Title: General models of ecological diversification. II. Simulations and empirical applications

Author: Philip M. Novack-Gottshall

Accepted 12/15/2015 for publication in *Paleobiology*

RRH: SIMULATING ECOLOGICAL DIVERSIFICATION MODELS

LRH: PHILIP M. NOVACK-GOTTSHALL

Keywords: Functional trait, Functional diversity; Morphological disparity; Community assembly rule; Monte Carlo simulation; Stochastic simulation; Classification tree; Regressive partitioning; Model selection; Multi-model inference; Ordovician

Abstract.—Models of functional ecospace diversification within life-habit frameworks (functional-trait spaces) are increasingly used across community ecology, functional ecology, and paleoecology. In general, these models can be represented by four basic processes, three that have driven causes and one that occurs through a passive process. The driven models include redundancy (caused by forms of functional canalization), partitioning (specialization), and expansion (divergent novelty), but they also share important dynamical similarities with the passive neutral model. In this second of two companion articles, Monte Carlo simulations of these models are used to illustrate their basic statistical dynamics across a range of data structures and implementations. Ecospace frameworks with greater numbers of characters (functional traits) and ordered (multi-state) character types provide more distinct dynamics and greater ability to distinguish the models, but the general dynamics tend to be congruent across all implementations. Classification tree methods are proposed as a powerful means to select among multiple candidate models when using multivariate data sets. Well-preserved Late Ordovician (type Cincinnatian) samples from the Kope and Waynesville Formations are used to illustrate how these models can be inferred in empirical applications. Initial simulations overestimate the ecological disparity of actual assemblages, confirming that actual life habits are highly constrained. Modifications incorporating more realistic assumptions (such as weighting potential life habits according to actual frequencies and adding a parameter controlling the strength of each model's rules) provide better correspondence to actual assemblages. Samples from both formations are best fit by partitioning (and to lesser extent redundancy) models, consistent with a role for local processes. When aggregated as an entire formation, the Kope Formation pool remains best fit by the partitioning model, whereas the entire Waynesville pool is better fit by the redundancy model, implying greater beta diversity within this unit. The 'ecospace' package is provided to implement the simulations and to calculate their dynamics using the R statistical language.

Philip M. Novack-Gottshall. Department of Biological Sciences, Benedictine University, Lisle, IL 60532 . E-mail: pnovack-gottshall@ben.edu

Introduction

Understanding the causes of ecological diversification remains an important goal in paleontology, ecology, and evolutionary biology. Andrew Bush and I recently introduced (2012) several models of diversification in ecospace (functional trait-space) that are useful for conceptually representing a broad range of diversifications, whether at the scale of ecological assembly of communities or whether shaping entire biotas over evolutionary timescales. The models share many similarities with hypotheses used by community and functional ecologists (Villéger et al. 2010, 2011, Mouillot et al. 2013, Vogt et al. 2013, Gerisch 2014), as well as those used by paleontologists studying morphological, and increasingly ecological (functional), disparity (Dineen et al. 2014, Miller et al. 2014, Mitchell and Makovicky 2014, Dick and Maxwell 2015, Dineen et al. 2015, Knope et al. 2015). Most of these ideas invoke the processes of ecological canalization, specialization, and/or divergence during diversification. In a companion article (Novack-Gottshall 2016), I described four of these models in more detail and discussed their usefulness as general models of ecological diversification. The redundancy model occurs when successive species occupy similar regions in ecospace as previously existing species, and can be considered a form of ecological canalization. The partitioning model occurs when successive species progressively subdivide ecospace, such as occurs in niche partitioning and specialization. The expansion model occurs when successive species progressively explore novel portions of ecospace, such as occurs during niche divergence. Despite their mechanistic differences, these three models are all driven (*sensu* McShea 1994) by particular causes. In contrast, the fourth neutral model occurs as a passive model, in which ecospace is inhabited at random.

The ecospace that becomes structured during each model's implementation can be envisioned as a landscape (or multivariate ordination) defined by the life habits (functional trait combinations) of constituent organisms (Fig. 1). Because each model ecospace is associated with distinct distributions of life habits (ecological structure) that can be quantified using various disparity statistics, the models also suggest promise as the basis for quantitative tests if implemented through computer simulations. In the companion article, I summarized the expected dynamics for these models, using a range of metrics from the morphological disparity and functional diversity literature. Sensitivity analyses (Ciampaglio et al. 2001, Villier and Eble 2009, Mouchet et al. 2010) have demonstrated inherent strengths and weaknesses to each metric, and it remains to be demonstrated how sensitive these dynamics are to different implementations of the models and to different data structures (ecospace frameworks). More important, it remains unclear how to select among alternative models when using multiple, interdependent statistics (Burnham and Anderson 2002, Johnson and Omland 2004, Grueber et al. 2011).

In this second of two companion articles, I demonstrate how these models can be implemented as stochastic simulations. Sensitivity analyses are conducted to evaluate how differences in number of characters (functional traits), character types, and various implementations of the simulations affect resulting statistical dynamics. Because the life habits theoretically allowed by any ecospace framework will always be greater than those that exist in reality (either by logical, biological or other constraints), the simulation results suggest several ways they can be made more realistic, i.e., so that they better approximate actual assemblages. The use of classification trees (Breiman et al. 1993) trained on multivariate Monte Carlo data sets (simulations) is proposed as a novel method for selecting among multiple models. As a case study, these methods are applied to well-preserved fossil assemblages from the type Cincinnatian

(Upper Ordovician of Ohio, Indiana, and Kentucky). Because the samples hierarchically span multiple localities and stratigraphic levels, they offer useful tests of the generality of the models. In addition to its application for fossil samples, the methods proposed here have value to those studying functional diversity in modern communities, where debate continues regarding the best manner to test important hypotheses on ecological structure.

Implementing the Models using Monte Carlo Simulations

Statistical dynamics for the four models can be obtained using Monte Carlo simulations, which are provided here for a range of ecospace frameworks. The simulations were programmed using R (R Development Core Team 2015), and a package—'ecospace: Simulating Community Assembly and Ecological Diversification Using Ecospace Frameworks' (Novack-Gottshall 2015)—is provided to allow others to implement them for their own purposes. In the discussion below, single quotation marks (' ') refer to control parameters or variables called within the package functions.

Theoretical Ecospace Framework

Prior to each simulation, a theoretical ecospace framework is defined (with the function 'create_ecospace'), which specifies the realm of possible life habits that all species can inhabit during the simulation. This step is flexible, allowing any number of characters and character types used by functional ecologists (e.g., Mouchet et al. 2010, Villéger et al. 2011), paleoecologists (e.g., Bambach et al. 2007, Bush et al. 2007, Novack-Gottshall 2007), and those studying morphological disparity (e.g., Foote 1994, Ciampaglio et al. 2001). Examples include factors (e.g., diet can be autotrophic, carnivorous, herbivorous, or microbivorous), ordered factors (factors with a specified order, such as for mobility: habitual > intermittent > facultative > passive > sedentary), ordered numerics (discrete or continuous numeric values, such as body size), and binary/numeric character types. Binary states can be treated individually (present=1/absent=0) or can be treated as multiple states within a single character. For example, the character "reproduction" can be treated as including two states [sexual, asexual] with exclusively sexual habits coded as [0,1], exclusively asexual as [1,0], and hermaphrodites as [1,1]. This coding strategy is preferable over the use of discrete factors (sexual, asexual, hermaphroditic) in allowing flexibility for possible (and biologically frequently occurring) state combinations, while maintaining logical functional distances (in a Euclidean or other metric sense), such that hermaphrodites would be closer to sexual and asexual habits that either exclusive mode is to the other. (This example is trivial as it could easily be coded as an ordered factor, but such binary coding schemes can be extended to additional character states where such factorial options are no longer practical.) See Novack-Gottshall (2007) and Supplementary Appendix 1 for additional discussion and examples of the value of multiple binary states within a single character (functional category). When such multiple binary characters are included, subsequent simulations (and the ecospace framework that controls them) specify that "all-absences" are impossible (i.e., an organism cannot be neither sexual nor asexual, [0,0]). This behavior can be controlled with a 'constraint' parameter, which can allow any combination (e.g., [1,1,1]) or specify a maximum number of "multi-presences" (e.g., setting 'constraint=2' allows [1,0,0], [0,1,0], [0,0,1], [1,1,0], [1,0,1], and [0,1,1] as state combinations, but excludes [1,1,1]; setting 'constraint=1' only allows the first three of these combinations).

Another feature when defining the ecospace framework is the ability to weight character states, such that certain states occur more often than others. This constraint can be done using customized weights, or calculated automatically using empirical samples (such as a regional species pool) to assign weights based on their frequency of occurrence. A weight of zero excludes that character state from inclusion in the ecospace framework, and thus from simulations based on it. Although not explored here, these weights could be used, with some care, in linking covariation among states across different characters, and offer potential to extend these simulations to studies of morphological disparity, where non-applicable character states—e.g., glabella shape for a mollusk—are frequently encountered.

Simulations and Implementation as Algorithms

All simulations are implemented as Monte Carlo processes in which species are added iteratively to assemblages, with all added species having their character states specified by the model rules below (Fig. 1). Aside from these model rules (and whatever inherent constraints are imposed by the ecospace framework), there are no deterministic components to the stochastic simulations. Simulations begin with the seeding of a specified number of species, chosen at random (with replacement) from either the species pool (if provided) or following the neutral-rule algorithm (if not). Once seeded, the simulations proceed iteratively (character-by-character, species-by-species) by following the appropriate algorithm until terminated at a pre-specified species-richness value (sample size). A 'strength' parameter can be used to specify the probability that a simulation follows each rule, with possible values ranging from 1.0, in which the model rule is always followed, to 0.0, in which the model rule is never followed (and during which the simulation switches to the default neutral model rule.) Although not explored here, it might be profitable in future analyses to modify the model-selection analyses into an optimization routine so that the 'strength' parameter is treated as a free parameter to be estimated by the data, rather than chosen as an arbitrary value.

Neutral rule.—Iteratively add species independently, assigning characters character-by-character as random multinomial draws (with probabilities assigned by character-state weights, if assigned) from the theoretical ecospace framework. If weighting was assigned to the ecospace framework, character states are assigned in proportion to those weights, on average.

This "rule" is not technically a rule, but the absence of other rules: all further simulations default to this neutral algorithm when, for example, a simulation is implemented with a strength parameter of zero, and converge on this model when implemented at strengths less than one. It is important to note that the neutral rule is not a simple permutation (randomization) test (Kowalewski and Novack-Gottshall 2010), drawing at random from the species pool (except for the seed species, if a species pool is provided). Most tests in the functional ecology literature use such simple permutation tests, and such a test can be enacted in the neutral model by setting the number of seed species to a sufficiently large value, such as the size of the species pool. The life habit of each new species is built character-by-character from the realm of theoretically possible states allowed by the ecospace framework. Thus, new species can occupy combinations of character states that did not occur in the species pool (if provided). This is an important feature of the simulations, allowing the entire theoretical ecospace to be explored by the neutral model if given sufficient numbers of species. If it turns out that the actual species pool is structurally different from a similarly parameterized neutral model, it can be concluded that there is an important mechanism controlling how the actual sample was composed.

Redundancy rule.—Pick one existing species at random and create a new species using that species' characters as a template. A character is modified (using a random multinomial draw from the theoretical ecospace, as in the neutral rule) according to the strength parameter. When the strength parameter is set to unity, all species will be functionally identical to the original seed species; when set to zero, the simulation is identical to the neutral rule. Because new character states can be any allowed by the ecospace framework, there is the possibility of obtaining redundancy greater than that specified by strength parameters less than 1.0 (if, for example, the new randomly chosen character states are identical to those of the template species).

