, DT190, DT198	DT003	DT004, DT020	DT008, DT052, DT061, DT168
Randomly selected damage type	DT003	DT033	DT081	DT003	DT004	DT168

Rarefaction of Interactions.—The method of Dyer et al. (2010), which measures the diversity of interactions at an assemblage, can be implemented with any algorithm that performs rarefaction. We discuss considerations for coverage-based rarefaction of interactions in the Appendix.

We performed coverage-based rarefaction of interactions on data from Williamson Drive (Xu et al. 2018) and Colwell Creek Pond (Schachat et al. 2014). We conducted coverage-based rarefaction on the original dataset, and upon iteratively resampling each dataset to the original amount of surface area, and upon subsampling each dataset to 50% and 25% of the original surface area. (Surface area data were collected for each specimen at Williamson Drive but were not published with the damage type data. Therefore, the surface area assigned to each specimen was the mean value for the taxon to which it belongs.) We rarefied each vector of interaction counts to a sample coverage of 0.771, which is the maximum amount of coverage reached by all subsampled datasets.

Because the importance of sampling completeness is a major theme of this contribution, we wished to test the extent to which rarefaction of interactions is robust to incomplete sampling. To understand how rarefaction of interactions might perform on an angiosperm-dominated dataset with complete sampling, we simulated a vector of counts of interactions using the base-R function *rlnorm* with the settings meanlog=0 and sdlog=1.5. We chose this method because we found that it yielded a distribution of interaction frequencies that closely mirrors that seen at Willershausen. This procedure generated 3000 values, which we had to round to whole integers, because these values represent simulated counts. Upon removing the values that round down to 0, we had 2046 simulated unique interactions that were observed a total of 9597 times. These numbers are approximately double those seen in the Willershausen dataset, so we attributed these simulated interactions to 15,000 leaves, because this is approximately double the number in the Willershausen dataset.

We examined the validity and reliability of rarefaction of interactions in this simulated dataset by subsampling. We subsampled the interactions to one-half of the original count (4798), attributing these to one-half of the original number of leaves (7500). We then subsampled the interactions to one-quarter of the original count (2399), attributing these to one-half of the original number of leaves (3750). We rarefied each vector of subsampled interaction counts to a sample coverage of 0.726, which is the maximum amount of coverage reached by all subsampled datasets.

All rarefaction of interactions was carried out with the *estimateD* function in the R package iNEXT. All resampling and subsampling procedures were iterated 1000 times.

Results and Discussion

Sensitivity of Bipartite Network Metrics to Sampling Completeness

None of the 28 network-level metrics mentioned in previous studies of fossil herbivory (Currano et al. 2021; Swain et al. 2021, 2022) perform as unbiased estimators for the complete Willershausen dataset (Fig. 3). (An unbiased estimator is an estimator whose average value does not change in response to sampling completeness.) Two simple criteria for robustness to sampling completeness are that the 95% confidence intervals for all subsampling routines contain the mean estimate for the resampling routine and that the 95% confidence interval for the resampling routine contains the mean estimates for all subsampling routines. Coverage-based rarefaction of damage type diversity fulfills these two criteria (Fig. 4), but not a single network metric examined here does.

When the Willershausen data are restricted to only the 10 host plants with the highest number of leaves in the dataset (Fig. 5), 4/28 network metrics fulfill these criteria and thus perform comparably well to coverage-based rarefaction: togetherness for damage types, niche overlap for damage types, C score for damage types, and nestedness.

Only one network metric, C score for damage types, is among the best-performing metrics in both the complete and top 10 analyses of the Willershausen dataset. If the C score for damage types were found to be robust for the majority of available fossil herbivory

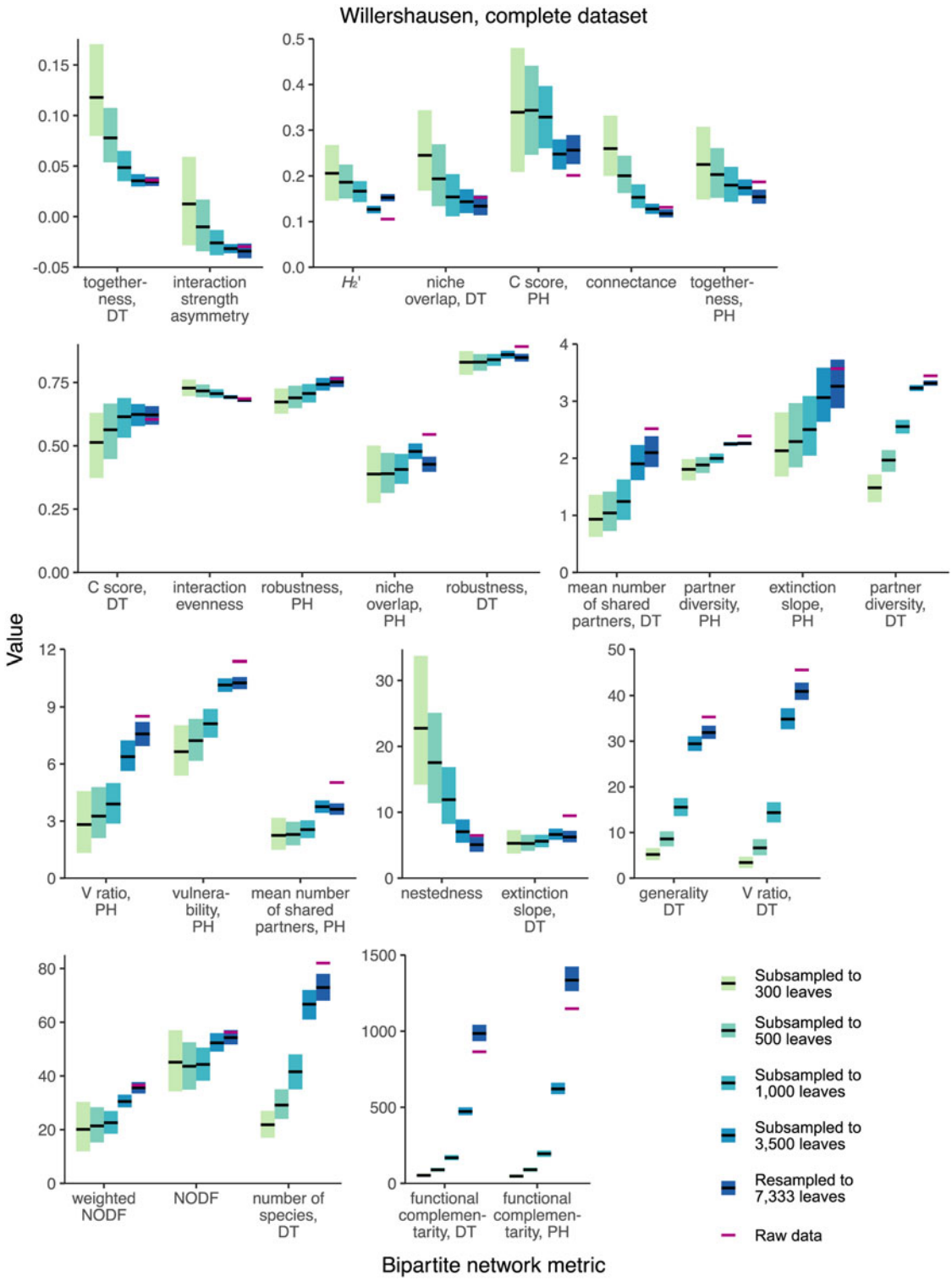

FIGURE 3. Mean values and 95% confidence intervals for bipartite network metrics generated by resampling and subsampling the cleaned Willershausen dataset in its entirety. DT, damage type; PH, plant host.

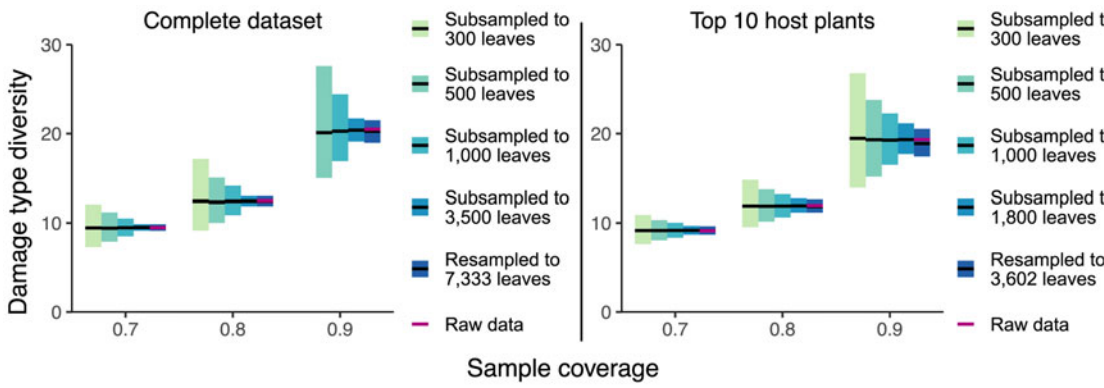

FIGURE 4. An example of a nearly unbiased estimator. Mean values and 95% confidence intervals for coverage-based rarefaction generated by resampling and subsampling the Willershausen dataset. Moreover, coverage-based rarefaction performs as a consistent estimator, in that estimates converge on the true value as sample size increases. No results are presented for 300 subsampled leaves from the complete dataset at sample coverage of 0.9, because some iterations of this sampling routine yielded an observed sample coverage below 0.9.

datasets, which are far less complete than Willershausen, a key question would still need to be answered: What does the C score tell us (Emer et al. 2016)? The C score has been described in the fossil herbivory literature as "the checkerboard (mutual presence/absence) nature of the interactions" (Swain et al. 2022: p. 243) and as "the randomness of species distribution across an ecosystem" (Currano et al. 2021: p. 8), but no outstanding paleontological questions that can be addressed with such a metric have been identified.

Apparent Robustness at Lower Sample Sizes.—For many metrics in both the complete and top 10 datasets, the mean estimate and the limits of the confidence intervals change little for the subsampling routines at 1000, 500, and 300 leaves. However, when the resampling routine and the subsampling routines with more than 1000 leaves are taken into account, it is clear that these metrics are biased by sampling incompleteness. The misleading appearance of a lack of bias in certain network metrics seen at lower levels of sampling makes intuitive sense. When a relatively large proportion of realized interactions are unobserved because only 1000 leaves have been sampled, the additional proportion of realized interactions that go unobserved at 500 or 300 leaves will make little difference for various metrics. These findings and this reasoning highlight the danger of evaluating the bias of network metrics by performing sensitivity analyses on smaller datasets. Therefore, any metrics that appear robust to subsampling routines performed on datasets smaller than that of Willershausen should be treated with extreme caution. For these same reasons, methods that quantify the extent to which bipartite network metrics are biased by sampling incompleteness (Swain et al. 2021) may well be unreliable, especially when applied to incomplete datasets.

