Title: Disentangling trait and phylogenetic signals in species co-occurrence: a Bayesian dyadic regression approach

Author: Kyle C. Rosenblad*[1]

1- Department of Integrative Biology, University of California, Berkeley – Berkeley, CA 94720

*Corresponding author. Email: kyle_rosenblad@berkeley.edu

ORCID: Kyle Rosenblad – 0000-0001-8778-7413

Acknowledgements: KCR was supported by the National Science Foundation Graduate Research Fellowship. David Ackerly, Bruce Baldwin, John Battles, Todd Dawson, and three anonymous reviewers provided feedback that improved the manuscript. Andrew Johnson assisted with a software development problem.

Data & code availability statement: Code and data supporting the analyses in this manuscript are available at https://anonymous.4open.science/r/dyadic_regression_for_co-occurrence-5263/README.md. If the manuscript is accepted, this will be converted to a permanent Zenodo repository. compnet source code is available at https://anonymous.4open.science/r/compnet-880D/README.md.

Conflicts of interest: None.

Keywords: co-occurrence, community assembly, competition, environmental filtering, traits, community phylogenetics, functional traits

Abstract

Aim: Understanding the factors shaping species co-occurrence is a central objective in ecology. Progress is hindered by correlations among species traits and phylogenetic relationships, which confound each other's signals. Regression-based approaches hold promise to ameliorate this issue, but the current toolkit lacks methods that combine suitable likelihoods for co-occurrence data with latent variables that account for non-independence, or pseudoreplication, among species pairs.

Innovation: I introduce an approach for estimating the effects of multiple trait and phylogenetic distances on species co-occurrence via Bayesian dyadic regression. Implemented in the compnet R package, this method disentangles the roles of correlated predictors by jointly estimating their conditional effects. compnet extends existing tools by incorporating latent variables that model dependence among species pairs, as well as response distributions suitable for co-occurrence data.

Main Conclusions: In mechanistic metacommunity simulations featuring trait correlations, compnet recovers variation in the strength of community assembly processes while avoiding false detections for ecologically irrelevant traits. In an empirical analysis of North American freshwater fish assemblages, the model reveals opposing effects on co-occurrence for two correlated traits, which appear weak or absent in models that do not account for the correlation. Simulation tests also show that the latent factor component, which is not included in other ecologically motivated dyadic regression tools, is important for avoiding pseudoreplication and associated risks of false detection. compnet offers a complementary new approach for analyzing co-

occurrence data, revealing the roles of correlated traits while reducing risks of spurious inference.

Introduction

A major goal in ecology and biogeography is to understand the factors shaping species co-occurrence (Diamond 1975, Webb et al. 2002, McGill et al. 2006, Vellend 2010). Many studies examine how co-occurrence varies with trait differences or phylogenetic relatedness. The aim is to identify mechanisms underlying patterns of divergence, whereby co-occurrence increases with pairwise trait or phylogenetic distance, or convergence, whereby co-occurrence decreases with distance (Webb et al. 2002, de Bello et al. 2025). However, a given pattern can reflect multiple processes, including environmental filtering, competitive interactions, facilitation, dispersal limitation, and recent speciation (Cavender-Bares et al. 2004, Mayfield & Levine 2010, Münkemüller et al. 2012, Pigot et al. 2015, Navarro-Cano et al. 2021). As new methods have emerged for analyzing co-occurrence patterns, opportunities have expanded for evaluating candidate mechanisms (Webb et al. 2002, Pigot et al. 2015, Araujo et al. 2025), but methodological gaps persist.

An enduring methodological challenge is how to account for correlations among species traits and phylogenetic relationships (Rohr et al. 2025). Different traits may play different roles in ecological community assembly, and correlations can confound their signals. For example, consider a fish metacommunity. Let gape (mouth) width be a divergence trait driving competitive niche differentiation (MacArthur & Levins 1967, Mihalitsis &

Bellwood 2017). Gape width disparities increase co-occurrence due to decreased overlap in prey use. Let body length be a convergence trait driving habitat filtering. Body length disparities decrease co-occurrence due to dissimilarity in abiotic niches (van der Valk 1981, Kirk et al. 2022). Lastly, let gape width and body length be correlated. To quantify the effect of species' differences in gape width on co-occurrence, it is misleading to simply examine variation in gape width disparities, as greater gape width disparities imply greater body length disparities (Fig. 1b). These effects counteract each other (Fig. 1b). However, if comparisons are only made among species pairs that are equally disparate in body length, then the underlying pattern emerges: co-occurrence increases with gape width disparities (Fig. 1c). Analogous reasoning applies when the aim is to quantify the effect of body length on co-occurrence—gape width must be accounted for. This example demonstrates that disentangling trait and phylogenetic effects on species co-occurrence can be viewed as a problem of stratification, or estimating conditional effects, in pairwise data. Without appropriate stratification, key patterns can be distorted, leading to incorrect inference regarding the processes shaping species co-occurrence.

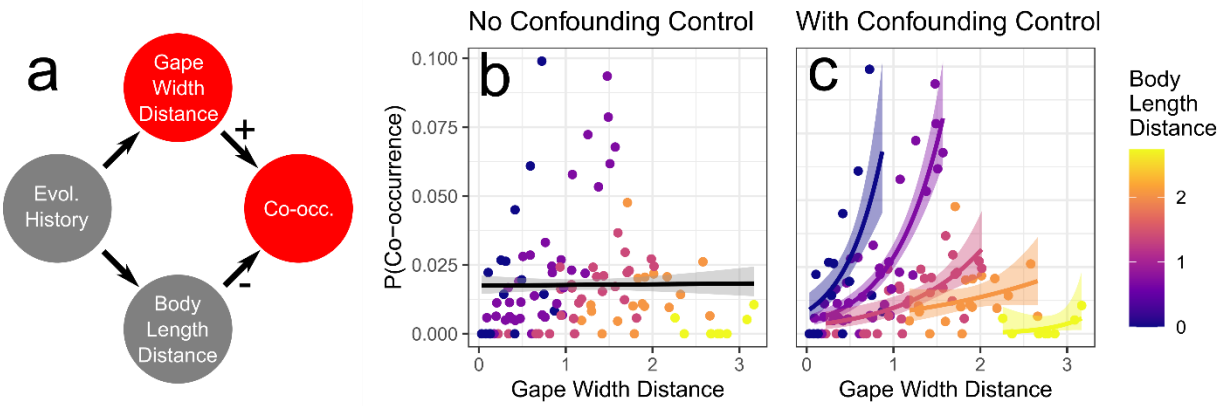


Fig. 1: *In a hypothetical example, controlling for a trait correlation reveals the signal of competitive niche differentiation in co-occurrence data.* **a)** *Conceptual diagram showing how correlated traits can confound inference of community assembly processes. Red highlights the effect of interest. Species that differ more in gape width are more likely to co-occur due to decreased overlap in prey use. Body length has the opposite influence due to habitat filtering. These traits are correlated due to evolutionary history.* **b)** *Naïve estimate of the effect of pairwise distances in gape width on co-occurrence probabilities. The confounding influence of body length obscures the effect.* **c)** *Stratifying by body length distance reveals the positive effect of gape width distance on co-occurrence.*