Partitioning rule.—Calculate distances between all pairs of species, using Euclidean distance if all characters are numeric/binary or ordered numeric types or using Gower distance if any character type is an ordered or unordered factor. There are two algorithms to choose from for the next step, a 'strict' (and default) partitioning implementation and a 'relaxed' implementation. In the strict (or "minimum distant-neighbor") implementation, identify the maximum distances between all pairs of species (the "most-distant neighbors"); the new species will have character states based on the pair of species that is the minimum of these distances. This implementation progressively fills in the largest parts of the ecospace that are least occupied between neighboring species, and on average partitions the ecospace in straight-line gradients between seed species. In the relaxed (or "maximum nearest-neighbor") implementation, identify nearest-neighbor distances between all pairs of species; the new species will have character states based on the pair that is the maximum of these distances. This implementation places new species in the most unoccupied portion of the ecospace that is within the cluster of pre-existing species, often the centroid. (See Fig. 1 for visual portrayal of these alternate implementations.) In both cases, the life habit for each new species is built as a resampled combination of the character states of the chosen neighbor pair, with probabilities controlled by the strength parameter. Ordered, multistate character partitioning (whether factor or numeric) can include any state equal to or between the observed states of existing species. Each newly assigned character is compared with the ecospace framework to confirm that it is an allowed state combination before proceeding to the next character; if the newly built character is disallowed from the ecospace framework (i.e., because it has "dual absences" [0,0], has been excluded based on the species pool, or is not allowed by the ecospace 'constraint' parameter), then the character-selection algorithm is repeated until an allowable character is selected. When simulations proceed to very large sample sizes (>100), this confirmatory process can require long computational times, and produce "new" intermediate species that are functionally identical to pre-existing species. This can occur, for example, when no life habits, or perhaps only one, exist that forms an allowable intermediate between the selected neighbors. This behavior mimics the "pathologically" tight packing of species that has been demonstrated in large sample-size simulations of niche partitioning (Kinzig et al. 1999). This behavior also causes dynamics at such sample sizes to share important similarities with weakened implementations of the redundancy model (see additional discussion below).

Expansion rule.—Calculate distances between all pairs of species, as done in the partitioning algorithm, and identify the species pair that is maximally distant. The newly added species has character states chosen at random that are equal to or more extreme (if ordered or numeric character types, or that are the same or different for unordered factors) than those in this species pair, with the strength parameter controlling the probability of following this rule. As with the partitioning rule, each newly assigned character is confirmed to be an allowed character-state combination before proceeding with remaining characters and species. Like the partitioning rule, the algorithm can take long computational times to run to completion for large sample sizes, and shares the similar property that functionally identical life habits may occur by virtue of saturation of ecospace-allowable life habits.

Other simulation assumptions and limitations.—The simulations explicitly assume that dispersal is guaranteed to all species, provided that new species have appropriate character states as proscribed by the ecospace framework and the model rules. This is an important distinction from Hubbell's neutral model (2001) and many other spatially structured models in community ecology, but consistent with how many species pools are defined in ecological studies (Cornell and Harrison 2014). Extinction, speciation, nor emigration is allowed during the course of a simulation (although they can play important roles in the definition of the original species pool). All species that immigrate to the assemblage remain there, with their character states retained, throughout the course of the simulation. Such ecological and evolutionary processes (character displacement, competitive exclusion, habitat filtering, etc.) are only present as implemented by the model rules governing addition of new species to the assemblage.

The framework and simulation algorithms currently do not incorporate relative abundance (cf., Bush et al. 2007, Deline 2009, Deline et al. 2012), life-history characteristics, population demographics, dispersal and local extinction, spatial structure, speciation and evolution, or other constraints on the regional pool. The incorporation of relative abundance could be especially worthwhile for ecological studies, as it would allow the simulations to better mimic the assembly of ecological communities. For example, abundance plays an important role in immigration from regional species pools (Hubbell 2001, Patzkowsky and Holland 2007, Cornell and Harrison 2014), and abundant taxa can have larger influences on the composition of functional traits within communities and they can have a greater impact on the functional identity of taxa that settle later (Fargione et al. 2003, Guillemot et al. 2011). Functional diversity studies increasingly incorporate relative abundance ((Villéger et al. 2008, Laliberté and Legendre 2010, Fontana et al. 2015), but the models used here are limited to presence/absence data, which some functional ecology studies (e.g., Ackerly and Cornwell 2007) have demonstrated produce similar results as those incorporating abundance data.

The simulations also currently do not incorporate phylogenetic structure in an explicit sense. The three driven models are Markovian in that pre-existing taxa affect which functional traits can enter at later stages of the simulation, but there is no explicit assumption of heritability or phylogenetic relatedness among simulated species. In contrast, the neutral model is non-Markovian in that all taxa are added independently of one another, without regard to their functional traits. The lack of such phylogenetic structure perhaps hinders the application of these models to diversification within individual, evolving lineages, but the models are intended as reasonable approximations of large-scale evolutionary diversifications involving many disparate lineages, and where the focus is on the ecospace structure of the overall biota. The existing simulations could be modified to include such features, and would prove worthwhile for future studies.

Calculating Functional Diversity and Ecological Disparity Statistics

As each simulation proceeds, the theoretical ecospace (functional-trait space) becomes progressively filled with new species having combinations of character states (life habits) selected by the model rules (Fig. 1). Because the mechanisms controlling community assembly vary among models, it is expected that the statistical dynamics will likewise vary. These model-specific dynamics can be calculated as a function of species richness. Here, eight widely used statistics are calculated, drawn from the morphological disparity and functional-diversity literature (Ciampaglio et al. 2001, Wills 2001, Villéger et al. 2008, Mouchet et al. 2010).

The four morphological (here ecological) disparity measures include the following (Foote 1993, Ciampaglio et al. 2001, Wills 2001). Unique trait-combination (life-habit) richness (H) measures the number of unique life habits within each sample. Mean distance (D) and maximum range (M) measure the mean and maximal distance, respectively, in functional-trait space among species, using Euclidean distance if all characters are numeric/binary or ordered numeric types, or using Gower distance (1971) if any character type is an ordered or unordered factor. Total variance (V) measures the sum of variances for each character state across species (Van Valen 1974); when using factor character types, this statistic cannot be measured and is excluded.

The four functional diversity measures include the following (Mason et al. 2005, Anderson et al. 2006, Villéger et al. 2008, Laliberté and Legendre 2010, Mouchet et al. 2010, Mouillot et al. 2013). Functional richness (FRic) measures the minimal convex-hull volume in multidimensional principal coordinates analysis (PCoA) trait-space ordination. Functional evenness (FEve) measures the evenness of minimum-spanning-tree lengths between species in PCoA trait-space. Functional divergence (FDiv) measures the mean distance of species from the PCoA trait-space centroid. Functional dispersion (FDis) measures the total deviance of taxa from the circle with radius equal to mean distance from PCoA trait-space centroid. FRic and FDiv use a subset (here set to 3 by default) of the PCoA axes for their calculation (because their convex-hull-volume calculation requires more species than functional traits), whereas the other two use the entire PCoA space.

These statistics are calculated with the function 'calc_metrics', wrapping around function 'dbFD' in the 'FD' package (Laliberté and Shipley 2014) to calculate functional-diversity statistics. The four functional diversity statistics are not calculated for samples having less than four unique life habits, for the same reason as just explained. The calculation of FRic can be computationally demanding, and has proven truculent to many users when dealing with idiosyncratic data structures (see help file in Laliberté and Shipley 2014). Some of these issues are resolved in the implementation here. Specifically, the package defaults to three PCoA axes ('m=3') as a generally useful and computationally tractable number of axes, defaults to Lingoes correction (Legendre and Anderson 1999) when non-Eucliden PCoA axes are calculated, has corrected (starting with 'FD' v. 1.0–12) a previously unnoticed rounding error in the calculation of PCoA scores which caused computation problems with highly redundant datasets, and has been optimized to prevent overwriting errors (of temporarily stored vertices files) when used in a parallel-processing computer environment. Although including more PCoA axes allows greater statistical power (cf., Villéger et al. 2011, Maire et al. 2015), the use of three here provides satisfactory resolution for the often functionally redundant data sets in these simulations. This choice is also warranted because such statistics are calculated directly from multivariate PCoA trait-spaces (instead of dendograms) and because the ecospace frameworks used here have few factor character types (which are especially sensitive to few axes). Three PCoA axes also provide the largest computationally tractable number to allow calculation of these statistics across all sample sizes and models considered.

Sensitivity analyses of these statistics (Ciampaglio et al. 2001, Mouchet et al. 2010, Novack-Gottshall 2010, Maire et al. 2015) have concluded that, despite much correlation among them, each statistic has its strengths for particular uses and contexts. H, M, V, FRic, and FDis are all useful measures of disparity (or dispersion of species within functional-trait space). D and FDiv are useful for characterizing internal structure (i.e., clumping or inhomogeneities within the trait-space). FEve is useful for characterizing the extent of spacing among species within the trait-space. Such sensitivity studies consistently recommend that all are useful for particular purposes, with selection left to the researcher, a solution that is decidedly indecisive. Below, I demonstrate that classification trees (trained on all statistics from simulated datasets) offer a productive solution to this subjectivity.

Model Selection

A difficulty in implementing model selection with these simulations is that every simulation iteration (i.e., each addition of a species) yields eight statistics. This lack of statistical independence is typically addressed through model-fitting procedures (such as likelihood-ratio tests or the Akaike information criterion [AIC], Anderson et al. 2000, Burnham and Anderson 2002, Johnson and Omland 2004, Grueber et al. 2011) that incorporate dependency relationships explicitly (such as for a multivariate normal model). However, these methods become prohibitively complicated when the statistics involved are not approximately normally distributed. In this case, H is a discrete binomial distribution. Depending on the model, several statistics are highly skewed, with D best fit by gamma distributions and M and V best fit by Weibull distributions (confirmed using maximum likelihood goodness-of-fit tests); in contrast, the four functional diversity statistics are all reasonably well fit by normal distributions. Most functional ecology studies thus use one (or a few) of these statistics for a particular test, subjecting them to a simple permutation test to reject a particular null hypothesis. Because all functional diversity /disparity metrics contribute potentially useful information on functional (ecospace) structure (Ciampaglio et al. 2001, Villier and Eble 2009, Mouchet et al. 2010), there is value in retaining all suitable metrics when possible. The obstacle is doing so while simultaneously considering multiple candidate models (Johnson and Omland 2004, Grueber et al. 2011).

Here I propose a novel and simple method to conduct model-fitting in such cases. Classification trees are a powerful and flexible tool for classifying complex datasets with non-linear, localized, and other complicated relationships among variables (Breiman et al. 1984, Cutler et al. 2007), and they are being increasingly used by ecologists (De'ath and Fabricius 2000, De'ath 2002, Sullivan et al. 2006, Cutler et al. 2007, Boyer 2010, Durst and Roth 2012) and paleontologists (Finnegan et al. 2012, 2015). The basic algorithm is to recursively subdivide heterogeneous data sets (e.g., composed of different classes/models) into subsets that are as homogeneous as possible, using whatever values of variable(s) are most powerful to do so. The result, often portrayed visually as a decision tree, is a statistical model in which values of predictor variables are used to classify the original class/model identities. Studies demonstrate that the method (and its continuous-variable counterpart regression trees) performs as well as non-linear (GAM and GLM) regressions and constrained ordination, and is exceptionally well suited to scenarios involving complex (heterogeneous) classifications (Breiman et al. 1993, Prasad et al. 2006, Cutler et al. 2007). Performance can be improved using "random forest" variants, in which replicate trees are made, each time leaving out random subsets of variables and data, until a stable model solution is achieved (Prasad et al. 2006, Cutler et al. 2007). This variant also prevents overfitting of the model, and allows ranking of predictor variables by their overall importance. (Use of alternate tree-building algorithms below confirmed that random forests performed best for the current analyses.) When applied to new data sets (either a validation/test set to test the tree's classification performance or empirical samples), each resulting tree in the forest casts a vote for which class (here, model) each sample should be classified. The proportion of votes for particular models, which scale from zero to one, can be used as a measure of support for particular candidate models, analogous to how Akaike weighting allows candidate models to be compared by their relative support (Anderson et al. 2000, Johnson and Omland 2004, Grueber et al. 2011).