Implications for Other Assemblages.—At any amount of sampling that is realistic for studies of fossil herbivory, the results of bipartite network analysis are biased by sampling completeness. The finding that certain metrics are "relatively robust" (Swain et al. 2021) is an inevitability by chance alone given presentation of dozens of metrics (Currano et al. 2021; Swain et al. 2022). Even when we limit our analysis to the 10 most abundant host plants at Willershausen, the mean estimates at 300 and 500 leaves for the best-performing metrics (Swain et al. 2021) either lie beyond (NODF, H_2', connectance, and niche overlap PH) or just barely fall within (niche overlap DT) the 95% confidence interval generated with the resampled dataset. Estimates of these metrics at different sampling intensities are even more discordant for the complete Willershausen dataset.

Paleoecologists have only recently begun to implement Bayesian methods to distinguish true absences of interactions from failures to

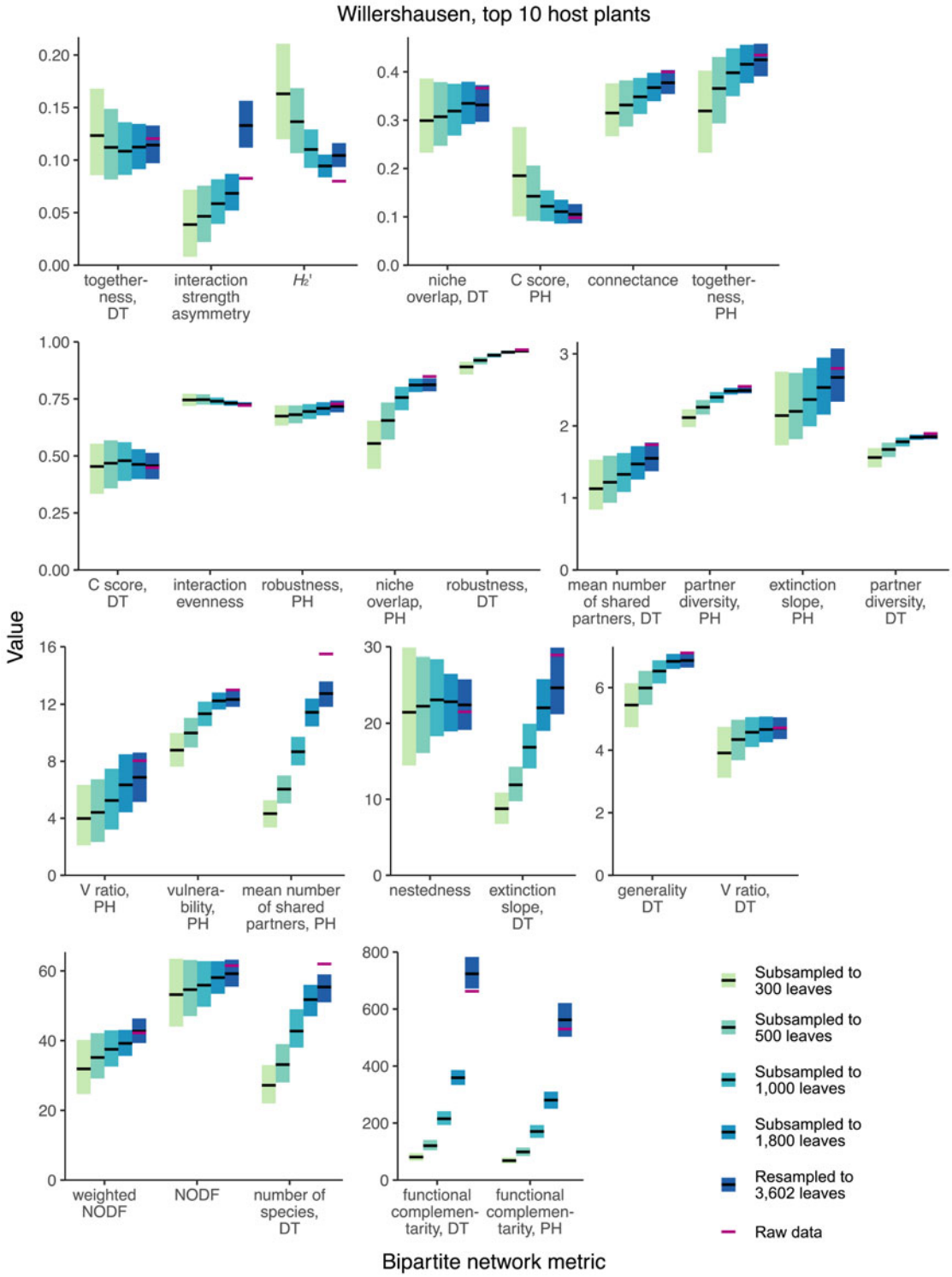

FIGURE 5. Mean values and 95% confidence intervals for bipartite network metrics generated by resampling and subsampling data for the 10 host plants at Willershausen represented by the highest numbers of leaves. DT, damage type; PH, plant host.

detect those interactions. In a recent contribution, the authors used a Bayesian framework to estimate the presence or absence of drilling predation on different molluscan species (Smith et al. 2022). This allowed the authors to estimate the number of zeros—that is, specimens on which predation was not detected—for which predation was truly absent. Whereas Smith and colleagues were interested in the true prevalence of only one type of feeding trace (drilling predation), fossil herbivory studies often encompass a wide array of traces, as denoted by different damage types. Thus, in fossil herbivory studies, the implementation of a Bayesian frame such as that used by Smith and colleagues would require estimation of the prevalence of a quantity of damage types exceeding the number of specimens from which many plant taxa are known in a given assemblage. For example, at Willershausen, 85 damage types have been observed, and more than 100 host plant taxa are known from only 5–85 specimens. The promise of Bayesian methods is undoubtedly great, and it is not possible to predict the advances that will be seen over the next few decades, but the high ratio of possible damage types to number of specimens per host plant taxon constitutes a major challenge.

Alternatives to Bipartite Networks

Beta Diversity.—Our calculations of balanced turnover and abundance gradients for the two dominant host plants at Willershausen show that these metrics are valid and reliable under the resampling routine and under the routine in which the dataset was subsampled to 3500 leaves (Fig. 6A). At lower levels of sampling, the abundance gradient metric yields a similar mean value, but with much wider confidence intervals. The balanced variation metric becomes less valid and reliable at lower levels of sampling. Unsurprisingly, estimates of balanced turnover and abundance gradients are most valid and reliable when coverage is high.

Among the datasets generated by iteratively resampling the Willershausen data and by subsampling the data to 3500 leaves, coverage estimates do not overlap, but estimates of balanced turnover and abundance gradients overlap almost perfectly (Fig. 6A). However, estimates become much less reliable when the Willershausen dataset is subsampled to only 1000 leaves, and the levels of coverage for *Zelkova ungeri* and *Fagus. sylvatica* fall to 0.91 and 0.86, respectively.

The Colwell Creek Pond data yield much more valid and reliable results (Fig. 6B). This is perhaps unsurprising, because coverage of the second most-abundant host plant is higher at Colwell Creek Pond than at Willershausen. Whereas it is very rare for two host plants within a single assemblage to have such high sample coverage—0.990 for *Auritifolia waggoneri* and 0.989 for *Taeniopteris* spp.—our findings suggest that valid and reliable estimates of balanced turnover and abundance gradients are achievable for those rare assemblages with two host plants that are nearly completely sampled.

The Williamson Drive data yield results that are even more valid and reliable than those for Colwell Creek Pond (Fig. 6C). This is a bit surprising: although the most dominant host plant at Williamson Drive, *Macroneuropteris scheuchzeri*, has sample coverage of 0.991, the second most-dominant host plant, *Sigillariophyllum* foliage, has sample coverage of only 0.948. This is quite a bit less than that of *Taeniopteris* spp. at Colwell Creek Pond, and we do not yet have enough paleontological data to evaluate the consequences of this reduced sample coverage. For Williamson Drive, balanced variation and abundance gradients essentially perform as unbiased and consistent estimators to nearly the same extent as does coverage-based rarefaction (Fig. 4). Further analyses are needed to determine exactly why these two metrics perform somewhat better for the Paleozoic data than for Willershausen—richness of damage types may be a key determinant—and particularly why these metrics perform better for Williamson Drive than for Colwell Creek Pond.

Nevertheless, it is clear that these two components of beta diversity are a preferable alternative to bipartite network metrics. They are more valid and reliable than nearly any bipartite network metric that has been examined for fossil herbivory (Currano et al. 2021; Swain et al. 2022). Their meanings are clear, as is the difference between them. They can be calculated for pairwise comparisons among host plants, or can be used to generate a single

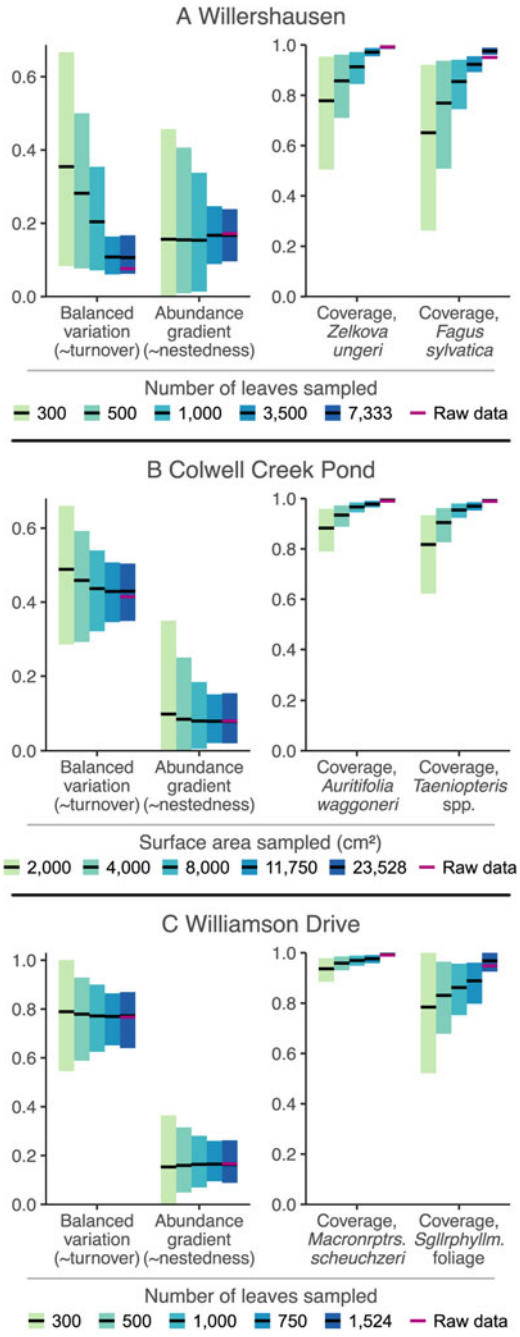

FIGURE 6. Mean values and 95% confidence intervals for beta-diversity metrics generated by resampling and subsampling data for the two most abundant host plants from (A) Willershausen, (B) Colwell Creek Pond, and (C) Williamson Drive. At the highest sample sizes, represented in dark blue, the data were resampled rather than subsampled. *Macronrptrs.scheuchzeri* = *Macroneuropteris scheuchzeri*; *Sgllrphyllm.* = *Sigillariophyllum.*

value for an entire assemblage (Baselga and Orme 2012; Baselga 2017), and can thus be used whether an assemblage contains 2 or 20 host plants with nearly complete sampling.