A powerful approach for estimating conditional effects while controlling for other variables is regression (Harrell 2001, McElreath et al. 2020). In the fish metacommunity example, the units of analysis are species pairs, and the response variable reflects their propensity to co-occur. The predictors are species' pairwise distances in gape width and body length. Each predictor's coefficient represents the effect of varying that predictor while holding others constant (Harrell 2001). This approach mimics the stratified study design in Fig. 1, enabling estimation of the effect of gape width differences on co-occurrence, with body length controlled for (Arif & MacNeil 2023). Thus in principle, a regression-based approach might resolve the issue of correlated traits. However, pairwise co-occurrence data may violate assumptions of standard approaches like GLMs (Minhas et al. 2019). Specialized features may be required to avoid pseudoreplication of species pairs that share a member (e.g., pairs {*i*, *j*} and {*i*, *k*}) and to account for higher-order dependence among observations—e.g., apparent attraction

between species *i* and *j* driven by shared repulsion from *k*. Additionally, the statistical properties of co-occurrence data may necessitate a specialized distribution for the response variable (Veech 2013, Arita 2016, Mainali et al. 2022).

The current methodological toolkit includes some regression-based and regression-like approaches, but each has important limitations in this use case. Partial Mantel tests have been used to estimate associations between co-occurrence and trait or phylogenetic distances (Legendre & Legendre 2012, Elliott & Davies 2017), but these methods can yield inflated false detection rates (Harmon & Glor 2010) and only estimate correlations rather than regression-based effect sizes. Joint species distribution models (JSDMs) can also incorporate traits and phylogenies (Pollock et al. 2014, Ovaskainen et al. 2017). However, species pairs are not the units of anaylsis, and JSDMs are not typically parameterized to directly estimate trait or phylogenetic distance effects on co-occurrence. Phylogenetic generalized linear mixed models (PGLMMs) can directly model associations between co-occurrence and trait or phylogenetic distances, but the relevant model parameters are nested in the covariance matrices of random effects, making interpretation a challenge (Ives & Helmus 2011). Additionally, two separate model parameterizations must be used to account for the possibility that a given trait distance might be positively or negatively associated with co-occurrence. Thus to evaluate N different effects, 2^N PGLMMs must be fit. Regression-free methods are also available, including null model analysis of community-level metrics such as mean nearest neighbor distance (Webb et al. 2002, de Bello et al. 2025). These methods do not typically include a way to disentangle the effects of multiple correlated

133 factors, although strides have been made toward more process-explicit null models
134 (Peres-Neto et al. 2001, Pigot et al. 2015). To my knowledge, no existing method allows
135 direct estimation of conditional effects of multiple trait and phylogenetic distances on co-
136 occurrence while accounting appropriately for dependence among species pairs.
137
138 One approach with potential to fill this gap is dyadic regression. These models are
139 explicitly structured for estimating the effects of multiple predictor variables on a
140 pairwise outcome variable, such as species co-occurrence (Minhas et al. 2019).
141 Pseudoreplication among pairs can be handled via species-level random effects and a
142 latent variable approach known as a latent factor model (Warner et al. 1979, Hoff et al.
143 2002, Hoff 2007). Existing dyadic regression tools, which were developed for the social
144 sciences, can incorporate these features (Hoff 2015, Hoff et al. 2024), but these tools
145 were not designed for co-occurrence data and do not include an appropriate likelihood.
146 Recent ecological work has also highlighted dyadic regression as a promising approach
147 for co-occurrence analysis (Newbury 2024). The example analyses presented use
148 species-level random effects, as well as a Fisher's noncentral hypergeometric (FNCH)
149 likelihood, which has a theoretical justification for modeling co-occurrence data (Veech
150 2013, Arita 2016, Mainali et al. 2022). However, the latent factor model was not
151 included, and past work suggests models without this feature often inadequately
152 account for higher-order dependence among species pairs. This omission can lead to
153 inappropriately narrow uncertainty estimates around effect sizes and thus, potentially,
154 inflated risks of false detection (Hurlbert 1984, Hoff 2007, Minhas et al. 2019). Overall,
155 dyadic regression holds promise as a general approach for disentangling trait and

phylogenetic distance effects on species co-occurrence. However, no existing implementation readily combines a likelihood suited to co-occurrence data with appropriate random effects and latent variables to account for dependence among species pairs.

Here I introduce an approach for disentangling the factors shaping species co-occurrence via Bayesian dyadic regression. The model combines an FNCH likelihood with species-level random effects and a latent factor model used to account for pseudoreplication among species pairs. I also introduce an alternative likelihood using the binomial distribution to address practical limitations of FNCH models identified in simulation testing. I evaluate the performance of both model structures in a suite of simulation studies, demonstrating that these models are well calibrated and recover meaningful effect sizes from mechanistic simulations, while maintaining appropriate false detection rates. In an empirical example with North American freshwater fishes, I show that co-occurrence increases with gape width disparities and decreases with body length disparities, and that these effects are only revealed by accounting for the correlation between these traits. I implement the method in compnet, an open-source R package.