The typical work flow to implement classification trees as a form of model-fitting includes the following steps. See Supplementary Appendix 2 for technical details on how these steps were implemented in analyses below.

1. Simulate training data sets: Run many simulations of the candidate models (here, the four ecological diversification models) to be considered, calculating relevant summary statistics (here, the eight disparity and functional diversity statistics) for each iteration. This large Monte-Carlo data set serves as a training data set to train the classification tree.
2. Build classification tree: Use random-forest classification to identify which combination of variables and their statistics are most predictive of each original model: Model ~ S + H + D + M + V + FRic + FEve + FDiv + FDis
3. Test tree with validation data set: Test the tree's performance using validation data sets, additional simulation data sets of known model that were not included when training the tree. A general (non-overfit) classification tree should perform approximately as well on a validation data set as it did on its training data set.
4. Classify empirical samples using the random-forest tree: Submit empirical samples (with calculated disparity /diversity statistics but unknown model identity) to the tree for classification, treating proportion of votes cast for each model as a relative measure of model support.

Methodological analyses (Supplementary Fig. 9) demonstrate that classification trees perform well as model-fitting methods. When conducted across a variety of ecospace frameworks (spanning different numbers of characters, seed species, and character types, and incorporating weighting by empirical species pools), trees were able to classify 75–97% of their training data sets correctly (Supplementary Appendix 2); validation tests showed similar performance, with ~2% fewer samples being classified correctly. When applied to more complex cases, such as the ability to distinguish nine models of subtly distinct dynamics (neutral model, four [including both strict and relaxed partitioning versions] process-driven models with 100% rule-following, and the four process-driven models with 50% rule-following), the training trees still classified 90% of training data correctly (84% of validation data), even when the dynamics of the models were quite similar. This performance extends across all sample sizes, with the classification tree still sufficiently powerful to distinguish these models >33% of the time correctly at sample sizes as small as 6 (the minimum sample size when the models could be classified, as 5 species were seeded without any assembly rules). In nearly all cases, the correct model also received the largest proportion of classification votes.

The trees were also successful in identifying samples produced from simulations quite foreign from those used to train the sample (Supplementary Fig. 9). When trained on the nine 50% and 100% rule-following models, the resulting tree generally classified validation data sets produced at 95% and 90% rule-following as members of the 100% models, with some votes for the alternate 50% rules. Validation samples produced at 75% generally were classified with mixed support for both the 100% and 50% models. (This generality declines at sample sizes greater than 30–40 as dynamics reach stable asymptotic values; at larger sample sizes, it is worthwhile to consult more complex classification trees to identify the strength of these models.)

These results demonstrate that classification tree votes can provide a measure of model-selection support, and this support can be extended, at least in this case, beyond the particular training models. Whether trained on 9 training sets (as in Supplementary Fig. 9), or on larger numbers (13- or 21-model training sets, if adding 75%- or 75%-, 90%-, and 95%-strength models), the correct model consistently achieves support levels greater than 0.20, from which it is possible to recommend this support value as a benchmark of "substantial" support for the correct model, and values above 0.40 tend to be associated with "unambiguously strong" support for the correct model. (In other words, support values of 0.2 and 0.4 here can serve analogous roles as the 0.1 and 0.9 Akaike-weight benchmarks used in AIC-based model selection.) Although 0.4 may seem low compared to 0.9, the difference can be explained by the large number of candidate models [9, 13, or 21] being considered here, and the fact that the correct model in these tests often achieves support values much greater than 0.4 whereas poorly supported models tend to receive far fewer votes. It is recommended that future analyses using different model implementations run sensitivity analyses to validate the generality of these benchmarks, as they likely depend on the number and dynamical similarity of models being considered.

The classification-tree approach to model selection shares several similarities with the methods recently termed approximate Bayesian computation (ABC) in that model selection can proceed without a known likelihood function or an understanding of the dependence between variables (Beaumont 2010, Slater et al. 2012). However, ABC is primarily used when estimating one or a few parameters, and often within structurally similar or hierarchically nested models; they are not well suited for choosing among models with distinct parameters (such as is the case here). ABC also performs poorly when considering many summary statistics simultaneously, the "curse of dimensionality" (Beaumont 2010) (and also the case here, where eight interdependent statistics are used). In contrast, classification trees are powerful and flexible methods that were developed to classify such heterogeneous data sets using many predictive variables (Breiman et al. 1993). Although they have not been used before for multivariate model selection, they perform remarkably well and appear to be well suited to this goal.

Simulation Dynamics and Sensitivity to Different Ecospace Frameworks

The four models can be distinguished by examining the dynamics of the functional diversity/disparity statistics as a function of species richness. These dynamics are examined here across a range of ecospace-framework simulations in order to understand their general dynamics, their sensitivity to different ways of creating ecospace frameworks, and their statistical power in model inference when models are known to operate in simulations.

Popular ecospace frameworks.—Figure 2 illustrates the dynamics for two influential ecospace frameworks, that of Bush and Bambach (Bambach et al. 2007, Bush et al. 2007, Bush and Bambach 2011, Bush et al. 2011) and Novack-Gottshall (2007), which have been adapted for a variety of studies (e.g., Xiao and Laflamme 2009, Bush and Novack-Gottshall 2012, Ros et al. 2012, Bush and Pruss 2013, Laflamme et al. 2013, Dineen et al. 2014, Aberhan and Kiessling 2015, Dick and Maxwell 2015, Dineen et al. 2015, Mondal and Harries 2015, O'Brien and Caron 2015). Bambach's original ecospace framework (1983, 1985) consisted of three factor characters (diet, activity/motility, and tiering, with the last ordered) and 11 character states, allowing 48 possible unique life habits. The Bush and Bambach ecospace framework retained these three characters, but added additional character states (totaling 19) and reformulated motility as an ordered factor. With the addition of osmotrophy as a diet category (Laflamme et al. 2009, Bush and Bambach 2011, Bush et al. 2011), this yields 252 possible unique life habits. The Novack-Gottshall framework more finely subdivides life-habit dimensions according to substrate relationships, microhabitat, motility, diet, and foraging habits, and is modified here to include 18 characters (6 ordered numeric and 12 multi-state binary) and 37 character states, yielding nearly 3.6 trillion possible life habits. (See Supplementary Appendix 1 for list of characters and states.) In all three frameworks, many of these theoretically possible life habits are unlikely to ever be realized given logical or functional constraints (cf., Bambach et al. 2007).

Model dynamics for both ecospace frameworks are congruent, despite their rather different structures. In both cases, life-habit richness (H) increases as a function of species richness (S) for all models except redundancy, with the values for strict partitioning (and for Bush and Bambach, also relaxed partitioning) reduced compared to the other models. The neutral and expansion model slope for Bush and Bambach is also reduced slightly compared to that of Novack-Gottshall, as the smaller realized ecospace becomes saturated more quickly. Trends for mean distance (D) are also similar, with neutral remaining stable, expansion increasing slightly (more so for Novack-Gottshall), and the others declining asymptotically to varying extents. Total variance (V) is highly correlated with D for Novack-Gottshall, although the presence of factor character types in Bush and Bambach precludes its measurement. Maximum range (M) is quite distinct across frameworks, with all Bush and Bambach simulations quickly filling the available ecospace. The Novack-Gottshall framework, in contrast, does not reach asymptotic limits within 50 species, with the rank order of model dynamics generally paralleling that for the other disparity metrics (D and V): expansion has the greatest range values, followed closely by neutral, significantly greater than both partitioning models, and with redundancy remaining constant at low range. Functional richness (FRic) increases for all models, except redundancy for Novack-

Gottshall which decreases, with the greatest values for expansion and neutral models, intermediate for strict partitioning, and least for redundancy. For Bush and Bambach, relaxed partitioning produces the largest FRic values of all models, whereas it remains intermediate for Novack-Gottshall. Functional evenness (FEve) remains nearly constant for neutral and expansion and declines asymptotically for the other models, and the rank order is similar in both cases: expansion and neutral at top, followed by both partitioning models, and redundancy at bottom. Functional divergence (FDiv) has generally non-linear dynamics. Its dynamics increase in both frameworks for redundancy whereas they decrease for other models, with relaxed partitioning generally decreasing at the fastest rate (although eventually crossing neutral and expansion models into an intermediate value at high sample sizes for Novack-Gottshall). Functional dispersion (FDis) has similar dynamics to D (and V), although the dynamics for the expansion and neutral models increase asymptotically instead of remaining constant.

That these two quite different frameworks share generally similar dynamics is good news for those seeking to apply these statistics for model inference. The most important differences include nearly continuous overlap between the neutral and expansion models in the Bush and Bambach framework, whereas the expansion model generally has greater values and more distinct dynamics for Novack-Gottshall. This is an important consideration if the goal (cf., Bush et al. 2007) is to distinguish an active process (expansion) from a passive process (neutral). Trend dynamics generally also have smaller error margins for the Novack-Gottshall framework, which enables more powerful model selection.

This statistical power can be evaluated using classification tree methods, by training a random forest tree on the simulation statistics and counting the proportion of an independently simulated (validation) data set that is classified correctly. The Bush and Bambach framework was able to correctly identify 73% of validation samples correctly. (See Supplementary Appendix 2 for details and classification rates.) Greatest variable importance (in order from most to moderate importance) was awarded to H, FEve, FDis, D, and FRic. This ranking is consistent with Figure 2A, where these dynamics generally overlap least with those of other models. Confusion matrices, which illustrate both the frequency that the correct model was selected as well as which models they were misclassified, reiterate that the tree had difficulty distinguishing the expansion from neutral dynamics in the Bush and Bambach framework, with ~50% of expansion samples misclassified as neutral (and vice versa). In contrast, the two partitioning models were correctly classified more than 80% of the time (and usually confused with the alternate partitioning parametrization), and the redundancy model was nearly always properly classified.

The Novack-Gottshall framework provided greater power to distinguish the models, with 89% classified correctly using a validation data set. Variable importance rankings rated FEve the most important, followed in order by D, H, FDiv, and FDis, four of which were also considered most important in the Bush and Bambach tree. The confusion matrix also demonstrates some difficulty distinguishing the expansion and neutral models, but to a much lesser extent: ~75% of expansion models were correctly distinguished from the neutral models (and vice versa), with 98% of partitioning and exactly 100% of redundancy models correctly classified. Thus, the more multidimensional Novack-Gottshall framework, albeit more complicated and less intuitively appealing for general summary applications, provides greater statistical power for inferring important ecological and evolutionary models.

These results can be generalized by examining other ecospace frameworks, offering recommendations for what forms of ecospace framework choices are most likely to provide informative results for future analyses. The most important differences between these two ecospace frameworks concern differences in number of characters and character types, thus it is worth examining the effect of these variables. In addition, analyses also examine the effect that number of seed species has on dynamics and the ability to distinguish resulting models.