Host Specificity.—The results of our resampling and subsampling procedures demonstrate that the traditional method for assigning host-specificity scores is strongly biased by sampling completeness: at lower levels of sampling, the host breadth of a damage type inevitably decreases (Fig. 7). For example, in the Colwell Creek Pond resampling routines, we treated each iteration in which the generalist DT032 or DT120 damage type was restricted to only one host plant taxon as a false positive finding of specialization. DT032 appeared on only one host plant taxon in 2.72% of iterations; DT120, in 3.78%. When a finding of specialization requires a damage type to appear on three or more specimens, following the convention established by Wilf and Labandeira (1999), the false positive rate falls to 0.93% for DT032 but remains at 3.34% for DT120.

The inadequacy of the three-specimen threshold for designation of a damage type as "specialized" is shown by the frequencies of false positive results (Fig. 8). These frequencies appear to follow lognormal distributions. For DT032, which was observed on fewer leaves than DT120, $\sigma > 1$ such that the greatest proportion of false positive results occur when this damage type is observed on only one specimen. However, for DT120, $\sigma < 1$ such that 4.7% of false positive results occur when this damage type is observed on only one specimen, 8.7% occur when this damage type is observed on four specimens, and 4.9% occur when this damage type is observed on nine specimens. Thus, the three-specimen threshold protects against only a small fraction of false positives.

Rarefaction of Interactions.—Coverage-based rarefaction of interactions performs as an unbiased and consistent estimator: as sampling completeness decreases, the mean estimate changes negligibly while confidence intervals widen (Fig. 9). Resampled estimates and confidence intervals are often invalid for rarefaction of interactions, because the number of single-

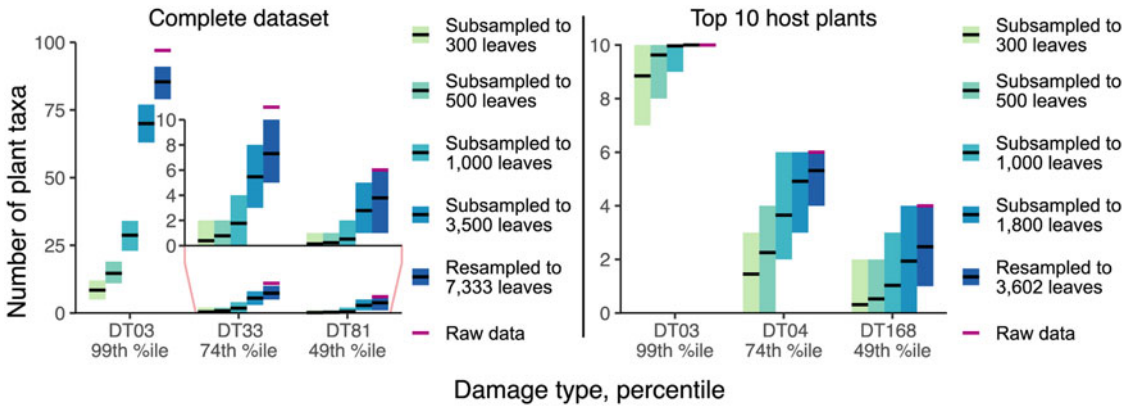

FIGURE 7. Mean values and 95% confidence intervals for the number of plant taxa on which various damage types appear, calculated with the Willershausen dataset.

tons in a resampled dataset tends not to exceed the number of singletons in the original dataset. The number of singletons is one of the main determinants of estimated sample coverage; thus, resampled datasets tend to have higher estimated coverage than the original datasets. This means that coverage-based rarefaction will generate lower estimates for resampled data than for subsampled data. This is abundantly clear for rarefaction of interactions in the simulated dataset and is also quite notable for Williamson Drive. The estimation of confidence limits from iteratively sampled data should therefore be performed with subsampled, rather than resampled, data whenever the mean estimate generated with

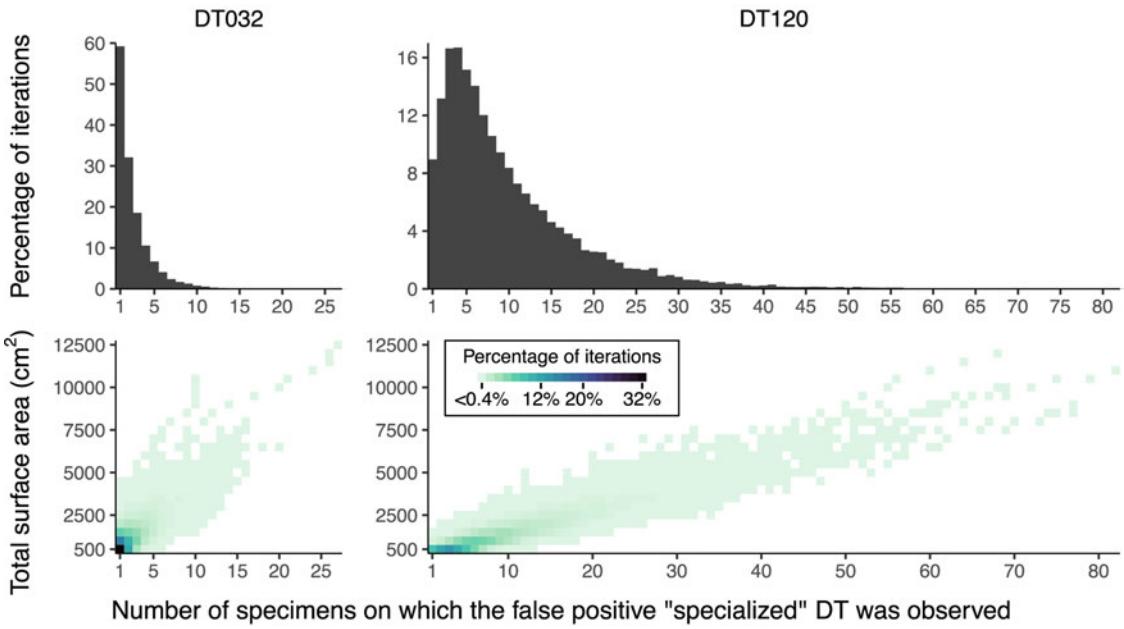

FIGURE 8. False positive results of "specialized" damage generated by iteratively resampling data from Colwell Creek Pond. We treated each iteration in which DT032 or DT120 was observed on only one host plant taxon as a false positive. The heat maps show the percentage of iterations for each amount of subsampled surface area in which a false positive result was recovered, arranged by the number of specimens on which the damage type was observed. The histograms show the summed percentages, by number of specimens.

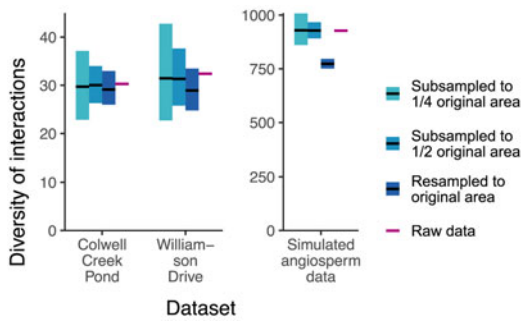

FIGURE 9. Mean values and 95% confidence intervals for coverage-based rarefaction of interactions. The datasets presented here are Williamson Drive and Colwell Creek Pond, both from the Permian of Texas (rarefied to a sample coverage of 0.771) and a simulated dataset that mimics the patterns seen among angiosperms at Willershausen (rarefied to a sample coverage of 0.726).

resampled data is clearly invalid. The methodology of coverage-based rarefaction of interactions is illustrated in Figure 10.

An Example of Bipartite Network Metrics and the Potential for Metric Hacking

While it has been argued that bipartite network metrics allow a more finely resolved, "in-depth" understanding of the relationships between host plants and damage types (Swain et al. 2021), others argue that the multiple comparisons presented in many network studies often contain spurious results (Webber et al. 2020). To evaluate which of these two views of multiple comparisons in network studies is applicable to fossil herbivory datasets, we calculated bipartite network metrics for one of the most iconic and intensely studied series of assemblages in this discipline: Paleocene and Eocene floras of the western interior of North America (Wilf and Labandeira 1999; Currano et al. 2008, 2010). The finding of increased insect herbivory at the PETM is supported by quantitative measures of herbivorized leaf area (Currano et al. 2016) and by damage type diversity, whether rarefied by number of leaves (Currano et al. 2010)—an older practice shown to be biased by differences in leaf surface area among host plant taxa (Schachat et al. 2018)—or rarefied by sample coverage (Schachat et al. 2022). Changes in herbivory at the EECO have not been examined as thoroughly (Currano et al. 2019), but the logic about climate, nutrient availability, and herbivory used to describe the PETM (Currano et al. 2008, 2010) ought to apply to the EECO as well.

When the 28 bipartite network metrics considered here are calculated for the Paleocene–Eocene assemblages of the Bighorn Basin and Wind River Basin (Fig. 11), none of these metrics yield extreme values for the PETM

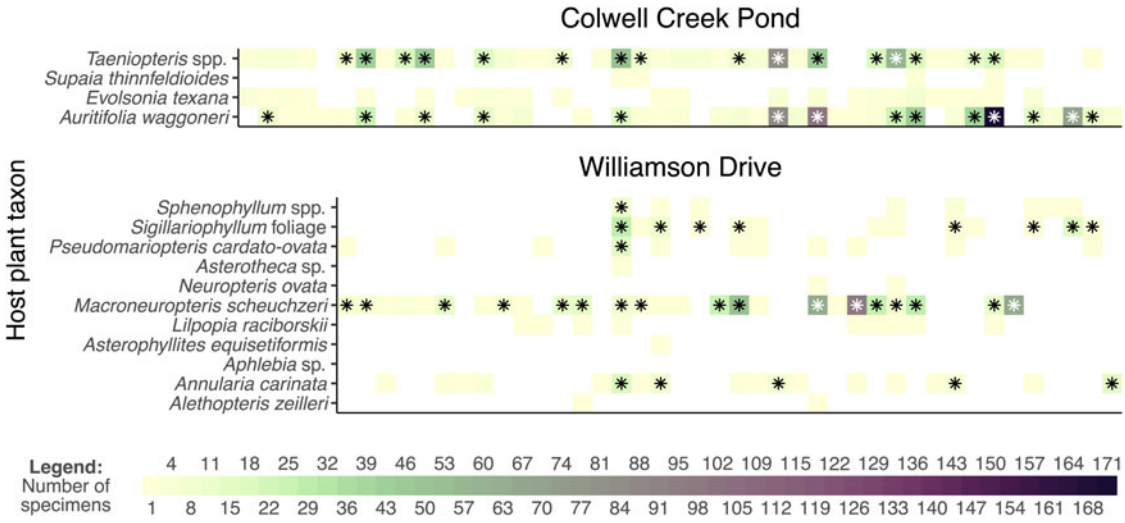

FIGURE 10. Comparison of the raw and rarefied interaction data from Colwell Creek Pond and Williamson Drive. Each column of each graph represents a damage type. The heat maps show the prevalence of each interaction, and the asterisks denote interactions that remain after rarefying data from each assemblage to a sample coverage of 0.771.