Materials & Methods

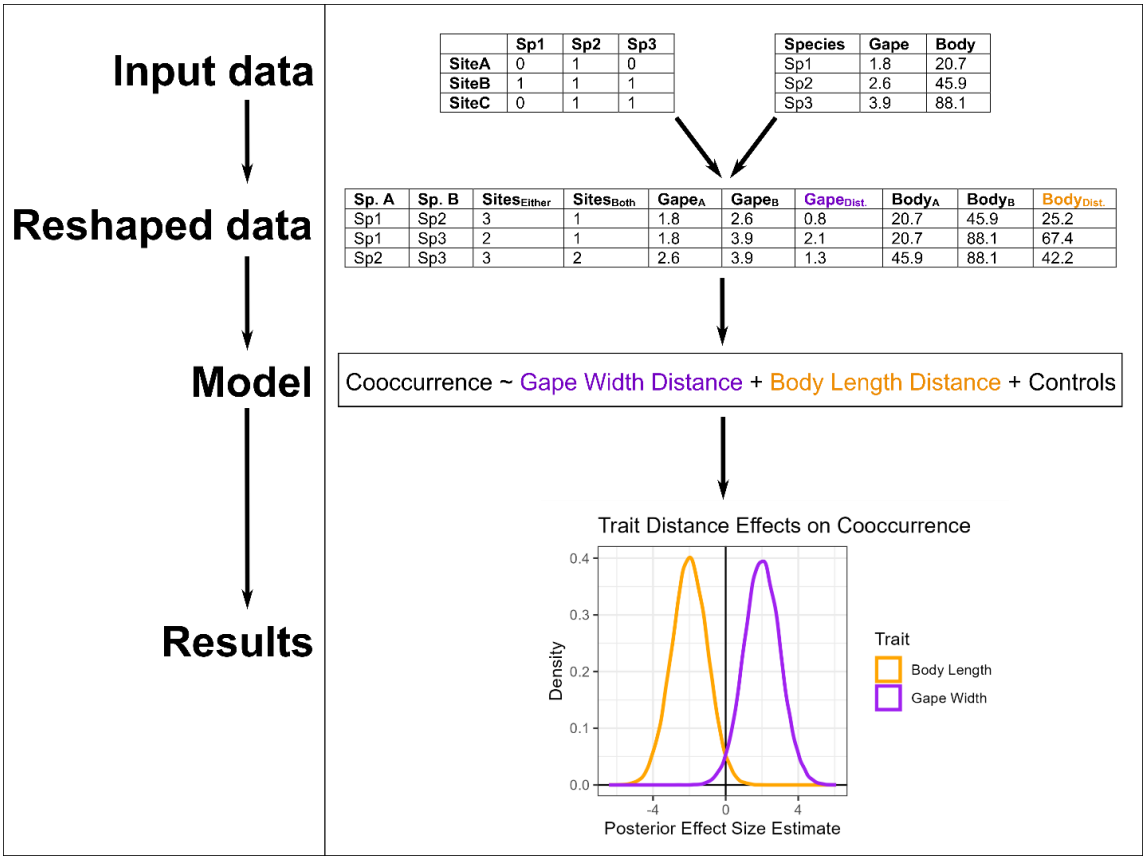
Overview

Model inputs include up to three data types. A species-by-site presence–absence matrix is used to characterize species co-occurrence. Species' trait and phylogenetic

relationships are characterized using a species trait matrix and/or a matrix of pair-level attributes (such as phylogenetic distances) (Fig. 2).

These inputs are combined into a new dataset, with each row representing a species pair. Columns include the identities of each species, their individual occurrence frequencies (i.e., counts of occupied sites), their joint occurrence (i.e., the count of sites where both occur), and any associated trait values or pairwise distances.

This dataset forms the basis for a regression analysis, which estimates how co-occurrence varies with trait and phylogenetic distances. Specialized random effects and latent variables are incorporated to handle dependence among species pairs. (For example, species pairs that share a member might be expected to behave similarly.) These specialized features are described in the Statistical Framework section below.



193

194 **Fig. 2:** *Analytical workflow. Input data include a species-by-site presence-absence*

195 *matrix and a species trait matrix. A matrix of pair-level attributes, such as phylogenetic*

196 *distances, can also be used. These inputs are converted into a new matrix, in which the*

197 *rows are species pairs, and the columns include co-occurrence data, species trait*

198 *values, and pairwise distances. This matrix serves as the input to a regression model,*

199 *which estimates how co-occurrence varies with each trait or phylogenetic distance.*

200 *Positive associations with trait or phylogenetic distance are known as divergence, and*

201 *negative associations are known as convergence. Each of these patterns can reflect a*

202 *variety of community assembly processes.*

203

204 **Statistical framework**

The units of analysis are species pairs or “dyads”. Predictors can include species *i*’s trait values, species *j*’s trait values, and pairwise phylogenetic and trait distances (i.e., absolute value differences). Categorical traits (e.g., flower color) can be incorporated with an interaction term that equals 1 when both species have the same trait value and 0 when they do not. The response variable follows Fisher’s noncentral hypergeometric distribution (Mainali et al. 2022) or “FNCH” henceforth. The FNCH distribution uses a co-occurrence affinity parameter to quantify how strongly the count of co-occurrence sites deviates above or below an expectation based on random site selection by each species. FNCH thus offers a principled way to account for prevalence effects, whereby species’ marginal occurrence frequencies influence co-occurrence counts. For example, a pair of species that both occur at every site in the dataset will have 100% co-occurrence regardless of any attraction or repulsion mechanism. To test ecological hypotheses regarding drivers of co-occurrence, the co-occurrence affinity parameter is modeled as a function of pairwise trait or phylogenetic distances.

Dyadic regression models must also account for non-independence among species pairs. For example, pairs {*i*, *j*} and {*i*, *k*} might behave similarly because they both contain species *i* (Warner et al. 1979). These species-level effects can be modeled as mean-zero offsets drawn from a univariate normal distribution (Warner et al. 1979, Newbury 2024). Higher-order patterns of non-independence can also arise, e.g., “the enemy of my enemy is my friend” (Hoff et al. 2002, Minhas et al. 2019). These patterns can be accounted for using latent factors—i.e., unmeasured species traits inferred from the data, which capture complex patterns of attraction or repulsion. See Hoff et al.

(2002) and Hoff (2007) for theoretical details, and see Appendix S1 for details regarding my approach to latent factor estimation.

The full FNCH model structure is:

$$F_{i\&j} \sim FNCH\left(F_i, N - F_i, F_j, e^{\alpha_{ij}}\right)$$

$$\alpha_{ij} = \beta_0 + \boldsymbol{\beta_p^T x_{ij}} + \boldsymbol{\beta_s^T}\left(\boldsymbol{x_i} + \boldsymbol{x_j}\right) + a_i + a_j + \boldsymbol{u_i^T \Lambda u_j}$$

$$a_1, \ldots a_n \sim i.i.d.\, N(0, \sigma_a^2)$$

$$\boldsymbol{u_i} \sim MVN(0, \boldsymbol{I_K}) \text{ for } i = 1 \ldots n$$

$F_{i\&j}$ is the number of sites where species *i* and *j* co-occur. F_i and F_j are the numbers of sites containing species *i* and *j*, respectively. N is the total number of available sites. α_{ij} is co-occurrence affinity. In the linear predictor, β_0 is a universal intercept. $\boldsymbol{x_{ij}}$ is a vector of pair-level attributes—e.g., trait or phylogenetic distances between species *i* and *j*—with coefficients $\boldsymbol{\beta_p^T}$, which contains the key parameter(s) of interest for inferring influences on co-occurrence. $\boldsymbol{x_i}$ and $\boldsymbol{x_j}$ are vectors of species-level attributes, such as individual species' trait values, with coefficients $\boldsymbol{\beta_s^T}$. a_i and a_j are species-level random intercepts with variance σ_a^2. n is the number of species. $\boldsymbol{u_i}$ and $\boldsymbol{u_j}$ are length-K vectors of species-level latent factor values. $\boldsymbol{\Lambda}$ is a K-by-K diagonal matrix of "factor loadings", which specify the strength of attraction or repulsion between species, as determined by their latent factor values. The model rank, K, is chosen by the user. I follow Minhas et al. (2019) in suggesting that users begin with rank 0 and increase until statistics given

by the ‘gofstats’ function in the compnet R package are sufficient. ‘gofstats’ are percentiles describing the strength of non-independence among observations in the observed data, compared to the expected distribution of data from the fitted model. Extreme values reflect strong non-independence—e.g., here I use an upper threshold of 0.9.