Number of ecospace characters.—Ecospace frameworks with more characters and states ought to have greater opportunities for the models to be distinguished, and thus be more powerful frameworks. Simulations— implemented using ecospace frameworks with 5, 15, and 25 characters of mixed character types—bear this out, but only to a modest extent (Supplementary Fig. 6). The dynamics across simulations are generally similar, in much the same way as for the frameworks discussed above (Fig. 2). In particular, the smaller 5-character simulation (Supplementary Fig. 6A) strongly resembles the dynamics of the Bush and Bambach framework (Fig. 2A) whereas the larger 15-character simulation strongly resembles the dynamics of the Novack-Gottshall framework (Fig. 2B), attesting that many of the differences result from differences in the numbers of characters in each framework. However, the classification rates improve only modestly as additional characters are added (83%, 85%, and 86% of models, respectively), suggesting that number of characters is not the primary driver of performance between the two frameworks. This conclusion reiterates one made in prior sensitivity analyses of functional diversity (Maire et al. 2015), which likewise found that number of characters accounted for limited improvement in performance. Examination of confusion matrices (Supplementary Appendix 2) demonstrates that most of the improvement in the larger simulations was the result of improved performance at distinguishing the partitioning models, both from each other and from the other models, made possible because the larger number of characters allows for greater opportunities for intermediate life habits.

Character types.—Another important difference between the two frameworks considered above is their character types, with the Bush and Bambach framework using ordered and unordered factors and the Novack-Gottshall framework using ordered and unordered numeric (binary) characters. Maire et al. (2015), using sensitivity analyses involving simulated data sets, concluded that character type carried the second-most importance in the performance of functional-trait spaces (ecospace frameworks), second only to whether statistics were calculated via a multidimensional ordination versus a functional dendogram, a comparison not considered in the current study (where all functional-diversity metrics are calculated using the more powerful ordination method). They demonstrated that continuous and ordinal (ordered numeric) character types performed best, and discrete categorical (unordered factor) types performed worst. This conclusion is partially borne out here (Supplementary Fig. 7) using simulated ecospace frameworks built from 15 characters of varying types: factors, ordered factors, ordered numeric, and binaries (unordered numeric). Again, dynamics are generally similar across simulations, with the greatest difference occurring with ordered factors (Supplementary Fig. 7B), which has rather distinct dynamics from the other frameworks for D, M, FDiv, and FDis, and to a lesser extent for FRic. The dynamics for factors and ordered and unordered numeric (binaries) are all quite similar. Thus, there is negligible difference caused by distance metric used, with Gower distance used for both factor types and Euclidean distance for both numeric types. Classification rates on ordered factors were also substantially improved (94% of trained models classified correctly) compared to the other character types (78% for unordered factors, 79% for ordered numerics, and 81% for binaries), with most of the difference due to improved ability of the ordered-factor tree to distinguish the neutral and expansion models. (See Supplementary Appendix 2 for additional details, including confusion matrices.) Overall, ordered character types, and especially ordered factors, perform better than the other character types, although the improvement is modest.

Number of seed species in simulations.—Actual ecological communities do not normally start with pre-specified numbers of initial colonizing "seed" species; thus, selecting how many to choose in a simulation should reflect knowledge of the ecological and evolutionary scenarios one seeks to mimic. It is worthwhile to evaluate the effect that this arbitrary choice might have on resulting simulations. Supplementary Figure 8 demonstrates that the effect can be substantial for some models, although the dynamics for the neutral and expansion models are essentially invariant across implementations tested here. The greatest difference occurs for FDiv, with the partitioning dynamics dramatically changing behavior across the three scenarios: when seeded with three species, the dynamics are relatively constant at low values; when seeded with five species, the strict version decreases and the relaxed version remains constant after an initial decrease; and when seeded with ten species, both trends decrease at different rates. In all cases, however, the strict version has values greater than the relaxed one. The statistic M and to a lesser extent FRic also vary with number of seed species, with the redundancy and partitioning values substantially suppressed when seeded with three (and to a lesser extent five) species and substantially increased when seeded with ten species. Although some dynamics vary dramatically due to differences in seed species, the classification rates are less variable. The simulation seeded with three species classified 85% of training data correctly (or 82% when excluding the redundancy model, in which many statistics were absent because the functional-diversity statistics require a minimum of five unique life habits), 85% for the simulation seeded with five species, and 75% for the simulation seeded with ten species (see Supplementary Appendix 2 for details, including confusion matrices). Confusion matrices demonstrate that the neutral and expansion models are most difficult to distinguish in all three simulations, with the ten-seed-species simulation also showing greater difficulty distinguishing the other models. Thus, it is worthwhile—assuming the ecological and evolutionary context justifies it—using moderately few numbers of species to seed the simulations because using larger numbers impedes the simulations from following the model rules. Although not entirely justified by the overall classification tree results, the dynamics suggest that using too few species to seed the simulations can suppress the dynamics of the partitioning and redundancy models, which are explicitly constrained to exploring the part of the ecospace seeded by the initially occurring species.

Which statistics are most valuable?—Another benefit of using random-forest classification trees as a form of model selection is that they provide an explicit mechanism to evaluate which statistics are most valuable in classifying the models. As discussed earlier, this is accomplished by excluding some variables at random from the tree-training algorithm (somewhat analogous to bootstrap support in cladistics), which prevents overfitting of the model and aids evaluating the relative variable contribution to the final tree model (Prasad et al. 2006, Cutler et al. 2007, Strobl et al. 2007). Variable importance rankings for the classification trees used with these simulations considered FEve the most valuable statistic, on average, for distinguishing the models, followed by D, H, and FDis. In models where V could be included, it was considered the fourth most important variable. FDiv and FRic consistently had intermediate (but worthwhile) value, and M and S consistently had low value. The poor ranking of S is reassuring, as it is the independent variable in all models, and only useful in combination with other statistics. This ranking is intuitive, as the high-value statistics tended to have the least overlap among models in the simulations presented above (Fig. 2, Supplementary Figs. 6–8).

Although functional ecologists focus almost exclusively on their newly invented statistics, those traditionally used in morphological disparity are often just as powerful (or more so), and also computationally more straightforward to measure. Thus, they ought to be considered more frequently in functional diversity studies. However, a benefit of using classification trees as a basis for model selection is that one ultimately does not have to choose which statistics to include (cf., Mouchet et al. 2010). The tree algorithm is sufficiently powerful and flexible to "learn" which variables to rely on in each circumstance: it always sees the forest through their trees.

Model Inference of Late Ordovician Samples and Effect of Geographic and Temporal Scale

Empirical samples.—These methods are applied to a case study of well-preserved fossil biotas from the type Cincinnatian (Late Ordovician) Kope and Waynesville Formations of southwest Ohio, northern Kentucky, and southeast Indiana. Lithologically, these units are composed of shales interbedded with thin limestone tempestites, representing offshore, soft-substrate conditions below storm-wave base in the Cincinnati Arch epeiric sea (Holland 1993). Taphonomic conditions are excellent, with preservation by obrution deposits, and most assemblages include evidence of fossils preserved in life position, articulated specimens, and autochthonous and parautochthonous communities (Frey 1987a, b, Schumacher and Shrake 1997, Hughes and Cooper 1999, Aucoin et al. 2015). The shelly and trace-fossil biotas are also extensively well studied, with much attention given to their life habits (Pojeta 1971, Richards 1972, Frey 1987a, b, 1989, Brandt et al. 1995, Lescinsky 1995, Brandt 1996, Feldmann 1996, Sandy 1996, Schumacher and Shrake 1997, Leighton 1998, Gaines et al. 1999, Meyer et al.

2002, Morris and Felton 2003, Novack-Gottshall and Miller 2003, English and Babcock 2007, Freeman et al. 2013). These two formations were selected for analysis as a case study because they represent exemplary preservation within a single depositional environment, but the results are likely to apply to other type Cincinnatian units.

218 collections from the Kope and Waynesville Formations (totaling 2322 occurrences; min=2, mean=8.2, median=8, max=23 taxa) were downloaded from the Paleobiology Database (www.paleobiodb.org), only including those that included the entire biota. Source references include Dalvé (1948), Browne (1964), Frey (1987a, 1988, 1989), Novack-Gottshall and Miller (2003), and Holland and Patzkowsky (2007). The size of each collection was confirmed to represent a typical field sample (i.e., a hand sample, slab, or small bedding plane). All stratigraphic members were included, as they are defined more by biofacies than by major differences in depositional environment (Holland et al. 2001, Holland and Patzkowsky 2007).

92 Kope samples were collected from eight stratigraphic sections (roughly equivalent to outcrops) spanning the Fulton, Economy, Southgate, and McMicken Members; the 126 Waynesville samples represented fifteen sections spanning the Fort Ancient, Clarksville, and Blanchester Members and their informal "trilobite shale" and "*Treptoceras duseri* shale" facies. Analyses were conducted at multiple scales: individual samples (representing autochthonous local communities), stratigraphic sections, members, and each formation in aggregate. Because the analyses hierarchically span multiple localities and stratigraphic levels, they offer useful tests of the generality of the models. If the ecological models presented above provide good fits for the functional structure within individual samples, then we might expect different structures (and models) to be recorded at the spatially and temporally mixed scales of the larger aggregate levels where regional and evolutionary processes are more likely to play out.

Ecospace framework and ecological disparity.—Life habits (functional traits) of 237 unique Kope and Waynesville taxa were coded into a modified ecospace framework from Novack-Gottshall (2007) that included 18 characters (6 ordered numeric and 12 multi-state binary) and 37 character states. Given the sensitivity analyses above, the large number of characters and the presence of these character types ought to be sufficient for distinguishing the relevant models. Life habits were coded according to inferences of functional morphology, body size, ichnology, *in situ* preservation, biotic associations recording direct interactions, and interpretation of geographic and depositional environment patterns, in consultation with 67 peer-reviewed, published references, many studying this fauna directly. Body-size measurements were made using the methods of Novack-Gottshall (2008b), primarily from Feldmann (1996), supplemented with the *United States Geological Survey Professional Paper* volume 1066 series and the *Treatise on Invertebrate Paleontology*. Ordered numeric states (e.g., body volume, mobility, and distance from seafloor) were rescaled as five discrete, equidistant bins (seven for body volume) so that they ranged from zero to unity. See Supplementary Appendix 1 for list of characters and states, and an example of how life-habit (functional-trait) inferences were coded, using the Ohio state fossil, the trilobite *Isotelus maximus*.

Of the 237 taxa in the species pool (Supplementary Appendix 3), 101 were identified to (and coded at) species-level and 136 were coded at genus-level, with just 7 of these taxa having uncoded (unreliably inferred) states. The relative completeness of this coding is important, as sensitivity analyses have demonstrated that missing data (traits or species) can lead to inaccurate functional-diversity measurements (Pakeman 2014); this impact is also mitigated here by having replicate samples and the use of a model-selection routine that explicitly incorporates variation among simulation trials. Indeterminate taxa (e.g., trepostome bryozoan indet. or *Platystrophia* sp.) that occurred within individual samples (but were excluded from the aggregate 237-taxon species pool unless their occurrence was the sole member of that taxon, see below) were coded for a particular state only when all other members of that taxon within the Kope-Waynesville species pool unanimously shared that common state; otherwise, the state was listed as NA (missing). Although there are differences in taxa and functional traits between the Kope and Waynesville Formation species pools (especially resulting from the Richmondian invasion, Holland and Patzkowsky 2007, 2009), a single aggregate pool is used here for simplicity. Use of formation-specific species-pool models do not alter the current results.