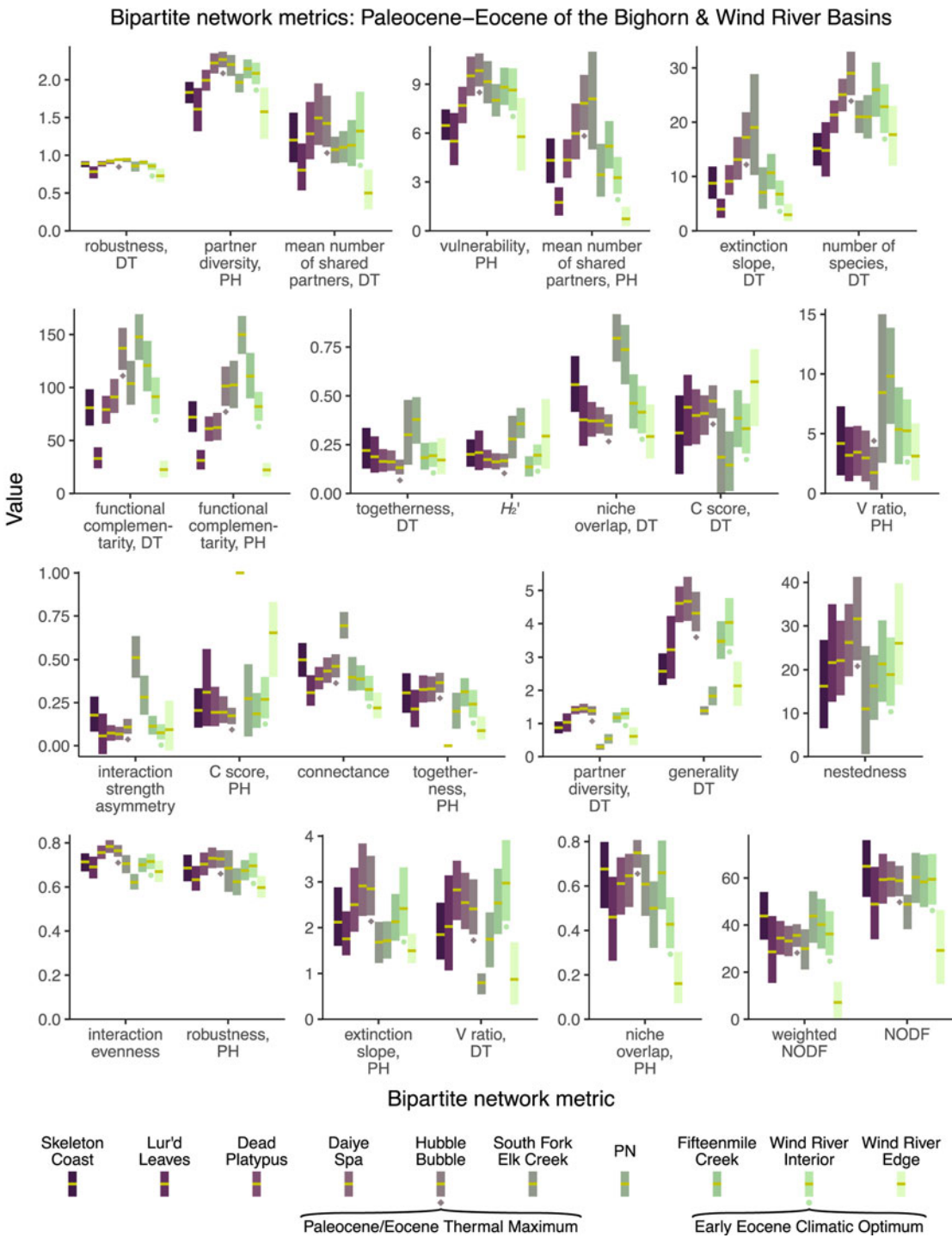

FIGURE 11. Mean values and 95% confidence intervals for bipartite network metrics, generated by subsampling each dataset to 300 leaves. DT, damage type; PH, plant host.

Hubble Bubble assemblage (Currano et al. 2008) or the EECO Wind River Interior assemblage (Currano et al. 2019). If these metrics are taken at face value, rather than being dismissed due to their susceptibility to sampling bias, they suggest that extreme climate change does not have a perceptible impact on plant–insect interactions. For a variety of metrics (interaction strength asymmetry, the C score for host plants, connectance, togetherness, partner diversity for damage types, generality for damage types), it is not the assemblage deposited during the PETM, but the assemblage deposited just afterward, that yields the most extreme values. This assemblage, South Fork of Elk Creek, was immediately noted for having only two host plants preserved in meaningful quantities (Currano et al. 2008; Currano, 2009): a peculiarity that has not been ascribed with ecological significance (Currano et al. 2008, 2010; Currano 2009). However, this long-known peculiarity appears to be driving temporal patterns in approximately one-quarter of bipartite network metrics. (For all other assemblages shown in Fig. 11, the mean number of host plant taxa in each subsampling iteration ranges from 4.7 to 11.9.)

Different combinations of these metrics support different narratives. Of the 28 bipartite network metrics, approximately one-third suggest that the PETM and EECO had dissimilar impacts on the relationship between host plants and damage types, approximately one-third suggest that the PETM and EECO had similar impacts, and approximately one-third yield inconclusive results (Fig. 11, Table 2). The PETM itself yields a variety of possible conclusions. More than two-thirds of these metrics suggest that the relationship between host plants and damage types did not drastically change from the very late Paleocene to the PETM, and less than one-quarter are inconclusive (Fig. 11, Table 2). The only two metrics that suggest a drastic change in the relationship between host plants and damage types at the PETM—functional complementarity for host plants and for damage types—are the two metrics that show the greatest amount of spread overall (Figs. 3, 5, 11). Moving from the PETM into the Eocene, more than one-quarter of these metrics suggest that the

TABLE 2. The variety of narratives about the PETM supported by different combinations of bipartite network metrics.

Intervals being compared	Metrics that suggest little or no difference	Metrics that suggest a drastic difference	Metrics with inconclusive results
PETM (Hubble Bubble; Currano et al. 2008) and EECO (Wind River Interior; Currano et al. 2019)	robustness DT, mean number of shared partners DT, interaction strength asymmetry, partner diversity DT, generality DT, robustness PH, extinction slope PH, V ratio DT, weighted NODF, NODF	mean number of shared partners PH, extinction slope DT, number of species DT, functional complementarity DT, connectance, togetherness PH, interaction evenness, niche overlap PH	partner diversity PH, vulnerability PH, functional complementarity PH, togetherness DT, H2, niche overlap DT, C score DT, V ratio PH, C score PH, nestedness
Latest Paleocene (Daiye Spa; Currano et al. 2008) and PETM (Hubble Bubble; Currano et al. 2008)	partner diversity PH, mean number of shared partners DT, vulnerability PH, togetherness DT, H2, niche overlap DT, C score DT, V ratio PH, C score PH, connectance, togetherness PH, partner diversity DT, generality DT, nestedness, interaction evenness, robustness PH, extinction slope PH, V ratio DT, weighted NODF, NODF	functional complementarity DT, functional complementarity PH	robustness DT, mean number of shared partners PH, extinction slope DT, number of species DT, interaction strength asymmetry, niche overlap PH
PETM (Hubble Bubble; Currano et al. 2008) and earliest Eocene (South Fork of Elk Creek; Currano et al. 2008)	robustness DT, partner diversity PH, vulnerability PH, mean number of shared partners PH, extinction slope DT, functional complementarity PH, robustness PH, weighted NODF	number of species DT, niche overlap DT, interaction strength asymmetry, C score PH, connectance, togetherness PH, partner diversity DT, generality DT, interaction evenness, extinction slope PH, V ratio DT, NODF	mean number of shared partners DT, functional complementarity DT, togetherness DT, H2, C score DT, V ratio PH, nestedness, niche overlap PH

relationship between host plants and damage types did not change from the PETM to its immediate aftermath, more than one-third suggest that this relationship did indeed change, and more than one-quarter are inconclusive (Fig. 11, Table 2).

The only metric that returns a more extreme value for the PETM than for the two assemblages that immediately predate and postdate it—that is, the mean value for the PETM lies beyond the 95% confidence intervals for any of these four other assemblages—is "number of species, DT." We have presented this metric here as if it were a bipartite network metric, because it was previously reported as such (Currano et al. 2021; Swain et al. 2022), and because it is calculated with the *networklevel* function in the bipartite package in R (Dormann et al. 2008). However, this is not truly a bipartite network property, in that it does not respond to the distribution of damage types among the host plants.

Bipartite network properties fail to identify the PETM as an anomaly. This finding necessitates a reckoning as to whether bipartite network analysis provides additional nuance and context to traditional metrics such as the herbivory index and rarefied damage type diversity, or alternatively, whether these metrics are too biased at realistic sample sizes to provide results that warrant interpretation. If the canonical notion of uniquely intense and diverse insect herbivory at the PETM is erroneous, that notion should of course be challenged. But, for the many reasons detailed earlier, the various narratives that emerge from bipartite network analysis that contradict the accepted influence of the PETM on insect herbivory are quite likely artifacts of sampling incompleteness and unevenness.

Conclusions

The challenge of linking host plants to damage types through bipartite network analysis is threefold. First, sampling incompleteness does not simply cause increased uncertainty, as is the case for consistent and unbiased estimators such as the herbivory index or coverage-based rarefaction of damage type diversity; instead, sampling incompleteness typically leads to inaccurate, misleading results. Second, the wide variety of bipartite network metrics creates many opportunities for HARKing. Those opportunities are exacerbated by the unclear meanings of these metrics. And third, many damage types violate both the taxonomic species concept and the trophic species concept, depriving specialization of its ecological meaning in this context.

No amount of sampling completeness can remove the potential for HARKing presented by bipartite network analysis, but our results show that alternative methods that are insusceptible to HARKing can be used to evaluate host specificity, to compare component communities, and to measure the diversity of interactions at an assemblage. Rarefied interaction richness and the components of beta diversity are much more likely than bipartite network metrics to perform as unbiased and consistent estimators and do not require complete sampling of damage types across all host plants at an assemblage. Much essential information is still lacking: the exact sample coverage required for valid measurement of abundance gradients, balanced variation, and the diversity of interactions; as well as the surface area data required for evaluation of host specificity, which are unavailable for most published assemblages. However, the first step is understanding which analyses are meaningful and which measurements are needed for those analyses to be valid.

At present, there are a number of large gaps in our knowledge of fossil herbivory. First is the nearly complete lack of Pennsylvanian or Jurassic assemblages examined for herbivory and the lack of early to mid-Cretaceous assemblages. Second is the general lack of assemblages examined from tropical latitudes. Third is the widespread lack of surface area measurements, which are necessary for evaluating the intensity of herbivory (Schachat et al. 2018). Fourth is the widespread lack of counts of the number of times that each damage type appears on each leaf, termed "feeding event occurrences." These data can be used to evaluate various hypotheses about the causes of increased herbivory (Schachat et al. 2022). In light of the limited amount of time that paleontologists are able to spend collecting fossil herbivory data, we believe that addressing these four gaps is the most important use of investigator effort.

Acknowledgments

We thank three reviewers who provided helpful feedback that improved our article. We thank Conrad Labandeira for extensive feedback. We thank all researchers who have collected fossil herbivory data, especially Benjamin Adroit and colleagues who collected and shared the Willershausen dataset.