Alternative model versions

Although the FNCH likelihood has a theoretical basis (Veech 2013, Arita 2016, Mainali et al. 2022), it imposes a relatively rigid relationship between the mean and variance of co-occurrence counts. In practice, ecological community data with many unmodeled sources of variation might strain this mean-variance relationship. To account for this possibility, I also developed a model with a binomial likelihood. Rather than modeling co-occurrence affinity, the binomial model focuses on the probability of co-occurrence—i.e., the probability that both species occur at a site, given that at least one is present. Binomial generalized linear mixed models (GLMMs) have been studied extensively, and fit problems arising from unmodeled patterns in the data (e.g., overdispersion, quantile deviations) are often addressed well by adding appropriate random effects (Gelman & Hill 2007). The binomial model structure is:

Below is the binomial dyadic regression model structure:

$$F_{i\&j} \sim Binomial\left(F_{i\&j} + F_{i|j}, P_{ij}\right)$$

$$Logit\left(P_{ij}\right) = \beta_0 + \boldsymbol{\beta_p^T x_{ij}} + \boldsymbol{\beta_s^T}\left(\boldsymbol{x_i} + \boldsymbol{x_j}\right) + a_i + a_j + \boldsymbol{u_i^T \Lambda u_j}$$

$$a_1, \ldots a_n \sim i.i.d.N(0, \sigma_a^2)$$

$$\boldsymbol{u_i} \sim MVN(0, \boldsymbol{I_K}) \text{ for } i = 1 \ldots n$$

$F_{i|j}$ is the number of sites containing only species *i* or only species *j*. P_{ij} is the probability that species *i* and *j* will both occur at a site, given that at least one of the pair is present.

Following the same motivation, I also explored the performance of an alternative FNCH model, which includes an observation-level random intercept drawn from a mean-zero normal distribution. This approach has often been shown to ameliorate fit problems commonly found in GLMMs, such as overdispersion (Gelman & Hill 2007).

R implementation

I implement the above models in the compnet R package, which also includes functions for diagnosing, summarizing, and visualizing model outputs. To install compnet, first follow the installation instructions for the 'rstan' package used for model fitting (Stan Development Team 2024) at https://github.com/stan-dev/rstan/wiki/RStan-Getting-Started. Then, in R, install 'devtools' if needed and run devtools::install_github("[anonymized]/compnet"). Documentation and tutorials are available on the package website at [anonymized].

Model checking

One desirable property of a regression model is the capacity to simulate data that mimic key properties of its input data. Using each of my three model versions (binomial, basic

FNCH, and FNCH with observation-level random effects), I analyzed 20 randomly chosen datasets simulated from a mechanistic metacommunity simulation model (see "Metacommunity simulation study" section below) and checked DHARMa's quantile residual diagnostics (Hartig 2022), as well as compnet's 'gofstats' diagnostic described above in "Statistical framework". See details in Appendix S2.

Runtime

As a simple assessment of computational demands, I logged the runtime for each compnet model used in the above model checking procedure. All models were run on an ASUS StudioBook PC with 32 GB RAM and 12 processor cores (Intel[R] Core[TM] i7-9750H CPU @ 2.60GHz [2.59 GHz]). I analyzed the relationship between runtime and model version using a generalized linear mixed model with a Gamma error distribution and a log link function, which I fit using the brms R package (Bürkner 2017). The linear predictor included a fixed effect for the categorical predictor "model version", as well as a random intercept for "dataset".

Simulation-based calibration

A well-calibrated statistical model should, on average, recover correct parameter values from datasets that were simulated by a model with the same parameter structure. I assessed this capacity for my models in a simulation-based calibration analysis (Talts et al. 2020). See Appendix S3 for details.

Metacommunity simulation study

Simulations

It is also valuable to assess how models perform when confronted with data generated by a more complex simulation model, which may better approximate nature. Toward this aim, I simulated and analyzed metacommunity datasets from modified versions of Thompson et al.'s (2020) temporally explicit, mechanistic simulation model. See Appendix S4.

False detection analysis

An effective approach for modeling co-occurrence data should avoid detecting strong associations with traits that do not influence co-occurrence. To evaluate my method's false detection rates, I simulated 100 metacommunity datasets that included a "dummy trait", which played no role in metacommunity dynamics. The dummy trait was correlated with a "dominance trait", which influenced species' competitive ability. (Species co-occur more frequently when they have similar values of the dominance trait, as coexistence is less stable between weak and strong competitors [Mayfield & Levine 2010].) I examined my model's posterior estimates for the effect size of pairwise distances in the dummy trait. If more than 97.5% of posterior samples were positive, or 97.5% negative, I treated this as detection of a positive or negative effect, respectively. Based on this threshold, I reasoned that false detections should not substantially exceed 5%. Although significance testing is not traditionally a priority in Bayesian statistics (McElreath 2020), two-tailed significance tests with a 5% threshold are frequently used in practice, and this procedure is designed to provide a practical comparison against this widely used standard.

In addition to the above analyses, I also analyzed the simulated metacommunity data using dyadic regression models with rank 0—i.e., with no latent factor model. Past work suggests omitting the latent factor model may amount to pseudoreplication (Hurlbert 1984, Hoff 2007, Minhas 2019), which could yield elevated false detection rates due to inappropriately small uncertainty surrounding the model's effect size estimates.

Effect size analysis

It is also important to confirm that effect size estimates capture meaningful variation in the data-generating process. To assess this capacity in my models, I ran another set of 100 metacommunity simulations, in which I systematically varied the parameter governing the strength of competition. Competitive outcomes were influenced by two correlated species traits: the dominance trait (see above) and the "niche differentiation trait". The niche differentiation trait defines species' differences in resource use, such that pairs with more disparate values are more likely to co-occur (MacArthur & Levins 1967). I tested the expectation that compnet's effect size estimates should be larger with increasing competition strength in the simulation model. Effect sizes are expected to be positive for the niche differentiation trait and negative for the dominance trait.

To assess compnet's performance in detecting phylogenetically structured patterns, I ran an alternate set of metacommunity simulations, in which competitive niche differentiation was determined by phylogenetic distance, such that competitive effects were weaker between more distant relatives (Cavender-Bares et al. 2004). As above, I

tested the expectation that compnet's effect size estimates for phylogenetic distance should be greater in datasets that were simulated with stronger competition.

Empirical example

For an empirical case study, I used data from Kirk et al. (2022). This study, focused on freshwater fishes of the Missouri River basin in western North America, used presence-absence and trait data to explore how habitat filtering varies across environmental gradients. I chose this example to illustrate how a system showing signals of habitat filtering, as found by Kirk et al. (2022), may also reveal signals of other community assembly processes when dyadic regression is used to disentangle the roles of correlated traits. I focused on two traits that show a strong correlation: body length and gape width. Body length is a "life history trait", often showing ties to physiological and demographic rates (Kirk et al. 2022). In contrast, gape width is a "trophic trait", often shaping prey consumption (Kirk et al. 2022). Based on Kirk et al. (2022), I hypothesized that body length would exhibit trait convergence due to habitat filtering. I also hypothesized that, in a model controlling for body length, gape width would show trait divergence due to competitive niche differentiation of prey resources.

Kirk et al.'s (2022) analysis spans roughly 2,000 km in the western United States. I filtered the data to a smaller sub-basin, Horse Creek, spanning roughly 100 km. I took this step because, when the spatial scale of analysis is too large, the pattern typically interpreted as a signal of competitive niche differentiation—i.e., divergent trait values among co-occurring species—can instead reflect recent speciation (Pigot & Etienne

2015). Here, I assume that recent speciation has negligible influence on species' presence-absence patterns at the ~100 km scale. After my filtering step, the data set contained 14 species and 46 sites. To assess the sensitivity of my results to the choice of drainage basin, I repeated the analysis for each of the five most-sampled basins in Kirk et al.'s (2022) data set: Powder, Laramie, Cheyenne, South Platte, and Kansas.