Eight ecological disparity and functional diversity statistics were calculated for the 218 Kope and Waynesville samples, and for section, member, and formation-level aggregates of these samples. The routine used to produce temporal and geographical aggregates only retained indeterminate taxa if no other members of the same taxon were found in the aggregate; this logic also applied to genus-level occurrences, excluding them when a named species of the genus was present. This left 135 unique taxa (representing 83 unique life habits) in the Kope Formation aggregate and 174 unique taxa (93 unique life habits) in the Waynesville Formation aggregate. Eight samples were excluded from model-fitting protocols because they had fewer than four life habits, the minimum needed to calculate the functional-diversity statistics.

Simulations, constraints, and model selection.—Two criteria must be met when considering candidate models for these samples (Burnham and Anderson 2002, Johnson and Omland 2004). The first concerns adequacy: do the models predict values of similar magnitude to those found in the samples? The second concerns model selection: among adequate alternative candidates, which model best represents the empirical samples? "Best" in this context can be defined in multiple ways, such as model simplicity, maximum likelihood, or Chi-square or

Kolmogorov-Smirnoff statistics (Burnham and Anderson 2002), but classification trees trained (and validated) on simulated data sets are used here because of the non-independent, multidimensional nature of the model statistics.

Five seed species were used in the simulations, as this approximated (the 12th percentile) the sample size of the smallest samples, and this value allows calculation of all functional-diversity statistics and provides powerful model discrimination using classification trees. To obtain adequate models, simulations were implemented iteratively, altering constraints of the ecospace framework to produce models with statistics similar to those found in empirical samples and aggregates (Fig. 3). Note that simulations proceeded in identical manner in each simulation; only constraints delimiting the nature of the theoretically possible ecospace were modified. As discussed below, the alterations needed to achieve realistic statistics are themselves informative as to the types of constraints that may exist in defining the realized Kope/Waynesville ecospace.

The simulations used when building the models in Figure 2B used an ecospace framework that allowed at most "two presences" for each binary character ('constraint=2'; see "Theoretical Ecospace Framework" above). These models all fail the first test of adequacy. Except for H, FEve and FDiv, the model simulations predict disparity and functional diversity statistics much larger than the values found in the actual Kope and Waynesville samples (compare with points in Fig. 3A). It is unwise to select the best model when all perform poorly. The inadequacy is attributable to the relatively unconstrained framework, which allows for 930 billion unique life habits to occur in simulations, many of which are unlikely or impossible to occur in nature (such as animals that feed using any two foraging combinations among ambient, filter, attachment, mass, and raptorial habits). The highly multidimensional nature of this framework provides so many opportunities for unrealistic life habits that empirical samples are bound to be depauperate by comparison.

Simulations in Figure 3A further constrain the framework so that life habits can inhabit only a single binary character state ('constraint=1'; e.g., only infaunal or epifaunal are allowed but not semi-infaunal). This additional constraint diminishes all statistics by a negligible amount, but the overall dynamics are largely unchanged, with the largest differences occurring in the partitioning simulations. Empirical samples continue to have statistics substantially below those predicted by these models, implying that even this more constrained framework—allowing only 1.13 billion unique life habits—is too permissive to represent life habits that actually occur in nature. For example, it "permits" animals the size of humans to crawl between grains of sediment, a logically impossible life habit. However, many of the life habits created by this framework are biologically possible, even if they require some creativity to envision an organism inhabiting such a life habit. As an example, the framework could happen upon a "buried" microscopic, fluid-feeding carnivore that feeds facultatively attached to its prey on hard substrates while living far above the sea floor. Although seemingly implausible, this is a typical life habit for numerous parasites (cf., Huntley and Scarponi 2012). The exposure of such unlikely (or at least rarely fossilized) life habits (see Novack-Gottshall 2007 for others) is not a flaw or bias in the simulations or analyses, and can reveal important constraints that structure actual biotas (Raup 1966, Seilacher 1970, Thomas and Reif 1993).

The generally unchanged dynamics in Figures 3A and 2B, despite their ability to represent orders-of-magnitude different numbers of life habits, also demonstrate that the inclusion of large numbers of potentially illogical or biologically unreasonable life habits has little effect on resulting dynamics. The imposed "single-presence" constraint is also biologically unrealistic, as 52% of the Kope/Waynesville taxa are coded using "two-presence" characters (e.g., corals reproduce both sexually and asexually, semi-infaunal bivalves are common, and many crinoids, brachiopods, and bryozoans are known to attach to lithic and biotic substrates). Different kinds of constraints must be imposed to allow the models to approximate actual Ordovician samples.

The next simulations relax the "one-presence" constraint to allow for "two-presence" characters, but use the Kope/Waynesville aggregate pool to weight the inhabitation of character states. For example, because 87% of the pool taxa are coded as epifaunal, 9% as infaunal, and 4% as semi-infaunal, simulated life habits within the neutral model (and seed species for all models) will mimic these proportions, on average. Similarly, because no taxa are coded as simultaneously autotrophic and carnivorous, that combination is removed as a possible character state across all simulations, as are other such non-occurring ones. Imposing such empirical weighting still allows for 57 billion unique life habits, and allows the ecospace framework to mimic important features of the empirical species pool, albeit indirectly. For example, although the simulation rules do not specify actual predator:prey or producer:consumer trophic-group ratios (Van Valkenburgh 1988, Dunne et al. 2008, Mitchell et al. 2012, Hatton et al. 2015), size-diet relationships (Van Valkenburgh and Molnar 2002, Codron et al. 2013), or other regulating factors, the weighting allows the simulation to indirectly mirror these empirical ecological rules. As a result of this weighting, most of the simulated life habits also correspond to biologically reasonable life habits commonly inhabited by animals, living or extinct. Neutral simulations yield life habits differing by ~4 states (out of 37 in the framework) from ones actually occurring in the Late Ordovician species pool (after allowing the simulation to proceed through completion at 50 species per sample). The dynamically restricted models produce life habits differing by less than 1 state from actual Ordovician life habits for redundancy (with 95% rule-following) and strict partitioning simulations, and ~2 states for the relaxed partitioning simulations. The expansion model, as intended, explores the most disparate (and slightly unreasonable) life habits, but still produces life habits that differ by ~10 states from actual Ordovician life habits. Thus, the life habits produced by the simulations, when weighted by empirical species pools, are biologically quite reasonable approximations of reality. (It should be emphasized, however, that the model-selection method only considers the disparity structure of each assemblage, and not the identity of simulated life habits.) Further analysis of the nature of these constraints is beyond the scope of this study, but it seems reasonable that the implementation of empirical weighting is a practical solution to restrict the theoretical ecospace to those approximate regions locally allowed by whatever combination of historical, biological (adaptational), and structural constraints shaped the species pool (Raup 1966, Seilacher 1970, Thomas and Reif 1993).

Implementing these weightings also result in simulations (Fig. 3B) that reflect adequately the statistics of empirical samples across all eight statistics, with dynamics substantially diminished compared to the unweighted simulations. Statistical dynamics are also quite altered, although the rank order of models remains generally unchanged: expansion (also the least altered, with asymptotic values negligibly diminished from prior unweighted implementations) has the largest disparity/diversity values, followed by neutral (now dynamically distinct from expansion), the two partitioning implementations, and with redundancy at lowest values. The distinct model dynamics also result in improved performance for the classification tree, with 96% of the training data sets classified correctly (compared to 91% for the one- and two-constraint simulations), primarily resulting from improved ability to distinguish the expansion and neutral models (see Supplementary Appendix 2 for details).

Although the empirical samples adequately overlay model dynamics, they follow the mean trends only rarely, occurring more frequently in the region between the passive neutral model and the other active models. This could indicate that the best models to represent the samples are not 100% implementations of the models, but instead weakened versions of the models. This is illustrated in Figure 3C where eight additional "weaker" simulations are added, in which the model rules are followed on average 90% and 50% of the time. In other words, for the 50%-strength implementation, a computational "coin flip" was used in the assignment for each taxon's character states, whether to assign it at random (weighted by the species pool) or whether to use the model rules to assign it (but still only allowing states present in the species pool). The neutral model represents a 0% simulation, in which no active model rules are followed, and model dynamics coalesce toward the neutral model as they become weaker and weaker. Most empirical samples overlay the weakened model dynamics. Although many of the nine models share similar dynamics (more so given the often overlapping variation around each mean trend line), the classification tree displays a remarkable ability to distinguish them, with 90% of training data sets classified correctly, and 73% correct when tested against independently created validation data sets.

The classification tree used below in model selection of samples is simpler, using just the 100% and 50% rule-following models in Figure 2C as training data sets; it classified 92% of training data sets and 78% of validation data sets correctly (Supplementary Appendix 2). Although the training data sets here represent just two strength levels (50% and 100%), validation samples produced at different strengths (75%, 90%, and 95%) indicate the simple "two-strength" classification tree is able to provide generalizable approximations for other model-strength implementations (Supplementary Appendix 2; Supplementary Fig. 9).

Performance remains strong at sample sizes as low as six, the minimum sample size when the models could be classified (i.e., because five were seeded without any assembly rules). (It is reassuring that if the simulated assemblages with five or less species are submitted to model selection, they are overwhelmingly, and correctly, classified as representing neutral models.) More complicated trees (i.e., those trained on 75%, 90%, and 95%-strength models) are used to confirm the strength of models, especially for larger samples, where these more complex trees are better suited for distinguishing their less dissimilar dynamics.

The majority of Kope and Waynesville samples are classified (Fig. 4, Supplementary Appendix 4) as representing partitioning models (73% and 74% respectively), with most classified as 100%-strength versions of the model (51% and 43%, respectively). Some samples are classified as weak (50%-strength) redundancy models (14% and 16% respectively), and negligible numbers are classified as neutral or expansion models. A Chi-squared test confirms that the model classifications are statistically indistinguishable between formations (X^2=48, df=36, *p*-value=0.087). Average model support (measured in terms of proportion of tree-classification votes cast) for the "best" model is 0.50, above the threshold of 0.20 in which support can be considered substantial and 0.40 when it can be considered unambiguously strong (see Supplementary Fig. 9). Only one sample had winning support below this threshold, a small seven-taxon Waynesville sample deemed neutral with support of 0.18. The relatively low numbers of samples classified as weakened partitioning models (less than 5% of samples) also suggests that most Kope and Waynesville samples are best represented by strong (>75%, and likely >90%) partitioning models, based on sensitivity analyses in Supplementary Figure 9. This is confirmed using more complicated classification trees: when additional strength simulations are included (i.e., when adding 90% as in Fig. 3C, or adding 75%, 90%, and 95% to the training data sets), most samples remain classified as strong (90, 95% and 100%) versions of strict partitioning, 100%-strength relaxed partitioning, and 75% and 90%-weakened redundancy models (Supplementary Appendix 4). Classification trees and simulations that used separate species pools for each formation (i.e., creating a Kope tree trained only on Kope species pool simulations) also produced similar model classifications, as expected given that 42% of genera and 69% of life habits are identical in each separate pool.

Representative samples classified by the four driven models are illustrated in Figure 5 on two-dimensional non-metric multidimensional scaling plots. As expected, samples classified as partitioning models demonstrate continuous gradations, either in a linear fashion for the strict partitioning version of the model or in a central cluster in the relaxed version. Samples classified as weakened redundancy demonstrate discrete clusters of life habits, often with multiple taxa sharing similar or identical life habits. At large sample sizes, both the partitioning and redundancy models yield many instances of taxa with nearly overlapping life habits, although those produced by partitioning tend to have more continuous gradients between these clusters (compare Figs. 1 and 5). Samples classified as expansion tend toward relatively large distances between life habits, often have the centroid empty, and encompass a broad range of the ecospace, even at small sample sizes.