Declaration of Competing Interests

The authors declare no competing interests.

Literature Cited

Adroit, B., V. Girard, L. Kunzmann, J.-F. Terral, and T. Wappler. 2018. Plant–insect interactions patterns in three European paleoforests of the late-Neogene—early-Quaternary. PeerJ 6:e5075.

Allen, S. E., A. J. Lowe, D. J. Peppe, and H. W. Meyer. 2020. Paleoclimate and paleoecology of the latest Eocene Florissant flora of central Colorado, USA. Palaeogeography, Palaeoclimatology, Palaeoecology 551:109678.

Anderson, J. 1999. Towards Gondwana alive, Vol. 1. Gondwana Alive Society, Pretoria.

Arceo-Gómez, G., C. Alonso, T.-L. Ashman, and V. Parra-Tabla. 2018. Variation in sampling effort affects the observed richness of plant–plant interactions via heterospecific pollen transfer: implications for interpretation of pollen transfer networks. American Journal of Botany 105:1601–1608.

Baselga, A. 2010. Partitioning the turnover and nestedness components of beta diversity. Global Ecology and Biogeography 19:134–143.

Baselga, A. 2017. Partitioning abundance-based multiple-site dissimilarity into components: balanced variation in abundance and abundance gradients. Methods in Ecology and Evolution 8:799–808.

Baselga, A., and C. D. L. Orme. 2012. betapart: an R package for the study of beta diversity. Methods in Ecology and Evolution 3:808–812.

Basset, Y. 1992. Host specificity of arboreal and free-living insect herbivores in rain forests. Biological Journal of the Linnean Society 47:115–133.

Basset, Y., G. Samuelson, and S. E. Miller. 1996. Similarities and contrasts in the local insect faunas associated with ten forest tree species of New Guinea. Pacific Science 50:157–183.

Beck, A. L., and C. C. Labandeira. 1998. Early Permian insect folivory on a gigantopterid-dominated riparian flora from north-central Texas. Palaeogeography, Palaeoclimatology, Palaeoecology 142:139–173.

Bennett, J. M., A. Thompson, I. Goia, R. Feldmann, V. Ştefan, A. Bogdan, D. Rakosy, M. Beloiu, I.-B. Biro, S. Bluemel, M. Filip, A.-M. Madaj, A. Martin, S. Passonneau, D. P. Kalisch, G. Scherer, and T. M. Knight. 2018. A review of European studies on pollination networks and pollen limitation, and a case study designed to fill in a gap. AoB PLANTS 10:ply068.

Bissonette, J. A. 2021. Big data, exploratory data analyses and questionable research practices: suggestion for a foundational principle. Wildlife Society Bulletin 45:366–370.

Blasco-Moreno, A., M. Pérez-Casany, P. Puig, M. Morante, and E. Castells. 2019. What does a zero mean? Understanding false, random and structural zeros in ecology. Methods in Ecology and Evolution 10:949–959.

Blüthgen, N. 2010. Why network analysis is often disconnected from community ecology: a critique and an ecologist's guide. Basic and Applied Ecology 11:185–195.

Blüthgen, N., F. Menzel, and N. Blüthgen. 2006. Measuring specialization in species interaction networks. BMC Ecology 6:9.

Blüthgen, N., J. Fründ, D. P. Vázquez, and F. Menzel. 2008. What do interaction network metrics tell us about specialization and biological traits. Ecology 89:3387–3399.

Brogniart, A. 1828. Prodrome d'une histoire des végétaux fossiles. F. G. Levrault, Paris.

Burkle, L., and R. Irwin. 2009. The importance of interannual variation and bottom–up nitrogen enrichment for plant–pollinator networks. Oikos 118:1816–1829.

Cairncross, B., and J. M. Anderson. 1995. Palaeoecology of the Triassic Molteno formation, Karoo basin, south Africa-sedimentological and palaeontological evidence. South African Journal of Geology 98:452–478.

Carpenter, F. M. 1997. Insecta. Pp. 184–193 *in* C. W. Shabica, and A. A. Hay, eds. Richardson's guide to the fossil fauna of Mazon Creek. Northeastern Illinois University, Chicago.

Carvalho, M. R., P. Wilf, H. Barrios, D. M. Windsor, E. D. Currano, C. C. Labandeira, and C. A. Jaramillo. 2014. Insect leaf-chewing damage tracks herbivore richness in modern and ancient forests. PLoS ONE 9:e94950.

Chaney, D. S., S. H. Mamay, W. A. DiMichele, and H. Kerp. 2009. Auritifolia gen. nov., probable seed plant foliage with comioid affinities from the Early Permian of Texas, U.S.A. International Journal of Plant Sciences 170:247–266.

Chao, A., and L. Jost. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. Ecology 93:2533–2547.

Cleal, C. J., C. H. Shute, and E. L. Zodrow. 1990. A revised taxonomy for Palaeozoic neuropterid foliage. Taxon 39:486–492.

Constenius, K. N., M. R. Dawson, H. G. Pierce, R. C. Walter, and M. V. Wilson. 1989. Reconnaissance paleontologic study of the Kishenehn Formation, northwestern Montana and southeastern British Columbia. Montana Geological Society: 1989 Field Conference Guidebook: Montana Centennial Edition: Geologic Resources of Montana 1:189–203.

Correia, P., A. R. Bashforth, Z. Šimůnek, C. J. Cleal, A. A. Sá, and C. C. Labandeira. 2020. The history of herbivory on sphenophytes: a new calamitalean with an insect gall from the Upper Pennsylvanian of Portugal and a review of arthropod herbivory on an ancient lineage. International Journal of Plant Sciences 181:387–418.

Costa, J. M., L. P. da Silva, J. A. Ramos, and R. H. Heleno. 2016. Sampling completeness in seed dispersal networks: when enough is enough. Basic and Applied Ecology 17:155–164.

Currano, E. D. 2009. Patchiness and long-term change in early Eocene insect feeding damage. Paleobiology 35:484–498.

Currano, E. D., P. Wilf, S. L. Wing, C. C. Labandeira, E. C. Lovelock, and D. L. Royer. 2008. Sharply increased insect herbivory during the Paleocene–Eocene Thermal Maximum. Proceedings of the National Academy of Sciences USA 105:1960–1964.

Currano, E. D., C. C. Labandeira, and P. Wilf. 2010. Fossil insect folivory tracks paleotemperature for six million years. Ecological Monographs 80:547–567.

Currano, E. D., R. Laker, A. G. Flynn, K. K. Fogt, H. Stradtman, and S. L. Wing. 2016. Consequences of elevated temperature and pCO_2 on insect folivory at the ecosystem level: perspectives from the fossil record. Ecology and Evolution 6:4318–4331.

Currano, E. D., E. R. S. Pinheiro, R. Buchwaldt, W. C. Clyde, and I. M. Miller. 2019. Endemism in Wyoming plant and insect herbivore communities during the early Eocene hothouse. Paleobiology 45:421–439.

Currano, E. D., L. E. Azevedo-Schmidt, S. A. Maccracken, and A. Swain. 2021. Scars on fossil leaves: an exploration of ecological patterns in plant–insect herbivore associations during the Age of Angiosperms. Palaeogeography, Palaeoclimatology, Palaeoecology 582:110636.

Dayvault, R. D., L. A. Codington, D. Kohls, W. D. Hawes, P. M. Ott, and D. Behnke. 1995. Fossil insects and spiders from three locations in the Green River Formation of the Piceance Creek Basin, Colorado. Pp. 97–116 *in* The Green River Formation in Piceance Creek and Eastern Uinta Basins. Grand Junction Geological Society, Grand Junction, Colo.

Deng, W., T. Su, T. Wappler, J. Liu, S. Li, J. Huang, H. Tang, S. L. Low, T. Wang, H. Xu, X. Xu, P. Liu, and Z. Zhou. 2020. Sharp changes in plant diversity and plant-herbivore interactions during the Eocene–Oligocene transition on the southeastern Qinghai-Tibetan Plateau. Global and Planetary Change 194:103293.

Ding, Q., C. C. Labandeira, and D. Ren. 2014. Biology of a leaf miner (Coleoptera) on *Liaoningocladus boii* (Coniferales) from the Early Cretaceous of northeastern China and the leaf-mining biology of possible insect culprit clades. Arthropod Systematics and Phylogeny 72:281–308.

Ding, Q., C. C. Labandeira, Q. Meng, and D. Ren. 2015. Insect herbivory, plant–host specialization and tissue partitioning on mid-Mesozoic broadleaved conifers of Northeastern China. Palaeogeography, Palaeoclimatology, Palaeoecology 440:259–273.

Diserud, O. H., and S. Engen. 2000. A general and dynamic species abundance model, embracing the lognormal and the gamma models. American Naturalist 155:497–511.

Dobruskina, I. A. 1995. Keuper (Triassic) Flora from middle Asia (Madygen, southern Fergana). New Mexico Museum of Natural History and Science Bulletin 5:1–49.

Donovan, M. P., P. Wilf, C. C. Labandeira, K. R. Johnson, and D. J. Peppe. 2014. Novel insect leaf-mining after the end-Cretaceous extinction and the demise of Cretaceous leaf miners, Great Plains, USA. PLoS ONE 9:e103542.

Dorado, J., D. P. Vázquez, E. L. Stevani, and N. P. Chacoff. 2011. Rareness and specialization in plant–pollinator networks. Ecology 92:19–25.

Doré, M., C. Fontaine, and E. Thébault. 2021. Relative effects of anthropogenic pressures, climate, and sampling design on the structure of pollination networks at the global scale. Global Change Biology 27:1266–1280.

Dormann, C. F., B. Gruber, and J. Fründ. 2008. Introducing the bipartite package: analysing ecological networks. R News 8:2–11.

Dormann, C. F., J. Fründ, N. Blüthgen, and B. Gruber. 2009. Indices, graphs and null models: analyzing bipartite ecological networks. Open Ecology Journal 2:7–24.

Douglas, S. D., and R. A. Stockey. 1996. Insect fossils in middle Eocene deposits from British Columbia and Washington State: faunal diversity and geological range extensions. Canadian Journal of Zoology 74:1140–1157.

D'Rozario, A., C. Labandeira, W. Y. Guo, Y. F. Yao, and C. S. Li. 2011. Spatiotemporal extension of the Euramerican *Psaronius* component community to the Late Permian of Cathaysia: in situ coprolites in a *P. housuoensis* stem from Yunnan Province, southwest China. Palaeogeography, Palaeoclimatology, Palaeoecology 306:127–133.

Dunne, J. A., C. C. Labandeira, and R. J. Williams. 2014. Highly resolved early Eocene food webs show development of modern trophic structure after the end-Cretaceous extinction. Proceedings of the Royal Society of London B 281:20133280.

Dyer, L. A., T. R. Walla, H. F. Greeney, J. O. Stireman III, and R. F. Hazen. 2010. Diversity of interactions: a metric for studies of biodiversity. Biotropica 42:281–289.

Emer, C., J. Memmott, I. P. Vaughan, D. Montoya, and J. M. Tylianakis. 2016. Species roles in plant–pollinator communities are conserved across native and alien ranges. Diversity and Distributions 22:841–852.