I analyzed the data with three binomial models: one including gape width and maximum body length, one with only gape width, and one with only body length. The purpose of this approach was to assess whether accounting for the correlation between these traits in a two-trait model would yield stronger effect size estimates than single-trait models (Arif & MacNeil 2023).

Software

All data simulation and analysis were done in R (R Core Team 2024). In addition to compnet, the following R packages were used in data simulation and analysis: brms (Bürkner 2017), geiger (Harmon et al. 2008), DHARMa (Hartig 2022), sf (Pebesma 2018 & 2023), and phytools (Revell 2024).

Results:

Model-checking

For the binomial model, residual diagnostics were good for 60% of data sets, fair for 25%, and poor for 15% (Table S1). FNCH models with observation-level random effects

yielded 5% fair diagnostics and 95% poor. FNCH models without observation-level random effects had 100% poor diagnostics.

In all subsequent analyses, FNCH models included observation-level random effects.

Runtime

Binomial models had shorter runtimes (GLMM posterior mean estimate 89 s, 95% CI 63-128 s) than FNCH models (mean estimate 178 s, 95% CI 124-255 s). See Fig. S1.

Simulation-based calibration

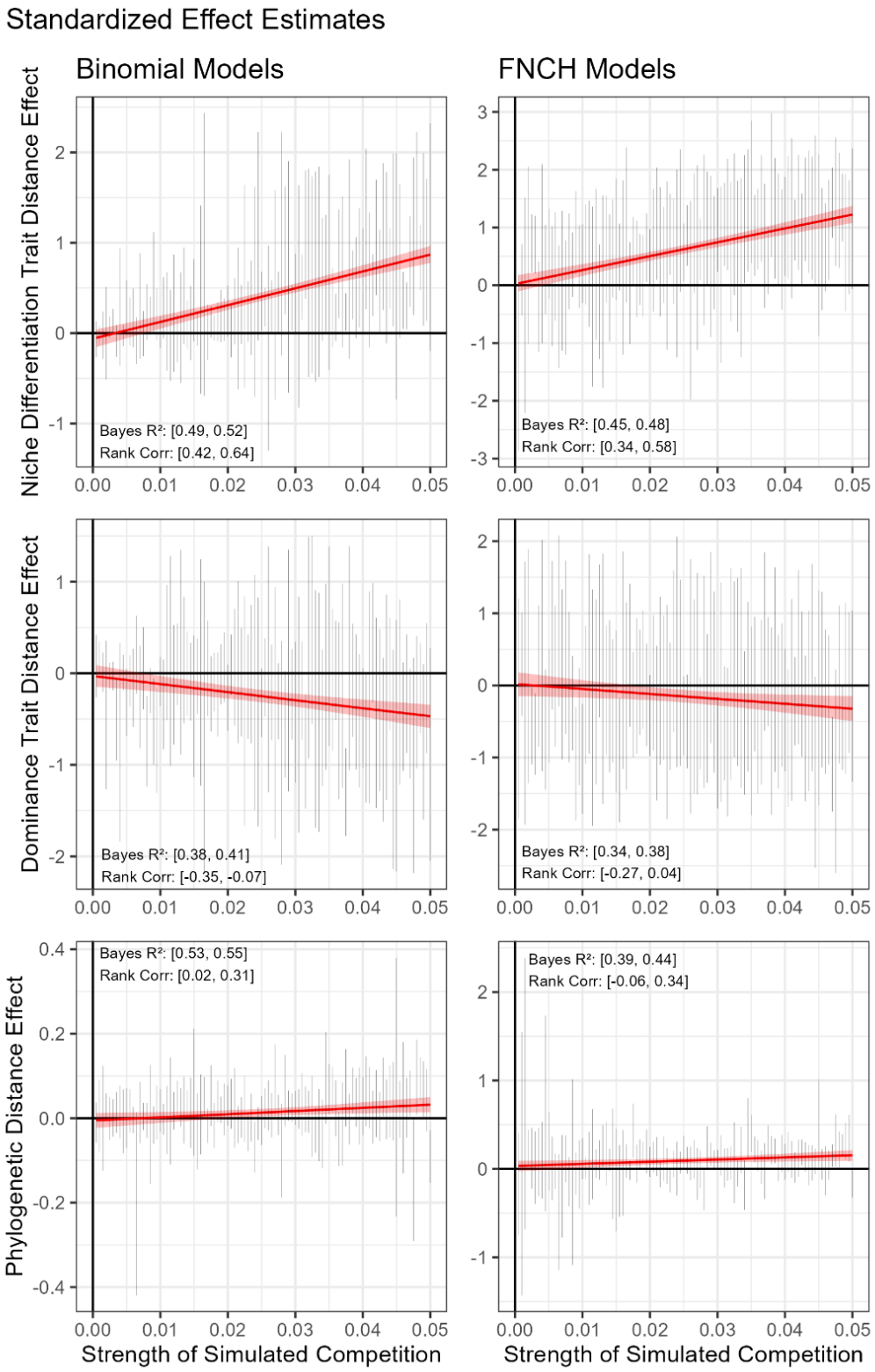
For binomial models, all 4 combinations of parameter values yielded mean parameter estimates (across 16 replicates) that differed negligibly from the true values (Appendix S5). For compnet models, results were similarly strong for all parameters except the latent factor term, which was poorly recovered.

Metacommunity simulation study

Effect size analysis

Both the binomial and FNCH models showed linear relationships, in the hypothesized directions, between simulated competition strength and standardized effect size estimates (Fig. 3). Effect size estimates increased with competition strength for the niche differentiation trait and for phylogenetic distance (a proxy for competitive niche differentiation), and estimates decreased with competition strength, becoming “more negative”, for the dominance trait. Across all trait and phylogenetic distances, the

433 binomial model's effect size estimates yielded a slightly tighter and more consistently
434 monotonic relationship with competition strength, as shown by Spearman's rank
435 correlation and Bayesian R^2 (see Fig. 3 inset text).
436


Standardized Effect Estimates
Binomial Models
FNCH Models
Niche Differentiation Trait Distance Effect
Dominance Trait Distance Effect
Phylogenetic Distance Effect
Bayes R²: [0.49, 0.52]
Rank Corr: [0.42, 0.64]
Bayes R²: [0.45, 0.48]
Rank Corr: [0.34, 0.58]
Bayes R²: [0.38, 0.41]
Rank Corr: [-0.35, -0.07]
Bayes R²: [0.34, 0.38]
Rank Corr: [-0.27, 0.04]
Bayes R²: [0.53, 0.55]
Rank Corr: [0.02, 0.31]
Bayes R²: [0.39, 0.44]
Rank Corr: [-0.06, 0.34]
Strength of Simulated Competition
Strength of Simulated Competition

437

Fig. 3: *Results from an analysis of compnet models' ability to estimate effect sizes reflecting the strength of simulated community assembly processes. Left panels show results for models with a binomial error distribution, and right panels show results for models with Fisher's noncentral hypergeometric distribution (FNCH). The x-axis shows the strength of competition in a metacommunity simulation model. The y-axis shows standardized estimates for effects of species' pairwise trait and phylogenetic distances on co-occurrence. In the simulations, co-occurrence increases with pairwise distances in the niche differentiation trait, as well as phylogenetic distances, and co-occurrence decreases with pairwise distances in the dominance trait. Gray vertical line segments show 95% credible intervals for compnet's effect size estimates. Red lines and 95% credible bands were estimated by a hierarchical Bayesian meta-analysis, which used the individual compnet models' outputs as its inputs. Inset text shows 95% credible intervals for each model's Bayesian R^2 and for the Spearman rank correlation between competition strength and effect size estimates.*

<u>False detection analysis</u>

In 100 binomial compnet models that included a "dummy trait", which played no role in the simulated data-generating process, 2 models yielded significant positive associations with co-occurrence, and 1 yielded a significant negative association. See Table 1. (Significance is defined by over 97.5% of posterior samples being on one side of zero, either positive or negative.) FNCH compnet models yielded no false detections in across 100 analyses.