If the model classifications indicate evidence for ecological structure within these generally autochthonous samples, then one might expect different models to emerge at larger temporal and spatial scales, where samples are aggregated into individual outcrops (sections) and stratigraphic members and formations (Patzkowsky and Holland 2003, Holland 2010, Tomašových and Kidwell 2010, Hautmann 2014). For example, if beta diversity is substantial, one might expect a preponderance of redundancy models, as samples with similar structure duplicate each other in aggregation. Alternatively, one might expect greater frequency of neutral models, as patterns at the larger scales increasingly reflect the regional species pool.

Such a transition is partially borne out here. The Kope and Waynesville stratigraphic sections have model classifications similar to those of their individual samples (Fig. 4), dominated by 100%-strength partitioning (albeit only the strict implementation) and the remainder largely weak redundancy. The sole section deemed neutral has the smallest sample size and lowest (but still substantial) model support of 0.34. These classifications are generally supported when using more complex classification trees, with all eight Kope sections best classified as 90–100% strict partitioning models, but a greater number of Waynesville sections classified as 50–90% redundancy models (Supplementary Appendix 4).

At the larger sample sizes of stratigraphic members and formations, it is possible to use more complex classification trees (i.e., adding 90%-strength training data sets or adding 75%, 90%, and 95- strength implementations) to conduct model selection. This is especially worthwhile because the dynamics of redundancy and both partitioning models are similar if even slightly weakened (i.e. 90–99%). This is evident by their overlapping dynamics in Figure 3C and when examining model classifications and confusion matrices on validation data sets (Supplementary Fig. 9, Supplementary Appendix 2). In particular, slightly weakened (90%-95%) redundancy models are typically misclassified by the simple tree as strong strict partitioning models at sample sizes greater than 30 (Supplementary Fig. 9), and moderately weak (75–90%) relaxed partitioning models are typically misclassified as strong relaxed partitioning and weak redundancy models (Supplementary Fig. 9). Note that these misclassifications tend to only occur at sample sizes greater than 30, where the more asymptotic dynamics make it easier for the simple classification tree to reject implementations that are different from those in the training data set. One might argue that these classifications point to indistinguishable dynamics for these models, regardless of whether named "redundancy" or "partitioning." Yet, the simple tree is able to correctly distinguish these models more than 33% of the time (Supplementary Fig. 9). In contrast, the more complex three-strength (50%, 75%, and 100%) and five-strength (50%, 75%, 90%, 95%, and 100%) trees classified 90% of training data sets correctly, and displays improved ability to distinguish all weakened implementations of each model.

When using the five-strength tree, all four Kope members are classified as strong (90–95%) strict partitioning models, consistent with model classifications at smaller scales (Fig. 4). The Waynesville members have divided support, with three best classified as 90–95% strict partitioning and two as 90% redundant models (Supplementary Appendix 4). Classifications using the three-strength tree are identical, but with the 95% votes classified instead as 90%-strength implementations.

The entire Kope Formation species pool, using the five-strength tree (Fig. 4), is classified as 90% strict partitioning (0.52 support), with substantial support for 90% redundancy model (0.34), whereas the Waynesville formation pool is classified as 90% redundancy (0.78 support) with the remainder for 95% strict partitioning. The three-strength tree provides similar classifications (Supplementary Appendix 4). This classification is intuitively evident when considering the ratio of unique life habits to taxa (H/S) in each formation: the Waynesville pool has substantially more true redundancy (81 of 174 taxa, or 47%, are functionally identical) compared to the Kope (with 52 of 135 taxa, or 37% functionally identical). The structure of each formation is visualized in Figure 5. Although they look quite similar (the most obvious difference being the presence of corals in the Waynesville, one of many new immigrants during the Richmondian invasion, Holland and Patzkowsky 2007, 2009), the classification trees are sufficiently sensitive to their structural differences to classify them accordingly.

When comparing classifications across spatially and temporally mixed scales, the Kope Formation remains consistently classified by strong (>90%) strict partitioning models. There is no evidence of change in ecospace structure across scales. In contrast, the Waynesville is typically classified as representing partitioning models (both relaxed and strict versions) at the scale of samples, switching to greater support for ~90% redundant models at larger scales, especially when considering the entire formation species pool. The consistent lack of support for neutral and expansion models at any scale is evidence that the ecospace structure of these Late Ordovician samples is decidedly nonrandom and restricted. Even when weighted by the life-habit traits known to occur in the Late Ordovician regional species pool, the ecospace actually inhabited by this biota remains substantially constrained.

Although the simulations do not incorporate phylogenetic structure (newly added species are added solely on account of their functional traits), it is promising that model selection is not a simple artifact of which taxa are present within a sample (Fig. 5). If it were, one might expect that samples classified as redundancy or partitioning be dominated by one or a few taxonomic groups, whereas samples classified as expansion might have greater taxonomic diversity. That is not the case here, where a wide range of major taxonomic groups is present at all scales, and where clusters of functionally similar species likewise tend to be taxonomically diverse. The same ecological structure reoccurs when taxonomically quite different biotas are analyzed.

Ecological and evolutionary theory underlying these models (see companion article Novack-Gottshall 2016) allows speculation as to the nature of processes structuring these biotas. At the scale of individual samples, partitioning dynamics are consistent with ecological niche-partitioning: co-existing taxa have similar (but not identical) life habits. Holland and Patzkowsky (2007, Patzkowsky and Holland 2007) demonstrated similar tight packing of species along onshore-offshore gradients in these intervals, even more so in the Waynesville samples. This partitioning could be manifested at the scale of local communities or through evolutionary processes at the regional scale, in which speciation produces multiple variants with slightly different specializations. Consideration of phylogenetic relatedness could offer a useful test of these claims (Gerhold et al. 2015), in particular whether newly evolving taxa within the regional species pool were more likely to have life habits different from those previously existing. Patzkowsky and Holland (2003) found no evidence for saturation within Cincinnatian samples, suggesting regional processes as the more important factor. This conclusion is tentatively supported here, especially for the Kope Formation, where the same model support across all scales could suggest a shared cause. Within the Waynesville sequence, Patzkowsky and Holland (2007) identified greater beta diversity, driven in large part by faunal incursions during the Richmondian invasion. This could explain the redundancy dynamics found within the Waynesville Formation and its members, in which the accumulation of many different life habits within individual samples results in functional duplication when aggregated.

These results, however, do not rule out an important role for local processes. Although not illustrated here, individual Kope and Waynesville samples, even those classified by the same model, do not share the same life-habits (or taxa). In other words, when plotted on a common ordination (such as for the partitioning samples in Fig. 5), taxonomic and functional composition varies substantially among samples. Model selection results suggest simply that most samples share a similar level of ecological disparity (Fig. 4), one best represented by the partitioning model. This consistency of structure suggests that local processes could still play an important role in regulating these communities. Among the many possible life habits available within the regional species pool, individual communities were preferentially composed of groups of generally similar but non-identical life habits, although the particular life habits present in any setting could vary a great deal.

Relevance to broader Phanerozoic trends in ecospace utilization

The general correspondence between empirical samples and the dynamics of the model simulations suggests that the simulations provide reasonable null models for understanding how ecospace is structured in extant and fossil biotas, at many spatio-temporal scales. Support for the partitioning model, and to a lesser extent redundancy, indicates that these Late Ordovician biotas had relatively constrained ecospace structures. Broader discussion of this model support is warranted, especially within the context of how these results might generalize to understanding Phanerozoic-wide trends in ecospace utilization. In particular, one might want to know whether these results are consistent with an expansion in ecospace utilization in later Phanerozoic biotas, as has been widely claimed (Bambach 1983, Vermeij 1987, Knoll and Bambach 2000, Bush et al. 2007, Novack-Gottshall 2007, Bush and Bambach 2011, Vermeij 2011, 2013). A conclusion cannot be made at this point, but speculation might be made based on the interactions of three factors: changes in the species pool used (which changes allowable ecospace traits and resulting simulation dynamics), measures of functional diversity/disparity in later biotas, and changes in taxonomic richness.

The most important of these factors is the species pool one chooses to use (cf., Cornell 1999, de Bello 2012). The analyses above used a narrowly defined one, the aggregate pool of species known from a single habitat in one small (tri-state) region during a short (~8 m.y.) time interval. Analyses using a Kope-only versus Waynesville-only pool had negligible effect, but future analyses should test the sensitivity of simulation dynamics to more distinct species pools. Use of a much larger pool would allow for a greater range of life-habits in the framework, which could increase dynamical values in simulations (although this remains to be determined). The effect of this change might be to increase classification-tree support for the partitioning and redundancy models, and especially so if using a Paleozoic-wide or Phanerozoic-wide species pool, given the greater range of epifaunal and infaunal tiering (Ausich and Bottjer 1982, Bottjer and Ausich 1986) and body sizes (Novack-Gottshall 2008a, Heim et al. 2015, Zhang et al. 2015) evolved by later biotas. Because the analyses above rescaled these ordered numeric characters, incorporating broader ranges would depress the empirical disparity statistics calculated here to some extent. However, the effect of this is likely minor, as post-Ordovician ranges would only add a single bin (given the logarithmic binning scale used in the framework) and the largest, most deeply infaunal, and tallest tiering animals were likely proportionally minor components of these biotas (which diminishes their effect for these extreme character states because of the empirical weighting used). It remains to be determined what pool is best suited to such analyses given the major evolutionary changes throughout the Phanerozoic, but the use of more inclusive pools would likely demonstrate that these Ordovician samples are even more functionally restricted (i.e., less ecologically disparate) compared to later Phanerozoic samples.

Species richness has also increased during the Phanerozoic, both globally (Sepkoski 1981, Alroy et al. 2008) and within individual assemblages (Bambach 1977, Powell and Kowalewski 2002, Bush and Bambach 2004, Kowalewski et al. 2006). The effect of this increase, by itself, is also likely to be limited here because the dynamics tend to reach asymptotic values at moderate (>20) species richness values, which means that simply increasing sample sizes, all things equal, will yield negligible changes to statistical values. The greater effect will be related to how these larger biotas utilize ecospace. Novack-Gottshall (2007) demonstrated that extant marine biotas had greater ecological disparity (measured as D, after accounting for differences in sample sizes) than Early–Middle Paleozoic ones, despite insignificantly greater number of life habits. Increasing values for statistics D and H through time could provide greater support for the expansion model for such biotas, although it remains uncertain whether this increase would be attenuated by using a larger species pool. It is worth reiterating that one would expect these statistics to increase with increasing species richness, even if ecospace was occupied by a passive process (i.e., that expected by the neutral model). The discussion above includes many subjective hypotheticals, and so this author is unwilling to make a wager on the outcome. Without conducting formal simulations and statistical analyses, it simply remains too early to predict whether the Phanerozoic trend would be better supported by the driven expansion model, the passive neutral one, or perhaps another. It seems a worthwhile goal to conduct such analyses.

Conclusions

Despite documentation of synoptic paleoecological trends across the Phanerozoic and speculation about their causes, we lack, in many—perhaps most—cases, specific quantitative claims on ecospace inhabitation that can be tested analytically. The models and analytical methods proposed here, none of which are entirely novel, offer potentially fruitful avenues for such tests, whether applied to individual assemblages or to the entire biosphere throughout the Geozoic. In particular, the analyses presented here, and in an accompanying article, lead to the following conclusions.