Feistmantel, O. 1886. The fossil flora of the Gondwana System—Part 2. The fossil flora of some of the coal fields in western Bengal. Memoirs of the Geological Survey of India, Palaeontologia Indica, series 12 4:1–71.

Feng, Z., J. Wang, R. Rößler, A. Ślipiński, and C. Labandeira. 2017. Late Permian wood-borings reveal an intricate network of ecological relationships. Nature Communications 8:556.

Florin, R. 1936. Die fossilen Ginkgophyten aus Franz-Joseph-Land nebst Erörterungen über vermeintliche Cordaitalesmesozoischen Alters. I. Allgemeiner Teil. Palaeontographica Abteilung B 81:1–72.

Forister, M. L., V. Novotny, A. K. Panorska, L. Baje, Y. Basset, P. T. Butterill, L. Cizek, P. D. Coley, F. Dem, I. R. Diniz, P. Drozd, M. Fox, A. E. Glassmire, R. Hazen, J. Hrcek, J. P. Jahner, O. Kaman, T. J. Kozubowski, T. a Kursar, O. T. Lewis, J. Lill, R. J. Marquis, S. E. Miller, H. C. Morais, M. Murakami, H. Nickel, N. a Pardikes, R. E. Ricklefs, M. S. Singer, A. M. Smilanich, J. O. Stireman, S. Villamarín-Cortez, S. Vodka, M. Volf, D. L. Wagner, T. Walla, G. D. Weiblen, and L. A. Dyer. 2015. The global distribution of diet breadth in insect herbivores. Proceedings of the National Academy of Sciences USA 112:442–447.

Fraser, H., T. Parker, S. Nakagawa, A. Barnett, and F. Fidler. 2018. Questionable research practices in ecology and evolution. PLoS ONE 13:e0200303.

Fründ, J., K. S. McCann, and N. M. Williams. 2016. Sampling bias is a challenge for quantifying specialization and network structure: lessons from a quantitative niche model. Oikos 125:502–513.

Gibson, R. H., B. Knott, T. Eberlein, and J. Memmott. 2011. Sampling method influences the structure of plant–pollinator networks. Oikos 120:822–831.

Goldwasser, L., and J. Roughgarden. 1997. Sampling effects and the estimation of food-web properties. Ecology 78:41–54.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40:237–264.

Grande, L. 1984. Paleontology of the Green River Formation, with a review of the fish fauna. Geological Survey of Wyoming Bulletin 63:1–333.

Grand' Eury, C. 1877. La flore carbonifère de la Loire et du centre de la France. Imprimerie Nationale, Paris.

Grass, I., D. G. Berens, F. Peter, and N. Farwig. 2013. Additive effects of exotic plant abundance and land-use intensity on plant–pollinator interactions. Oecologia 173:913–923.

Greenwalt, D., and C. Labandeira. 2013. The amazing fossil insects of the Eocene Kishenehn Formation in northwestern Montana. Rocks and Minerals 88:434–441.

Greenwood, D. R. 1991. The taphonomy of plant macrofossils. Pp. 141–169 *in* S. Donovan, ed. The processes of fossilization. Belhaven Press, London.

Henriksen, M. V., D. G. Chapple, S. L. Chown, and M. A. McGeoch. 2019. The effect of network size and sampling completeness in depauperate networks. Journal of Animal Ecology 88:211–222.

Hoffmann, F. 1827. P. 157 *in* C. Keferstein, ed. Deutschland, geognostisch-geologisch dargestellt, Vol. 4. Verlag des Landes-Industrie-Comptoirs, Weimar.

Hsieh, T. C., K. H. Ma, and A. Chao. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods in Ecology and Evolution 7:1451–1456.

Huang, D. 2016. The Daohugou Biota. China Scientific Books, Shanghai.

Hutson, M. 2018. Artificial intelligence faces reproducibility crisis. Science 359:725–726.

Jordano, P. 2016. Sampling networks of ecological interactions. Functional Ecology 30:1883–1893.

Kemp, J. E., and A. G. Ellis. 2017. Significant local-scale plant-insect species richness relationship independent of abiotic effects in the temperate Cape Floristic Region biodiversity hotspot. PLoS ONE 12:e0168033.

Knor, S., J. Prokop, Z. Kvaček, Z. Janovský, and T. Wappler. 2012. Plant–arthropod associations from the Early Miocene of the

Most Basin in north Bohemia—palaeoecological and palaeoclimatological implications. Palaeogeography, Palaeoclimatology, Palaeoecology 321–322:102–112.
Kotlaba, F. 1963. Tertiary plants from three new localities in southern Slovakia. Acta Musei Nationalis Pragae 19:53–74.
Kuppler, J., T. Grasegger, B. Peters, S. Popp, M. Schlager, and R. R. Junker. 2017. Volatility of network indices due to undersampling of intraspecific variation in plant–insect interactions. Arthropod-Plant Interactions 11:561–566.
Kustatscher, E., J. H. van Konijnenburg-van Cittert, C. V. Looy, C. C. Labandeira, T. Wappler, R. Butzmann, T. Fischer, M. Krings, H. Kerp, and H. Visscher. 2018. The Lopingian (late Permian) flora from the Bletterbach Gorge in the Dolomites, northern Italy: a review. Geo.Alp 14:39–61.
Labandeira, C. C. 1998. Plant–insect associations from the fossil record. Geotimes 43:18–24.
Labandeira, C. C. 2002. Paleobiology of middle Eocene plant–insect associations from the Pacific Northwest: a preliminary report. Rocky Mountain Geology 37:31–59.
Labandeira, C. C., and E. D. Currano. 2013. The fossil record of plant–insect dynamics. Annual Review of Earth and Planetary Sciences 41:287–311.
Labandeira, C. C., P. Wilf, K. R. Johnson, and F. Marsh. 2007. Guide to insect (and other) damage types on compressed plant fossils, Version 3.0. Smithsonian Institution, Washington, D.C.
Labandeira, C. C., S. L. Tremblay, K. E. Bartowski, and L. VanAller Hernick. 2013. Middle Devonian liverwort herbivory and antiherbivore defence. New Phytologist 200:247–258.
Labandeira, C. C., E. Kustatscher, and T. Wappler. 2016. Floral assemblages and patterns of insect herbivory during the Permian to Triassic of northeastern Italy. PLoS ONE 11:e0165205.
Labandeira, C. C., J. M. Anderson, and H. M. Anderson. 2018. Expansion of arthropod herbivory in Late Triassic South Africa: the Molteno Biota, Aasvoëlberg 411 site and developmental biology of a gall. Pp. 623–719 *in* L. H. Tanner, ed. The Late Triassic world: Earth in a time of transition. Springer International, Cham, Switzerland.
Lesquereux, L. 1872. Tertiary flora of North America. Pp. 304–318 *in* F. Hayden, ed. U.S. Geological Survey of Montana and adjacent territories. U.S. Government Printing Office, Washington, D.C.
Lewis, O. T., J. Memmott, J. Lasalle, C. H. C. Lyal, C. Whitefoord, and H. C. J. Godfray. 2002. Structure of a diverse tropical forest insect–parasitoid community. Journal of Animal Ecology 71:855–873.
Linnaeus, C. 1753. Species plantarum. Laurentius Salvius, Stockholm.
Liu, H.-Y., H.-B. Wei, J. Chen, Y. Guo, Y. Zhou, X.-D. Gou, S.-L. Yang, C. Labandeira, and Z. Feng. 2020. A latitudinal gradient of plant–insect interactions during the late Permian in terrestrial ecosystems? New evidence from southwest China. Global and Planetary Change 192:103248.
Lundgren, R., and J. M. Olesen. 2005. The dense and highly connected world of Greenland's plants and their pollinators. Arctic, Antarctic, and Alpine Research 37:514–520.
Lutz, H., U. Kaulfuss, T. Wappler, W. Loehnertz, V. Wilde, D. F. Mertz, J. Mingram, J. L. Franzen, H. Frankenhaeuser, and M. Koziol. 2010. Eckfeld maar: window into an Eocene terrestrial habitat in Central Europe. Acta Geologica Sinica 84:984–1009.
Maccracken, S. A., and C. C. Labandeira. 2020. The Middle Permian South Ash Pasture assemblage of north-central Texas: coniferophyte and gigantopterid herbivory and longer-term herbivory trends. International Journal of Plant Sciences 181:342–362.
MacGinitie, H. 1969. The Eocene Green River flora of northwestern Colorado and eastern Utah. 83:1–202.
Maia, L. F., A. R. Nascimento, and L. D. B. Faria. 2018. Four years host–parasitoid food web: testing sampling effort on trophic levels. Studies on Neotropical Fauna and Environment 53:132–142.
Mamay, S. H. 1989. *Evolsonia*, a new genus of Gigantopteridaceae from the Lower Permian Vale Formation, north-central Texas. American Journal of Botany 76:1299–1311.
Marcon, E., and B. Hérault. 2015. entropart: an R package to measure and partition diversity. Journal of Statistical Software 67:1–26.
Martínez-Delclòs, X., and J. Martinell. 1993. Insect taphonomy experiments. Their application to the Cretaceous outcrops of lithographic limestones from Spain. Kaupia 2:133–144.
Memmott, J., N. D. Martinez, and J. E. Cohen. 2000. Predators, parasitoids and pathogens: species richness, trophic generality and body sizes in a natural food web. Journal of Animal Ecology 69:1–15.
Meyer, H. W. 2003. The fossils of Florissant. Smithsonian Books, Washington, D.C.
Mokam, D. G., C. Djiéto-Lordon, and C.-F. Bilong. 2014. Patterns of species richness and diversity of insects associated With cucurbit fruits in the southern part of Cameroon. Journal of Insect Science 14:248.
Morris, J. 1845. Fossil flora. Pp. 245–250 *in* P. de Strzelecki, ed. Physical descriptions of New South Wales and Van Diemen's Land. Brown, Green and Longmans, London.
Morris, R. J., S. Gripenberg, O. T. Lewis, and T. Roslin. 2014. Antagonistic interaction networks are structured independently of latitude and host guild. Ecology Letters 17:340–349.
Nelson, N. C., K. Ichikawa, J. Chung, and M. M. Malik. 2021. Mapping the discursive dimensions of the reproducibility crisis: a mixed methods analysis. PLoS ONE 16:e0254090.
Neuwirth, E., and R. C. Brewer. 2014. RColorBrewer: ColorBrewer Palettes, R package version 1.1-2. https://cran.r-project.org/web/packages/RColorBrewer/index.html, accessed 10 October 2021.
Newberry, J. S. 1868. Notes on the later extinct floras of North America, with descriptions of some new species of fossil plants from the Cretaceous and tertiary strata. Annals of the Lyceum of Natural History of New York 9:1–76.
Novokshonov, V. G. 1997. Early evolution of scorpionflies (Insecta: Panorpida). Nauka, Moscow.
Novotny, V., Y. Basset, S. E. Miller, P. Drozd, and L. Cizek. 2002. Host specialization of leaf-chewing insects in a New Guinea rainforest. Journal of Animal Ecology 71:400–412.
Novotny, V., S. E. Miller, J. Leps, Y. Basset, D. Bito, M. Janda, J. Hulcr, K. Damas, and G. D. Weiblen. 2004. No tree an island: the plant–caterpillar food web of a secondary rain forest in New Guinea. Ecology Letters 7:1090–1100.
Novotny, V., S. E. Miller, Y. Basset, L. Cizek, K. Darrow, B. Kaupa, J. Kua, and G. D. Weiblen. 2005. An altitudinal comparison of caterpillar (Lepidoptera) assemblages on *Ficus* trees in Papua New Guinea: caterpillars along an altitudinal gradient. Journal of Biogeography 32:1303–1314.
Novotny, V., S. E. Miller, J. Hrcek, L. Baje, Y. Basset, O. T. Lewis, A. J. A. Stewart, and G. D. Weiblen. 2012. Insects on plants: explaining the paradox of low diversity within specialist herbivore guilds. American Naturalist 179:351–362.
O'Boyle, E. H., G. C. Banks, and E. Gonzalez-Mulé. 2017. The chrysalis effect: how ugly initial results metamorphosize into beautiful articles. Journal of Management 43:376–399.
O'Dea, R. E., T. H. Parker, Y. E. Chee, A. Culina, S. M. Drobniak, D. H. Duncan, F. Fidler, E. Gould, M. Ihle, C. D. Kelly, M. Lagisz, D. G. Roche, A. Sánchez-Tójar, D. P. Wilkinson, B. C. Wintle, and S. Nakagawa. 2021. Towards open, reliable, and transparent ecology and evolutionary biology. BMC Biology 19:68.
Oleques, S. S., J. Vizentin-Bugoni, and G. E. Overbeck. 2019. Influence of grazing intensity on patterns and structuring processes in plant–pollinator networks in a subtropical grassland. Arthropod-Plant Interactions 13:757–770.
Olesen, J. M., J. Bascompte, H. Elberling, and P. Jordano. 2008. Temporal dynamics in a pollination network. Ecology 89:1573–1582.