In models that lacked a latent factor component, false detection rates were substantially greater than the 5% threshold, at 25% for binomial models and 13% for FNCH.

Model version	**Negative False Detections**	**Positive False Detections**
Binomial with latent factor (rank 2)	2%	1%
Fisher's Noncentral Hypergeometric with latent factor (rank 2)	0%	0%
Binomial without latent factor	21%	4%
Fisher's Noncentral Hypergeometric without latent factor	13%	0%

Table 1: *Results of a simulation-based test of the method's false detection rates. Metacommunity data were simulated with a "dummy trait", which played no role in metacommunity dynamics, and then analyzed using compnet models. Negative and positive false detections occurred when the model's estimated posterior probability distribution for the effect size of the dummy trait were over 97.5% negative or positive, respectively. The analysis compared alternative likelihoods using the binomial distribution and Fisher's noncentral hypergeometric distribution.*

Empirical example

A two-trait compnet model found associations between co-occurrence and two correlated traits—gape width and body length—which were greatly attenuated in single-trait models that did not account for the correlation (Fig. 4). In the two-trait model, the standardized effect of gape width distance is 0.35 (Fig. 4b, blue; 95% CI 0.09, 0.60). For maximum body length, which is correlated with gape width (Pearson correlation 0.76), the standardized effect is -0.88 (Fig. 4c, blue; 95% CI -1.37, -0.34). In a single-trait model, the effect for gape width was much weaker, with substantial uncertainty on each side of zero (Fig. 4b, red; mean effect = 0.01; 95% CI -0.15, 0.18). The single-trait body length model yielded a weaker negative effect size at -0.35 (Fig. 4c, red; 95% CI -0.66, 0.01). A sensitivity analysis yielded qualitatively similar results for the five most-sampled drainage basins in Kirk et al'.s (2022) dataset (Fig. S2).

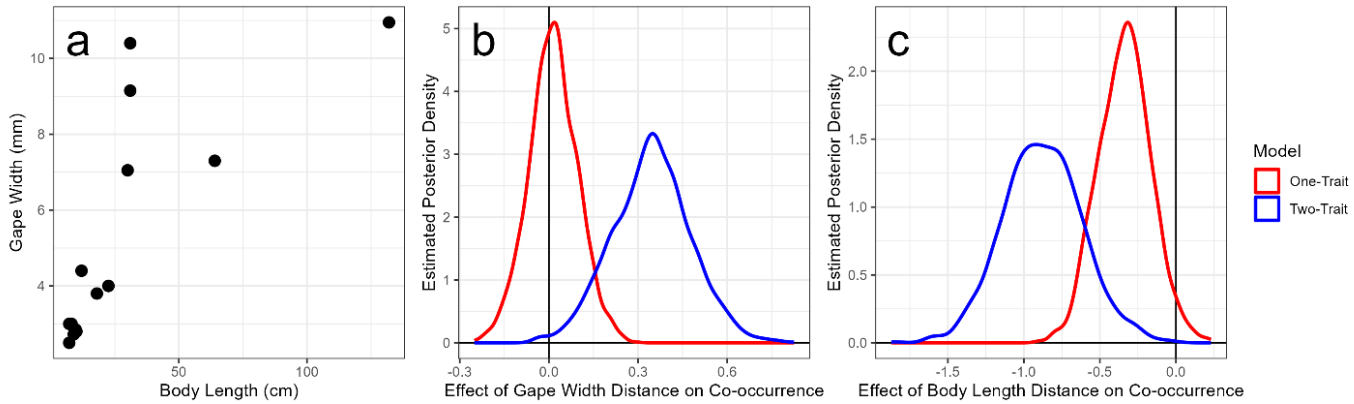


Fig. 4: *Effects of correlated traits on fish species co-occurrence are revealed by joint estimation in a two-trait dyadic regression model. Data are from the Horse Creek drainage basin (Wyoming, USA; Kirk et al. 2022).* **a)** *Species-level association between gape width and body length.* **b)** *Posterior standardized effect size estimates for gape width distance in a one-trait*

model (red) and a two-trait model with body length (blue). **c)** *Posterior standardized effect size estimates for body length distance in a one-trait model (red) and a two-trait model with gape width (blue).*

Discussion

Understanding the factors shaping species co-occurrence is a core goal in ecology. A persistent challenge is how to account for correlations among different factors, such as species traits and phylogenetic relationships, which can distort key patterns and yield incorrect conclusions regarding the underlying ecological processes (Rohr et al. 2025). Here I introduce a Bayesian dyadic regression approach, which disentangles the roles of correlated traits by estimating each trait's effect while conditioning on other variables. I implement my method in the compnet R package, which extends existing tools by combining an appropriate likelihood for co-occurrence data with a latent factor model that accounts for complex patterns of dependence among species pairs. In simulation testing, I show that compnet recovers effect sizes which reflect meaningful variation in the strength of community assembly processes, while minimizing false detections via the latent factor model. In a case study of North American freshwater fishes, I show that species are more likely to co-occur when they have similar body length and dissimilar gape width, and that these effects are only revealed when the correlation between these traits is accounted for. compnet introduces new, complementary functionality to the methodological toolkit for co-occurrence analysis, offering a way to disentangle the influences of correlated traits while avoiding spurious inference.

In a metacommunity simulation study, compnet models' effect size estimates reflected systematic variation in the strength of community assembly processes. Simulated competition was shaped by two correlated traits: the "niche differentiation trait", for which co-occurrence increases with trait disparities, and the "dominance trait", for which co-occurrence decreases with trait disparities. I found that effect size estimates for these traits were larger when simulated competition was stronger, with positive effect sizes for the niche differentiation trait and negative effect sizes for the dominance trait, as hypothesized (Fig. 3). In another analysis, phylogenetic distance served as a proxy for niche differentiation, and the analogous pattern emerged: stronger competition yielded greater estimates for the effect of phylogenetic distance on co-occurrence. These results demonstrate that compnet can detect meaningful effect sizes from datasets generated by more complex processes.

The model also avoids excessive false detections for traits that do not play a role in co-occurrence. In another set of 100 metacommunity simulations, I introduced a "dummy trait" that had no influence on metacommunity dynamics. Posterior effect estimates for pairwise dummy trait distances were nearly always distributed substantially on both sides of zero, indicating little belief in a positive or negative effect. As a quantitative assessment of false detection risk, I computed the proportions of datasets for which compnet's posterior effect estimates were over 97.5% positive or 97.5% negative—a practical Bayesian equivalent of a two-tailed 95% significance test. Binomial models yielded a 3% false detection rate, and FNCH models yielded no false detections. These

findings show that analyzing co-occurrence patterns with compnet carries a minimal, appropriately calibrated risk of erroneously detecting associations by chance.