1. Simulations of the redundancy, partitioning, expansion, and neutral models demonstrate dynamical consistency in functional diversity/disparity statistics across different data structures (ecospace frameworks), number and type of functional traits (characters), and implementations of the simulations. However, ecospace frameworks with greater numbers of functional traits, use of ordered factors, and modest numbers of seed species tend to be statistically more powerful for differentiating the models, especially the dynamically similar neutral and expansion models and the often-similar partitioning and redundancy models. The 'ecospace' package provides R functions to conduct simulations for any ecospace framework and to calculate a wide range of functional diversity/disparity statistics.
2. Classification trees are a powerful method for rigorously classifying these models in a multi-model inference framework (Breiman et al. 1993, Cutler et al. 2007), with relative support allocated to candidate models according to proportion of votes from the classification tree. Classification trees are successful in identifying models correctly, even when the statistical dynamics are similar and when tested with foreign data sets unlike those used to train the tree. The trees are also able to identify which statistics are most valuable. This method identifies number of unique life-habits (functional trait combinations, H) as the most important statistical discriminant, followed by functional evenness (FEve) and functional dispersion (FDis). However, because all metrics retain useful information on ecospace structure and tree algorithms perform well using large number of predictive variables, it is recommended that analyses of ecological disparity/functional diversity use all statistics when conducting model-selection analyses.
3. Comparison of stochastic simulation dynamics to those of Ordovician empirical samples demonstrates that actual fossil assemblages are substantially constrained in their inhabitation

of life habits, compared to what is possible in the theoretical ecospace framework. Although the identity of constraints are not analyzed here specific, they likely reflect a combination of logically impossible trait combinations, maladaptive strategies, inherent covariation among functional traits, ecologically meaningful restrictions to the regional species pool, and other factors. Incorporating constraints to the ecospace framework (such as limiting allowed life-habit combinations and weighting functional traits by those occurring in the empirical species pool) causes simulation dynamics to converge on more realistic values, allowing simple approximations for many of these constraints. However, doing so still demonstrates that these Late Ordovician biotas had substantially constrained ecospace structures.

4. Empirical application of the classification trees demonstrate that Late Ordovician (type Cincinnatian) samples from the Kope and Waynesville Formations are primarily best fit by partitioning models. When larger stratigraphic and temporal aggregates are analyzed, the entire Kope Formation pool remains best fit by the partitioning model, but the aggregate Waynesville pool is better fit by the redundancy model. This structural transition in the Waynesville Formation can be biologically interpreted by greater beta diversity in this unit, likely related to faunal incursions caused by the Richmondian invasion. However, the consistency of support for the partitioning model at small scales suggests an important role for local processes.
5. Most hypotheses regarding patterns in ecospace utilization across the Phanerozoic are superficially consistent with multiple models of ecological diversification, despite being caused by distinct processes. Most documented trends are equally consistent with passive processes. Statistical tests that consider alternative stochastic models must be conducted before we can confidently claim that ecological patterns across the Geozoic history of life

had driven causes. Similar concerns are shared with ongoing research in community ecology and functional ecology.

Acknowledgments

I thank S. C. Wang and G. Hunt for discussion on model fitting (it was S. C. Wang who originally recommended the use of classification trees for this purpose, albeit in a different context); A. Bush for discussion of ecological models; E. Laliberté and S. C. Wang for assistance programming in R; D. W. Bapst and G. Hunt for assistance building the R package; C. Ciampaglio, S. Villéger, and M. A. Wills for discussion of metrics; F. T. Kuserk, J. Ohles and staff at Reeves Library (Moravian College), where much of the manuscript was written; and D. Aleinikava, R. Meeker, and B. Ng for assistance, access, and funding of the Benedictine University High-Performance Computational Cluster. I also thank R. G. Browne, E. Dalvé, R. C. Frey, J. J. Sepkoski, Jr., S. M. Holland, and M. E. Patzkowsky for collecting and compiling the Paleobiology Database samples used here. Research and manuscript support was facilitated with a sabbatical leave provided by M. J. de la Cámara and Faculty Development (Benedictine University). This paper was strengthened by thoughtful reviews from M. Foote and S. M. Holland. This is Paleobiology Database Publication 244.

Literature Cited

Aberhan, M., and W. Kiessling. 2015. Persistent ecological shifts in marine molluscan assemblages across the end-Cretaceous mass extinction. Proceedings of the National Academy of Sciences 112(23):7207-7212.

Ackerly, D. D., and W. K. Cornwell. 2007. A trait-based approach to community assembly: partitioning of species trait values into within- and among-community components. Ecology Letters 10(2):135-145.

Alroy, J., M. Aberhan, D. J. Bottjer, M. Foote, F. T. Fursich, P. J. Harries, A. J. W. Hendy, S. M. Holland, L. C. Ivany, W. Kiessling, M. A. Kosnik, C. R. Marshall, A. J. McGowan, A. I. Miller, T. D. Olszewski, M. E. Patzkowsky, S. E. Peters, L. Villier, P. J. Wagner, N. Bonuso, P. S. Borkow, B. Brenneis, M. E. Clapham, L. M. Fall, C. A. Ferguson, V. L. Hanson, A. Z. Krug, K. M. Layou, E. H. Leckey, S. Nürnberg, C. M. Powers, J. A. Sessa, C. Simpson, A. Tomašových, and C. C. Visaggi. 2008. Phanerozoic trends in the global diversity of marine invertebrates. Science 321(5885):97-100.

Anderson, D. R., K. P. Burnham, and W. L. Thompson. 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64(4):912-923.

Anderson, M. J., K. E. Ellingsen, and B. H. McArdle. 2006. Multivariate dispersion as a measure of beta diversity. Ecology Letters 9(6):683-693.

Aucoin, C. D., B. F. Dattilo, C. E. Brett, and D. L. Cooper. 2015. Preliminary report on the Oldenburg "butter shale" in the Upper Ordovician (Katian; Richmondian) Waynesville Formation, USA. Estonian Journal of Earth Sciences 64:3-7.

Ausich, W. I., and D. J. Bottjer. 1982. Tiering in suspension feeding communities on soft substrata throughout the Phanerozoic. Science 216:173-174.

Bambach, R. K. 1977. Species richness in marine benthic habitats through the Phanerozoic. Paleobiology 3(2):152-167.

---. 1983. Ecospace utilization and guilds in marine communities through the Phanerozoic. Pp. 719–746. *In* M. J. S. Tevesz, and P. L. McCall, eds. Biotic Interactions in Recent and Fossil Benthic Communities. Plenum, New York.

---. 1985. Classes and adaptive variety: the ecology of diversification in marine faunas through the Phanerozoic. Pp. 191–253. *In* J. W. Valentine, ed. Phanerozoic Diversity Patterns: Profiles in Macroevolution. Princeton University Press, Princeton, NJ.

Bambach, R. K., A. M. Bush, and D. H. Erwin. 2007. Autecology and the filling of ecospace: key metazoan radiations. Palaeontology 50(1):1-22.

Beaumont, M. A. 2010. Approximate Bayesian computation in evolution and ecology. Annual review of ecology, evolution, and systematics 41(1):379-406.

Bottjer, D. J., and W. I. Ausich. 1986. Phanerozoic development of tiering in soft substrata suspension-feeding communities. Paleobiology 12:400-420.

Boyer, A. G. 2010. Consistent ecological selectivity through time in Pacific Island avian extinctions. Conservation Biology 24(2):511-519.

Brandt, D. S. 1996. Epizoans on *Flexicalymene* (Trilobita) and implications for trilobite paleoecology. Journal of Paleontology 70:442-449.

Brandt, D. S., D. L. Meyer, and P. B. Lask. 1995. *Isotelus* (Trilobita) "hunting burrow" from Upper Ordovician strata, Ohio. Journal of Paleontology 69:1079-1083.

Breiman, L. 2006. randomForest: Breiman and Cutler's random forests for classification and regression, Version 4.6-10. cran.r-project.org/web/packages/randomForest.

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. Classification and Regression Trees. Wadsworth & Brooks.

---. 1993. Classification and Regression Trees. Chapman and Hall/CRC press.

Browne, R. G. 1964. The coral horizons and stratigraphy of the Upper Richmond Group in Kentucky West of the Cincinnati Arch. Journal of Paleontology 38(2):385-392.

Burnham, K. P., and D. R. Anderson. 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach. Springer, New York.

Bush, A. M., and R. K. Bambach. 2004. Did alpha diversity increase during the Phanerozoic? Lifting the veils of taphonomic, latitudinal, and environmental biases. The Journal of geology 112(6):625-642.

---. 2011. Paleoecologic megatrends in marine Metazoa. Annual Review of Earth and Planetary Sciences 39:241-269.

Bush, A. M., R. K. Bambach, and G. M. Daley. 2007. Changes in theoretical ecospace utilization in marine fossil assemblages between the mid-Paleozoic and late Cenozoic. Paleobiology 33(1):76-97.

Bush, A. M., R. K. Bambach, and D. H. Erwin. 2011. Ecospace utilization during the Ediacaran radiation and the Cambrian eco-explosion. Pp. 111-134. *In* M. Laflamme, J. D. Schiffbauer, and S. Q. Dornbos, eds. Quantifying the Evolution of Early Life: Numerical Approaches to the Evaluation of Fossils and Ancient Ecosystems. Springer, New York.

Bush, A. M., and P. M. Novack-Gottshall. 2012. Modelling the ecological-functional diversification of marine Metazoa on geological time scales. Biology Letters 8(1):151-155.

Bush, A. M., and S. B. Pruss. 2013. Theoretical ecospace for ecosystem paleobiology: energy, nutrients, biominerals, and macroevolution. Pp. X-XX. *In* A. M. Bush, S. B. Pruss, and J. L. Payne, eds. Ecosystem Paleobiology and Geobiology. Short Courses in Paleontology 19. Paleontological Society and Paleontological Research Institute, Ithaca, NY.

Ciampaglio, C. N., M. Kemp, and D. W. McShea. 2001. Detecting changes in morphospace occupation patterns in the fossil record: characterization and analysis of measures of disparity. Paleobiology 27(4):695-715.

Codron, D., C. Carbone, and M. Clauss. 2013. Ecological interactions in dinosaur communities: influences of small offspring and complex ontogenetic life histories. PLoS One 8(10):e77110.

Cornell, H. V. 1999. Unsaturation and regional influences on species richness in ecological communities: a review of the evidence. Ecoscience 6:303-315.

Cornell, H. V., and S. P. Harrison. 2014. What are species pools and when are they important? Annual review of ecology, evolution, and systematics 45(1):45-67.

Cutler, D. R., T. C. Edwards, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. 2007. Random forests for classification in ecology. Ecology 88(11):2783-2792.

Dalvé, E. 1948. The fossil fauna of the Ordovician in the Cincinnati region. University Museum, Department of Geology and Geography, University of Cincinnati.

De'ath, G. 2002. Multivariate regression trees: a new technique for modeling species-environment relationships. Ecology 83(4):1105-1117.

De'ath, G., and K. E. Fabricius. 2000. Classification and regression trees: a powerful yet simple technique for ecological data analysis. Ecology 81(11):3178-3192.

de Bello, F. 2012. The quest for trait convergence and divergence in community assembly: are null-models the magic wand? Global Ecology and Biogeography 21(3):312-317.

Deline, B. 2009. The effects of rarity and abundance distributions on measurements of local morphological disparity. Paleobiology 35(2):175-189.

Deline, B., W. I. Ausich, and C. E. Brett. 2012. Comparing taxonomic and geographic scales in the morphologic disparity of Ordovician through Early Silurian Laurentian crinoids. Paleobiology 38(4):538-553.