Parker, T., H. Fraser, and S. Nakagawa. 2019. Making conservation science more reliable with preregistration and registered reports. Conservation Biology 33:747–750.

Pedersen, T. L., and F. Crameri. 2020. scico: Colour Palettes Based on the Scientific Colour-Maps, R package version 1.2.0. https://cran.r-project.org/web/packages/scico/index.html, accessed 10 October 2021.

Peguero, G., R. Bonal, D. Sol, A. Muñoz, V. L. Sork, and J. M. Espelta. 2017. Tropical insect diversity: evidence of greater host specialization in seed-feeding weevils. Ecology 98:2180–2190.

Pinheiro, M., B. E. de Abrão, B. Harter-Marques, and S. T. S. Miotto. 2008. Floral resources used by insects in a grassland community in southern Brazil. Brazilian Journal of Botany 31:469–489.

Ponomareva, G., V. Ponomarenko, and S. Naugolnykh. 1998. Chekarda—the locality of Permian fossil plants and insects. Permian University Press, Perm, Russia.

Potonié, H. 1893. Die Flora des Rotliegenden von Thuringen. Kongelige Preussische Geologie 9:1–298.

Prevec, R., C. C. Labandeira, J. Neveling, R. A. Gastaldo, C. V. Looy, and M. Bamford. 2009. Portrait of a Gondwanan ecosystem: a new late Permian fossil locality from KwaZulu-Natal, South Africa. Review of Palaeobotany and Palynology 156:454–493.

Pringle, R. M., and M. C. Hutchinson. 2020. Resolving food-web structure. Annual Review of Ecology, Evolution, and Systematics 51:55–80.

R Development Core Team. 2021. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Regel, E. A. von. 1868. P. 16 *in* A. L. P. P. de Candolle, ed. Prodromus systematis naturalis regni vegetabilis, sive, Enumeratio contracta ordinum generum specierumque plantarum huc usque cognitarium, juxta methodi naturalis, normas digesta, Vol. 2. Sumptibus Sociorum Treuttel et Würtz, Paris.

Reice, S. R. 1974. Environmental patchiness and the breakdown of leaf litter in a woodland stream. Ecology 55:1271–1282.

Ren, D., C. Shih, T. Gao, Y. Wang, and Y. Yao. 2019. Rhythms of insect evolution: evidence from the Jurassic and Cretaceous in northern China. Wiley, Hoboken, N.J.

Retallack, G. J. 1981. Middle Triassic megafossil plants from Long Gully, near Otematata, north Otago, New Zealand. Journal of the Royal Society of New Zealand 11:167–200.

Ribeiro, A. C., G. C. Ribeiro, F. G. Varejão, L. D. Battirola, E. M. Pessoa, M. G. Simões, L. V. Warren, C. Riccomini, and F. J. Poyato-Ariza. 2021. Towards an actualistic view of the Crato Konservat-Lagerstätte paleoenvironment: a new hypothesis as an Early Cretaceous (Aptian) equatorial and semi-arid wetland. Earth-Science Reviews 216:103573.

Root, R. B. 1973. Organization of a plant–arthropod association in simple and diverse habitats: the fauna of collards (*Brassica oleracea*). Ecological Monographs 43:95–124.

Schachat, S. R., C. C. Labandeira, J. Gordon, D. Chaney, S. Levi, M. N. Halthore, and J. Alvarez. 2014. Plant–insect interactions from Early Permian (Kungurian) Colwell Creek Pond, north-central Texas: the early spread of herbivory in riparian environments. International Journal of Plant Sciences 175:855–890.

Schachat, S. R., C. C. Labandeira, and D. S. Chaney. 2015. Insect herbivory from early Permian Mitchell Creek Flats of north-central Texas: opportunism in a balanced component community. Palaeogeography, Palaeoclimatology, Palaeoecology 440:830–847.

Schachat, S. R., C. C. Labandeira, and S. A. Maccracken. 2018. The importance of sampling standardization for comparisons of insect herbivory in deep time: a case study from the late Palaeozoic. Royal Society Open Science 5:171991.

Schachat, S. R., J. L. Payne, C. K. Boyce, and C. C. Labandeira. 2022. Generating and testing hypotheses about the fossil record of insect herbivory with a theoretical ecospace. Review of Palaeobotany and Palynology 297:104564.

Shcherbakov, D. E. 2008. On Permian and Triassic insect faunas in relation to biogeography and the Permian–Triassic crisis. Paleontological Journal 42:15–31.

Slater, B. J., S. McLoughlin, and J. Hilton. 2012. Animal–plant interactions in a Middle Permian permineralised peat of the Bainmedart Coal Measures, Prince Charles Mountains, Antarctica. Palaeogeography, Palaeoclimatology, Palaeoecology 363–364:109–126.

Slater, B. J., S. McLoughlin, and J. Hilton. 2015. A high-latitude Gondwanan Lagerstätte: the Permian permineralised peat biota of the Prince Charles Mountains, Antarctica. Gondwana Research 27:1446–1473.

Smith, D. M., and A. P. Moe-Hoffman. 2007. Taphonomy of Diptera in lacustrine environments: a case study from Florissant Fossil Beds, Colorado. Palaios 22:623–629.

Smith, J. A., J. C. Handley, and G. P. Dietl. 2022. Accounting for uncertainty from zero inflation and overdispersion in paleoecological studies of predation using a hierarchical Bayesian framework. Paleobiology 48:65–82.

Smith-Ramírez, C., P. Martinez, M. Nuñez, C. González, and J. J. Armesto. 2005. Diversity, flower visitation frequency and generalism of pollinators in temperate rain forests of Chiloé Island, Chile. Botanical Journal of the Linnean Society 147:399–416.

Su, T., J. M. Adams, T. Wappler, Y.-J. Huang, F. M. B. Jacques, Y. Liu, and Z. Zhou. 2015. Resilience of plant-insect interactions in an oak lineage through Quaternary climate change. Paleobiology 41:174–186.

Swain, A., L. E. Azevedo Schmidt, S. A. Maccracken, E. D. Currano, J. A. Dunne, C. C. Labandeira, and W. F. Fagan. 2021. Effects of sampling bias on robustness of ecological metrics in fossil plant-damage type association networks. Geological Society of America Abstracts with Programs 53: 212-3.

Swain, A., S. A. Maccracken, W. F. Fagan, and C. C. Labandeira. 2022. Understanding the ecology of host plant–insect herbivore interactions in the fossil record through bipartite networks. Paleobiology 48:239–260.

Trøjelsgaard, K., P. Jordano, D. W. Carstensen, and J. M. Olesen. 2015. Geographical variation in mutualistic networks: similarity, turnover and partner fidelity. Proceedings of the Royal Society of London B 282:20142925.

Unger, F. 1843. Chloris protogaea: Beiträge zur Flora der Vorwelt 3. W. Engelmann, Leipzig, Germany.

Vázquez, D. P., and M. A. Aizen. 2003. Null model analyses of specialization in plant–pollinator interactions. Ecology 84:2493–2501.

Wappler, T. 2010. Insect herbivory close to the Oligocene–Miocene transition—a quantitative analysis. Palaeogeography, Palaeoclimatology, Palaeoecology 292:540–550.

Wappler, T., E. D. Currano, P. Wilf, J. Rust, and C. C. Labandeira. 2009. No post-Cretaceous ecosystem depression in European forests? Rich insect-feeding damage on diverse middle Palaeocene plants, Menat, France. Proceedings of the Royal Society of London B 276:4271–4277.

Wappler, T., C. C. Labandeira, J. Rust, H. Frankenhäuser, and V. Wilde. 2012. Testing for the effects and consequences of mid Paleogene climate change on insect herbivory. PLoS ONE 7.

Webber, Q. M. R., D. C. Schneider, and E. Vander Wal. 2020. Is less more? A commentary on the practice of "metric hacking" in animal social network analysis. Animal Behaviour 168:109–120.

White, D. 1929. Flora of the Hermit Shale, Grand Canyon, Arizona, Vol. 405. Carnegie Institution of Washington, Washington, D.C.

Whittaker, R. H., and S. A. Levin. 1977. The role of mosaic phenomena in natural communities. Theoretical Population Biology 12:117–139.

Wilde, V., and H. Frankenhäuser. 1998. The Middle Eocene plant taphocoenosis from Eckfeld (Eifel, Germany). Review of Palaeobotany and Palynology 101:7–28.

Wilf, P., and C. C. Labandeira. 1999. Response of plant–insect associations to Paleocene–Eocene warming. Science 284:2153–2156.

Wilf, P., C. C. Labandeira, K. R. Johnson, and N. R. Cuneo. 2005. Richness of plant–insect associations in Eocene Patagonia: a legacy for South American biodiversity. Proceedings of the National Academy of Sciences USA 102:8944–8948.

Wilf, P., C. C. Labandeira, K. R. Johnson, and B. Ellis. 2006. Decoupled plant and insect diversity after the end-Cretaceous extinction. Science 313:1112–1115.

Wilson, M. V. H. 1978. Paleogene insect faunas of western North America. Quaestiones Entomologicae 14:13–34.

Wittry, J. 2006. The Mazon Creek fossil flora. Earth Science Clubs of Northern Illinois, Downers Grove, Ill.

Wolfe, J. A., and W. Wehr. 1987. Middle Eocene dicotyledonous plants from Republic, northeastern Washington. U.S. Geological Survey Bulletin 1597:1–25.