The false detection study also suggests the latent factor component of the model, which is not included in other ecologically-motivated dyadic regression approaches (Newbury 2024), may be crucial for avoiding pseudoreplication. In binomial and FNCH models without latent factors, false detection rates were 25% and 13% respectively, far above the 5% threshold. Theoretical work suggests latent factors are essential in accounting for higher-order forms of non-independence in dyadic data, such as apparent attraction between species *i* and *j* that is driven by shared repulsion from *k*. These results suggest that when the latent factors intended to capture these patterns are omitted, there may be a severe risk of false detections for ecologically irrelevant traits.

The new method's advantages are further illustrated in an empirical case study of fish communities in Wyoming, USA. Here, a multi-trait model detected increasing co-occurrence with differences in gape width—a potential indicator of prey resource partitioning (MacArthur & Levins 1967, Mihalitsis & Bellwood 2017, Kirk et al. 2022)—that is masked via a correlation with body length, which has the opposite influence. Similarly, the convergence pattern for body length is less than half as strong in a single-trait model that does not account for gape width. Trait convergence in body length is likely caused at least partly by habitat filtering, as indicated by Kirk et al.'s (2022) original analysis, although other mechanisms are possible (Mayfield & Levine 2010, Colwell & Winkler 1984, Münkemüller et al. 2012, Araujo et al. 2025). The above

patterns were consistent in a sensitivity analysis, which explored other drainage basins in the Kirk et al. [2022] dataset [Fig. S2]. These findings underscore the ability of multi-trait dyadic regression models to disentangle the roles of correlated traits in community assembly and species co-occurrence, as well as the risk of misleading conclusions from models that do not account for trait correlations.

In comparing alternative compnet model versions, I found that the binomial model's performance is equivalent to the FNCH model in some ways, and in other ways it offers advantages. Results were very similar between versions for simulation-based tests of effect size estimation (Fig. 3), where binomial models may have a slight advantage, and in tests of false detection rates (Table X), where FNCH models may have a slight advantage. However, binomial models were substantially more computationally efficient (Fig. S1), less likely to suffer fit problems like overdispersion (Table S1), and better able to recover latent variables in simulation-based calibrations (Appendix S5). Consequently, despite the FNCH model's theoretical basis (Mainali et al. 2022, Newbury 2024), I recommend the binomial model as a practical, computationally efficient alternative. FNCH models may also be a desirable option when they are computationally feasible and fit problems are avoided (e.g., overdispersion; see Materials & Methods for details). Although the binomial model does not include features explicitly tailored to account for prevalence effects—i.e., elevated co-occurrence among widespread species, irrespective of a true attraction mechanism—it does include species-level random effects that could, in principle, account for such patterns. The model's strong performance in simulation testing suggests this is likely the case.

In the theoretical realm, this new approach may contribute to the long-standing effort to link traits with coexistence of competing species. Many studies suggest intraspecific competition is stronger than interspecific competition (Adler et al. 2018)—the hallmark of competitive niche differentiation—yet the traits underpinning competitive coexistence remain elusive. Recent theoretical work suggests the relationships may be too complex to estimate directly (Levine et al. 2024), advocating instead for more labor-intensive "process-informed measurements" ("PIMs") like Tilman's (1980) R*. Trait-based null model studies have been cited as evidence that competitive niche differentiation rarely leaves a trace in typical co-occurrence datasets (Levine et al. 2024). However, compnet models can disentangle signals of competitive niche differentiation that may be hidden to other methods due to trait correlations, as demonstrated in the fish community example. Trait correlations are widespread in nature (Kraft et al. 2015, Kunstler et al. 2016, Díaz et al. 2016), suggesting that compnet may open many new opportunities to reveal previously undetected signals of competitive niche differentiation and other community assembly processes.

Conclusions

Here I present a new dyadic regression approach for quantifying associations between pairwise species co-occurrence and multiple predictors, such as traits and phylogenetic relationships. The model structure incorporates a unique combination of features tailored for analysis of species co-occurrence data, yielding strong performance in simulation testing and an empiricial case study. A major strength is the method's ability

to disentangle the influences of correlated trait and phylogenetic distances by jointly estimating their conditional effects. Another important feature is the latent factor model component, which is not included in other ecologically motivated dyadic regression implementations, and which strongly reduces the risk of pseudoreplication and false detection for ecologically irrelevant traits. A user-friendly R package is provided with documentation and tutorials at [https://[anonymized].github.io/compnet/index.html].

References

Adler PB, Smull D, Beard KH, Choi RT, Furniss T, Kulmatiski A *et al.* (2018) Competition and coexistence in plant communities: intraspecific competition is stronger than interspecific competition. *Ecology Letters*, 21, 1319–1329.

Anderson SC, Ward EJ, English PA, Barnett LAK, Thorson JT (2024) sdmTMB: an R package for fast, flexible, and user-friendly generalized linear mixed effects models with spatial and spatiotemporal random fields. *bioRxiv*, 2022.03.24.485545. https://doi.org/10.1101/2022.03.24.485545

Araujo ML, Coelho MTP, Gotelli NJ, Colwell RK, Rangel TF, Dambros CS (2025) In situ speciation of dispersal-limited taxa leads to strong phylogenetic community clustering. *Journal of Biogeography*, 52, e70005. https://doi.org/10.1111/jbi.70005

Arif S, MacNeil MA (2023) Applying the structural causal model framework for observational causal inference in ecology. *Ecological Monographs*, 93, e1554. https://doi.org/10.1002/ecm.1554

Bartholomew DJ, Knott M, Moustaki I (2011) *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley.

Bürkner PC (2017) brms: an R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80, 1–28. https://doi.org/10.18637/jss.v080.i01

Colwell RK, Winkler DW (1984) A null model for null models in biogeography. In: *Ecological Communities: Conceptual Issues and the Evidence* (eds Strong DR Jr, Simberloff D, Abele LG, Thistle AB), pp. 344–359. Princeton University Press, Princeton.

De Bello F, Fischer FM, Puy J, Shipley B, Verdú M, Götzenberger L *et al.* (2025) Raunkiæran shortfalls: challenges and perspectives in trait-based ecology. *Ecological Monographs*, 95, e70018. https://doi.org/10.1002/ecm.70018

Diamond JM (1975) Assembly of species communities. In: *Ecology and Evolution of Communities* (eds Cody ML, Diamond JM), pp. 342–444. Harvard University Press, Cambridge, MA.