Xiao, L., C. C. Labandeira, and D. Ren. 2022. Insect herbivory immediately before the eclipse of the gymnosperms: the Dawangzhangzi plant assemblage of northeastern China. Insect Science 29:1483–1520.

Xu, Q., J. Jin, and C. C. Labandeira. 2018. Williamson Drive: herbivory on a north-central Texas flora of latest Pennsylvanian age shows discrete component community structure, expansion of piercing and sucking, and plant counterdefenses. Review of Palaeobotany and Palynology 251:28–72.

Zemenick, A. T., R. L. Vanette, and J. A. Rosenheim. 2021. Linked networks reveal dual roles of insect dispersal and species sorting for bacterial communities in flowers. Oikos 130:697–707.

Appendix

Calculating *p*-Values for Host Specificity

The absolute amount of surface area examined should be taken into account when determining host specificity, because if the total amount of surface area is very small, the apparent restriction of a damage type to a particular clade of host plants will very possibly be an artifact of insufficient sampling. The relative amount of surface area should be taken into account, because this determines the probability that a damage type would falsely appear to be restricted to a particular clade of host plants.

Consider a hypothetical assemblage in which 100,000 cm^2 of surface area has been examined. If DT001 is restricted to a clade of host plants represented by a mere 500 cm^2 of surface area, and if DT001 is found on all 15 specimens belonging to the clade at this assemblage, then DT001 indeed appears to be specialized. This finding is supported by the large amount of surface area examined, by the moderately high number of specimens on which DT001 has been found, and by the small amount of relative surface area belonging to the plant clade in question, which confers a low probability that all detected incidents of DT001 would be restricted to this clade due to chance alone.

However, at Colwell Creek Pond, the host plant *Auritifolia waggoneri* accounts for greater than 60% of the broadleaf surface area examined. Therefore, especially if the total amount of surface area examined is low, a generalized damage type may appear to be restricted to *A. waggoneri* due to chance alone—particularly if the damage type is observed on only a few specimens. To test the frequency with which this sort of false positive finding of specialized herbivory may occur, we resampled the data from Colwell Creek Pond for the four host plant taxa from this assemblage that unambiguously meet the criteria for inclusion outlined by Swain et al. (2021): *A. waggoneri* (63% of total broadleaf surface area), *Taeniopteris* spp. (28%), *Evolsonia texana* (9%), and *Supaia thinnfeldioides* (1%). Our analysis focuses on two damage types, DT032 and DT120. Both of these damage types occur on all four of these host plants, with distributions that approximate the amount of surface area examined for each host plant: the majority of incidences of each damage type are on *A. waggoneri* (63%–89%), followed by *Taeniopteris* spp. (10%–25%), *E. texana* (1%–10%), and, finally, *S. thinnfeldioides* (1%–3%).

When a damage type is observed only on one clade of host plants at an assemblage, the surface area of those host plants can be used to test the null hypothesis that the damage type is restricted to a certain plant clade simply by chance. The proportion of all surface area examined at the assemblage that belongs to the clade in question—whether it is a genus or species, implying specialized host specificity, or a higher clade, implying intermediate specificity—can be raised to the number of specimens on which the damage type was observed. This process generates a *p*-value that can be used to test the null hypothesis of generalized host specificity. Consider an example in which a damage type appears to have an intermediate host specificity because it occurs only on plants belonging to the same order. If this order accounts for 40% of all surface area examined at the assemblage, and if the damage type has been observed on five specimens, the *p*-value for its host specificity is $0.4^5 = 0.01024$. This value is below 0.05, and thus the damage type has been observed on enough specimens to reject the null hypothesis of generalized host specificity. However, a correction for multiple comparisons, such as the

Bonferroni correction or the Benjamini–Hochberg correction, should be used if this procedure is carried out for more than one damage type.

These findings presented in our "Results and Discussion" section suggest that the more conservative Bonferroni correction should be used instead of the Benjamini–Hochberg correction when host-specificity *p*-values are calculated for multiple damage types. Surface area data from additional assemblages, with as much area as Colwell Creek Pond or more, are needed to determine whether the Benjamini–Hochberg correction will suffice.

Another fundamental, unresolved issue pertaining to the assignment of host-specificity scores is the definition of "specialized" and "intermediate" host specialization. If a damage type occurs on multiple genera within the same family, is it a specialized damage type, because it is restricted to one family, or is it an intermediate damage type, because it occurs on multiple genera? To our knowledge, this question has never been answered, leaving each team of authors to draw the boundaries between specialized, intermediate, and generalized host specificity wherever they please. To our knowledge, the locations of these boundaries are not typically articulated in publications, leading to a lack of reproducibility. Because the majority of herbivorous insects feed on plants belonging to a single family (Forister et al. 2015), we recommend that a damage type that occurs on a single family be considered "specialized" and that a damage type that occurs on multiple families within a single order be considered "intermediate."

We do not advocate assigning host-specificity scores to damage types. For reasons outlined in the "Introduction," specialist herbivores can be largely or entirely responsible for a "generalized" damage type. For reasons outlined in the "Results and Discussion," a "generalized" damage type can appear to be "specialized" due to sampling incompleteness. However, should any research teams continue to assign host-specificity scores, our method for generating *p*-values protects against false positive findings of specialized herbivory, and our recommended boundaries between specialized, intermediate, and generalized host specificity provide an objective, reproducible, working definition.

TABLE A1. A toy example of the input used for bipartite network analysis. For rarefaction of interactions (Dyer et al. 2010), the input would be a vectorized version of this matrix, which could take any of the following forms: [1 5 0 2 2 0 0 6 0 0 1 0 0 1 0 1 0 3 0 1], or [1 5 2 2 6 1 1 1 3 1], or [6 5 3 2 2 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0], or [6 5 3 2 2 1 1 1 1 1].

	DT001	DT002	DT003	DT004
Plant sp. 1	1	5	0	2
Plant sp. 2	2	0	0	6
Plant sp. 3	0	0	1	0
Plant sp. 4	0	1	0	1
Plant sp. 5	0	3	0	1

Considerations for Coverage-based Rarefaction of Interactions

The input used for bipartite network analysis and for rarefaction of interactions is essentially the same (Table A1). Bipartite network analysis uses a matrix in which each row represents a host plant, each column represents a herbivore (or, for fossil herbivory, a damage type), and each cell represents the number of times that a given interaction was observed. In the example shown in Table A1, DT001 was observed on one specimen belonging to plant sp. 1 and DT002 was observed on five specimens belonging to plant sp. 1. For rarefaction of interactions, the matrix is vectorized, or transformed into a single row. The information about particular host plants and damage types is removed, only the numbers of observations remain, the ordering of these observations does not matter, and it does not matter whether unobserved interactions with a value of 0 are retained in the vector.

This vector is then used for a subsampling procedure and can be subsampled to a threshold of sample coverage, as Schachat et al. (2022) have advocated. Whereas bipartite network analysis produces misleading results with incomplete sampling by treating rare, undetected interactions as true absences, rarefaction of interactions subsamples the observed interactions such that the rare, undetected interactions are removed from the dataset and thus cannot bias the results. Once the dataset for an assemblage reaches the coverage threshold to which all assemblages are subsampled, additional sampling completeness—revisiting an assemblage that already reaches a sample coverage of 0.9 and collecting additional data until sample coverage reaches 0.95—will

not change the results on average, in contrast to bipartite network analysis. This is because the progression of an unbiased sampling routine will lead to additional observations of common interactions while allowing the observation of new, rare interactions.

In a typical rarefaction analysis in the context of fossil herbivory, the input is a vector that contains the number of specimens upon which each damage type has been observed. For example, if DT001 and DT002 have each been observed on three specimens and DT003 has been observed on one specimen, the input vector would take the form of [3 3 1]. To rarefy the interactions rather than the damage type incidences in this toy example, if DT001 was observed on three specimens belonging to the same plant host and DT002 was observed on two different plant hosts, the input vector would take the form of [3 2 1 1]: the second "3" in the original vector, corresponding to DT002, has been split into a "2", representing two incidences of this damage type on one plant host, and a "1", representing an incidence of this same damage type on a different plant host.

There is a computational issue with increasing the number of values in an input vector that equal 1: this reduces sample coverage (Good 1953). Because scaling rarefaction curves by the number of leaves examined is an inadequate substitute for scaling by the amount of surface area examined (Schachat et al. 2018), coverage-based rarefaction is the only appropriate method for comparing assemblages that lack measurements of surface area. But the sampling completeness that is needed to rarefy damage type diversity (Schachat et al. 2022) will far fall short of the sampling completeness needed to rarefy the diversity of interactions. For example, when we iteratively subsampled the Willershausen dataset to 1000 leaves, sample coverage was as low as 0.599—a level at which comparisons will be grossly underpowered, as discussed by Schachat et al. (2022). Therefore, we evaluated rarefaction of interactions with a simulated dataset.

Host Plants with Sample Coverage above 0.99

The following is a nonexhaustive list of host plants censused for fossil herbivory, for which sample coverage is above 0.99. *Zelkova ungeri* from Willershausen (Adroit et al. 2018); *Macginitiea gracilis* Lesquereux, 1872 (Wolfe and Wehr 1987) from PN (Currano et al. 2010); *Heidiphyllum elongatum* Morris, 1845 (Retallack 1981) from Aasvoëlberg 311 (Labandeira et al. 2018); *Sphenobaiera schenckii* Feistmantel, 1886 (Florin 1936) from Birds River 111 (Labandeira et al. 2018); *Platanus raynoldsii* Newberry, 1868 from Mexican Hat (Wilf et al. 2006; Donovan et al. 2014); *Macroneuropteris scheuchzeri* from Williamson Drive (Xu et al. 2018); *A. waggoneri* from Colwell Creek Pond (Schachat et al. 2014); *Quercus* sp. Linnaeus, 1753 from Longmen (Su et al. 2015).

When coverage equals 1, this is typically misleading, as it most likely signifies that either no damage has been found on the host plant taxon in question (the *coverage* function in the entropart package calculates coverage of 1 when there is no damage) or that the sample size is much too small, which can spuriously lead to no singleton damage types. For example, Fabaceae sp. WW042 at the PN assemblage (Currano et al. 2010) is represented by 16 leaves. Three damage types are observed: DT002 is on two leaves, DT012 is on six leaves, and DT032 is on two leaves. Coverage equals 1. However, if the number of leaves with DT002 is experimentally reduced from two to one, coverage falls from 1 to 0.9111. The only host plant we are aware of for which coverage of 1 is not a spurious artifact is *Quercus* sp. from Longmen (Su et al. 2015). Twelve damage types were found on the 1027 leaves examined. All damage types were found on at least five leaves. If the number of leaf specimens with DT045 is experimentally reduced from five to four, coverage remains at 1. This suggests a rule of thumb for determining whether a high coverage estimate is an artifact: if coverage remains above 0.99 after one leaf specimen with the rarest non-singleton damage type is experimentally removed from the dataset, the coverage estimate is indeed robust. Notably, when we subsampled the Willershausen dataset to at least 1000 leaves and iterated this procedure 10,000 times, coverage never exceeded 0.9972. It therefore appears that all coverage estimates that equal 1 would become slightly lower—if not far lower—with additional sampling. Thus, a coverage estimate of 0.995 is a stronger indicator of complete sampling than is a coverage estimate of 1.