649 Gelman A, Hill J (2007) *Data Analysis Using Regression and Multilevel/Hierarchical*
650 *Models*. Cambridge University Press.
651
652 Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W (2008) GEIGER: investigating
653 evolutionary radiations. *Bioinformatics*, 24, 129–131.
654
655 Harmon LJ, Glor RE (2010) Poor statistical performance of the Mantel test in
656 phylogenetic comparative analyses. *Evolution*, 64, 2173–2178.
657
658 Harrell FE (2001) *Regression Modeling Strategies*. Springer, New York.
659
660 Hartig F (2022) *DHARMa: residual diagnostics for hierarchical (multi-level/mixed)*
661 *regression models*. R package version 0.4.6. https://CRAN.R-
662 project.org/package=DHARMa
663
664 Hoff PD (2007) Modeling homophily and stochastic equivalence in symmetric relational
665 data. In: *Advances in Neural Information Processing Systems*, 19. MIT Press.
666
667 Hoff PD (2015) Dyadic data analysis with AMEN. https://arxiv.org/abs/1506.08237
668
669 Hoff P, Fosdick B, Volfovsky A (2024) *amen: additive and multiplicative effects models*
670 *for networks and relational data*. R package version 1.4.5. https://CRAN.R-
671 project.org/package=amen

672

673 Hoff PD, Raftery AE, Handcock MS (2002) Latent space approaches to social network
674 analysis. *Journal of the American Statistical Association*, 97, 1090–1098.

675

676 Hurlbert SH (1984) Pseudoreplication and the design of ecological field experiments.
677 *Ecological Monographs*, 54, 187–211. https://doi.org/10.2307/1942661

678

679 Ives AR, Helmus MR (2011) Generalized linear mixed models for phylogenetic analyses
680 of community structure. *Ecological Monographs*, 81, 511–525.
681 https://doi.org/10.1890/10-1264.1

682

683 Kirk MA, Rahel FJ, Laughlin DC (2022) Environmental filters of freshwater fish
684 community assembly along elevation and latitudinal gradients. *Global Ecology and*
685 *Biogeography*, 31, 470–485. https://doi.org/10.1111/geb.13439

686

687 Kraft NJB, Godoy O, Levine JM (2015) Plant functional traits and the multidimensional
688 nature of species coexistence. *Proceedings of the National Academy of Sciences*, 112,
689 797–802.

690

691 Kunstler G, Falster D, Coomes DA, Hui F, Kooyman RM, Laughlin DC *et al.* (2016)
692 Plant functional traits have globally consistent effects on competition. *Nature*, 529, 204–
693 207.

694

695 Legendre P, Legendre LF (2012) *Numerical Ecology*, 3rd edn. Elsevier, Oxford.
696
697 Levine JI, An R, Kraft NJB, Pacala SW, Levine JM (2025) Why ecologists struggle to
698 predict coexistence from functional traits. *Trends in Ecology & Evolution*, 40, 147–158.
699
700 MacArthur R, Levins R (1967) The limiting similarity, convergence, and divergence of
701 coexisting species. *American Naturalist*, 101, 377–385.
702
703 Mayfield MM, Levine JM (2010) Opposing effects of competitive exclusion on the
704 phylogenetic structure of communities. *Ecology Letters*, 13, 1085–1093.
705
706 McElreath R (2020) *Statistical Rethinking*. CRC Press.
707
708 McGill BJ, Enquist BJ, Weiher E, Westoby M (2006) Rebuilding community ecology
709 from functional traits. *Trends in Ecology & Evolution*, 21, 178–185.
710
711 Mihalitsis M, Bellwood DR (2017) A morphological and functional basis for maximum
712 prey size in piscivorous fishes. *PLoS ONE*, 12, e0184679.
713
714 Minhas S, Hoff PD, Ward MD (2019) Inferential approaches for network analysis: AMEN
715 for latent factor models. *Political Analysis*, 27, 208–222.
716

717 Münkemüller T, de Bello F, Meynard CN, Gravel D, Lavergne S, Mouillot D *et al.* (2012)
718 From diversity indices to community assembly processes: a test with simulated data.
719 *Ecography*, 35, 468–480. https://doi.org/10.1111/j.1600-0587.2011.07259.x
720
721 Navarro-Cano JA, Goberna M, Valiente-Banuet A, Verdú M (2021) Phenotypic structure
722 of plant facilitation networks. *Ecology Letters*, 24, 509–519.
723 https://doi.org/10.1111/ele.13669
724
725 Ovaskainen O, Tikhonov G, Norberg A, Blanchet FG, Duan L, Dunson D *et al.* (2017)
726 How to make more out of community data? *Ecology Letters*, 20, 561–576.
727 https://doi.org/10.1111/ele.12757
728
729 Pebesma E, Bivand R (2023) *Spatial Data Science: With Applications in R*. CRC Press.
730 https://doi.org/10.1201/9780429459016
731
732 Pebesma E (2018) Simple features for R: standardized support for spatial vector data.
733 *R Journal*, 10, 439–446. https://doi.org/10.32614/RJ-2018-009
734
735 Peres-Neto PR, Olden JD, Jackson DA (2001) Environmentally constrained null models.
736 *Oikos*, 93, 110–120. https://doi.org/10.1034/j.1600-0706.2001.930112.x
737
738 Pigot AL, Etienne RS (2015) A dynamic null model for phylogenetic community
739 structure. *Ecology Letters*, 18, 153–163.

Pollock LJ, Tingley R, Morris WK, Golding N, O'Hara RB, Parris KM *et al.* (2014) Understanding co-occurrence by modelling species simultaneously with a JSDM. *Methods in Ecology and Evolution*, 5, 397–406. https://doi.org/10.1111/2041-210X.12180

R Core Team (2024) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/

Revell LJ (2024) phytools 2.0: an updated R ecosystem for phylogenetic comparative methods. *PeerJ*, 12, e16505.

Rohr M, Thuiller W, Chalmandrier L, Münkemüller T (2025) The role of observation scales, trait correlations, and competitive regimes. *Ecology and Evolution*, 15, e71272. https://doi.org/10.1002/ece3.71272

Stan Development Team (2024) RStan: the R interface to Stan. https://mc-stan.org/

Stephens M (2000) Dealing with label switching in mixture models. *Journal of the Royal Statistical Society B*, 62, 795–809.

Talts S, Betancourt M, Simpson D, Vehtari A, Gelman A (2020) Validating Bayesian inference algorithms with simulation-based calibration. *arXiv*.

763 Thompson PL, Guzman LM, De Meester L, Horváth Z, Ptacnik R, Vanschoenwinkel B *et*
764 *al.* (2020) A process-based metacommunity framework. *Ecology Letters*, 23, 1314–
765 1329. https://doi.org/10.1111/ele.13568
766
767 van der Valk AG (1981) Succession in wetlands—a Gleasonian approach. *Ecology*, 62,
768 688–696.
769
770 Veech JA (2013) A probabilistic model for analysing species co-occurrence. *Global*
771 *Ecology and Biogeography*, 22, 252–260. https://doi.org/10.1111/j.1466-
772 8238.2012.00789.x
773
774 Vellend M (2010) Conceptual synthesis in community ecology. *Quarterly Review of*
775 *Biology*, 85, 183–206. https://doi.org/10.1086/652373
776
777 Warner R, Kenny DA, Stoto M (1979) A new round robin ANOVA for social interaction
778 data. *Journal of Personality and Social Psychology*, 37, 1742–1757.
779
780 Webb CO, Ackerly DD, McPeek MA, Donoghue MJ (2002) Phylogenies and community
781 ecology. *Annual Review of Ecology and Systematics*, 33, 475–505.
782