Dick, D. G., and E. E. Maxwell. 2015. The evolution and extinction of the ichthyosaurs from the perspective of quantitative ecospace modelling. Biology Letters 11(7).

Dineen, A. A., M. L. Fraiser, and P. M. Sheehan. 2014. Quantifying functional diversity in pre- and post-extinction paleocommunities: A test of ecological restructuring after the end-Permian mass extinction. Earth-Science Reviews 136:339-349.

Dineen, A. A., M. L. Fraiser, and J. Tong. 2015. Low functional evenness in a post-extinction Anisian (Middle Triassic) paleocommunity: A case study of the Leidapo Member (Qingyan Formation), south China. Global and Planetary Change 133:79-86.

Dunne, J. A., R. J. Williams, N. D. Martinez, R. A. Wood, and D. H. Erwin. 2008. Compilation and network analyses of Cambrian food webs. PLoS Biol 6(4):e102.

Durst, P. A. P., and V. L. Roth. 2012. Classification tree methods provide a multifactorial approach to predicting insular body size evolution in rodents. American Naturalist 179(4):545-553.

English, A. M., and L. E. Babcock. 2007. Feeding behaviour of two Ordovician trilobites inferred from trace fossils and non-biomineralised anatomy, Ohio and Kentucky, USA. Memoirs of the Association of Australasian Palaeontologists (34):537.

Fargione, J., C. S. Brown, and D. Tilman. 2003. Community assembly and invasion: An experimental test of neutral versus niche processes. Proceedings of the National Academy of Sciences 100(15):8916-8920.

Feldmann, R. M. 1996. Fossils of Ohio.

Finnegan, S., S. C. Anderson, P. G. Harnik, C. Simpson, D. P. Tittensor, J. E. Byrnes, Z. V. Finkel, D. R. Lindberg, L. H. Liow, R. Lockwood, H. K. Lotze, C. R. McClain, J. L. McGuire, A. O'Dea, and J. M. Pandolfi. 2015. Paleontological baselines for evaluating extinction risk in the modern oceans. Science 348(6234):567-570.

Finnegan, S., N. A. Heim, S. E. Peters, and W. W. Fischer. 2012. Climate change and the selective signature of the Late Ordovician mass extinction. Proceedings of the National Academy of Sciences.

Fontana, S., O. L. Petchey, and F. Pomati. 2015. Individual-level trait diversity concepts and indices to comprehensively describe community change in multidimensional trait space. Functional Ecology 123(11):1391-1399.

Foote, M. 1993. Discordance and concordance between morphological and taxonomic diversity. Paleobiology 19:185-204.

---. 1994. Morphological disparity in Ordovician-Devonian crinoids and the early saturation of morphological space. Paleobiology 20:320-344.

Fordyce, D., and T. W. Cronin. 1993. Trilobite vision: a comparison of schizochroal and holochroal eyes with the compound eyes of modern arthropods. Paleobiology 19(3):288-303.

Fortey, R. A., and R. M. Owens. 1999. The trilobite exoskeleton. Pp. 537-562. *In* E. Savazzi, ed. Functional Morphology of the Invertebrate Skeleton. John Wiley and Sons, Ltd., New York.

Freeman, R. L., B. F. Dattilo, A. Morse, M. Blair, S. Felton, and J. Pojeta. 2013. The "curse of Rafinesquina": Negative taphonomic feedback exerted by strophomenid shells on storm-buried lingulids in the Cincinnatian Series (Katian, Ordovician) of Ohio. PALAIOS 28(6):359-372.

Frey, R. C. 1987a. The occurrence of pelecypods in Early Paleozoic epeiric-sea environments: Late Ordovician of the Cincinnati, Ohio area. PALAIOS 2:3-23.

---. 1987b. The paleoecology of a Late Ordovician shale unit from southwest Ohio and southeastern Indiana. Journal of Paleontology 61:242-267.

---. 1988. Paleoecology of *Treptoceras duseri* (Michelinoceratida, Proteoceratidae) from Late Ordovician of southwestern Ohio. New Mexico Bureau of Mines and Mineral Resources Memoir 44:79-101.

---. 1989. Paleoecology of a well-preserved nautiloid assemblage from a Late Ordovician shale unit, southwestern Ohio. Journal of Paleontology 63:604-620.

Gaines, R. R., M. L. Droser, and N. C. Hughes. 1999. The ichnological record in Ordovician mudstones: examples from the Cincinnatian strata of Ohio and Kentucky (USA). Acta Universitatis Carolinae, Geologica (1/2):163-166.

Gerhold, P., J. F. Cahill, M. Winter, I. V. Bartish, and A. Prinzing. 2015. Phylogenetic patterns are not proxies of community assembly mechanisms (they are far better). Functional Ecology 29(5):600-614.

Gerisch, M. 2014. Non-random patterns of functional redundancy revealed in ground beetle communities facing an extreme flood event. Functional Ecology 28(6):1504-1512.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. Biometrics 27(4):857-871.

Grueber, C. E., S. Nakagawa, R. J. Laws, and I. G. Jamieson. 2011. Multimodel inference in ecology and evolution: challenges and solutions. Journal of Evolutionary Biology 24(4):699-711.

Guillemot, N., M. Kulbicki, P. Chabanet, and L. Vigliola. 2011. Functional redundancy patterns reveal non-random assembly rules in a species-rich marine assemblage. PLoS One 6(10):e26735.

Hatton, I. A., K. S. McCann, J. M. Fryxell, T. J. Davies, M. Smerlak, A. R. E. Sinclair, and M. Loreau. 2015. The predator-prey power law: Biomass scaling across terrestrial and aquatic biomes. Science 349(6252).

Hautmann, M. 2014. Diversification and diversity partitioning. Paleobiology:162-176.

Hegna, T. A. 2010. The function of forks: *Isotelus*-type hypostomes and trilobite feeding. Lethaia 43(3):411-419.

Heim, N. A., M. L. Knope, E. K. Schaal, S. C. Wang, and J. L. Payne. 2015. Cope's rule in the evolution of marine animals. Science 347(6224):867-870.

Holland, S. M. 1993. Sequence stratigraphy of a carbonate-clastic ramp: the Cincinnatian Series (Upper Ordovician) in its type area. Geological Society of America Bulletin 105:306-322.

---. 2010. Additive diversity partitioning in palaeobiology: revisiting Sepkoski’s question. Palaeontology 53(6):1237-1254.

Holland, S. M., A. I. Miller, B. F. Dattillo, and D. L. Meyer. 2001. The detection and importance of subtle biofacies in lithologically uniform strata: the Upper Ordovician Kope Formation of the Cincinnati, Ohio region. PALAIOS 16:205-217.

Holland, S. M., and M. E. Patzkowsky. 2007. Gradient ecology of a biotic invasion: biofacies of the type Cincinnatian series (Upper Ordovician), Cincinnati, Ohio region, USA. PALAIOS 22(4):392-407.

---. 2009. The Richmondian invasion: understanding the faunal response to climate change through stratigraphic paleobiology. Type Cincinnatian (Upper Ordovician) outcrops, northern Kentucky, southwestern Ohio, and southeastern Indiana. Pp. 1-67. The Richmondian Invasion in the type Cincinnatian Series. Fieldtrip guidebook, Ninth North American Paleontological Convention. Cincinnati, OH.

Hubbell, S. P. 2001. The Unified Theory of Biodiversity and Biogeography. Princeton University Press, Princeton, NJ.

Hughes, N. C., and D. L. Cooper. 1999. Paleobiologic and taphonomic aspects of the "*Granulosa*" trilobite cluster, Kope Formation (Upper Ordovician, Cincinnati region). Journal of Paleontology 73(2):306-319.

Huntley, J. W., and D. Scarponi. 2012. Evolutionary and ecological implications of trematode parasitism of modern and fossil northern Adriatic bivalves. Paleobiology 38(1):40-51.

Johnson, J. B., and K. S. Omland. 2004. Model selection in ecology and evolution. Trends in Ecology & Evolution 19(2):101-108.

Kinzig, A. P., S. A. Levin, J. Dushoff, and S. Pacala. 1999. Limiting similarity, species packing, and system stability for hierarchical competition-colonization models. American Naturalist 153:371-383.

Knoll, A. H., and R. K. Bambach. 2000. Directionality in the history of life: diffusion from the left wall or repeated scaling of the right? Paleobiology (Supplement) 26(sp4):1-14.

Knope, M. L., N. A. Heim, L. O. Frishkoff, and J. L. Payne. 2015. Limited role of functional differentiation in early diversification of animals. Nature Communications 6.

Kowalewski, M., W. Kiessling, M. Aberhan, F. T. Fürsich, D. Scarponi, S. L. Barbour Wood, and A. P. Hoffmeister. 2006. Ecological, taxonomic, and taphonomic components of the post-Paleozoic increase in sample-level species diversity of marine benthos. Paleobiology 32(4):533-561.

Kowalewski, M., and P. Novack-Gottshall. 2010. Resampling methods in paleontology. Pp. 19-54. *In* J. Alroy, and G. Hunt, eds. Quantitative Methods in Paleobiology. Short Courses in

Paleontology 16. Paleontological Society and Paleontological Research Institute, Ithaca, NY.

Laflamme, M., S. A. F. Darroch, S. M. Tweedt, K. J. Peterson, and D. H. Erwin. 2013. The end of the Ediacara biota: Extinction, biotic replacement, or Cheshire Cat? Gondwana Research 23(2):558-573.

Laflamme, M., S. Xiao, and M. Kowalewski. 2009. Osmotrophy in modular Ediacara organisms. Proceedings of the National Academy of Sciences (U.S.A.) 106(34):14438-14443.

Laliberté, E., and P. Legendre. 2010. A distance-based framework for measuring functional diversity from multiple traits. Ecology 91(1):299-305.

Laliberté, E., and B. Shipley. 2014. FD: Measuring functional diversity from multiple traits, and other tools for functional ecology, Version 1.0-12.

Legendre, P., and M. J. Anderson. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological Monographs 69(1):1-24.

Leighton, L. R. 1998. Constraining functional hypotheses: controls on the morphology of the concavo-convex brachiopod *Rafinesquina*. Lethaia 31:293-307.

Lescinsky, H. L. 1995. The life orientation of concavo-convex brachiopods: overturning the paradigm. Paleobiology 21:520-551.

Maire, E., G. Grenouillet, S. Brosse, and S. Villéger. 2015. How many dimensions are needed to accurately assess functional diversity? A pragmatic approach for assessing the quality of functional spaces. Global Ecology and Biogeography 24(6):728-740.

Mason, N. W. H., D. Mouillot, W. G. Lee, and J. B. Wilson. 2005. Functional richness, functional evenness and functional divergence: the primary components of functional diversity. Oikos 111(1):112-118.

McShea, D. W. 1994. Mechanisms of large-scale evolutionary trends. Evolution 48(6):1747-1763.

Meyer, D. L., A. I. Miller, S. M. Holland, and B. F. Dattilo. 2002. Crinoid distribution and feeding morphology through a depositional sequence: Kope and Fairview Formations, Upper Ordovician, Cincinnati Arch region. Journal of Paleontology 76(4):725-732.

Miller, J. H., A. K. Behrensmeyer, A. Du, S. K. Lyons, D. Patterson, A. Tóth, A. Villaseñor, E. Kanga, and D. Reed. 2014. Ecological fidelity of functional traits based on species presence-absence in a modern mammalian bone assemblage (Amboseli, Kenya). Paleobiology 40(4):560-583.

Mitchell, J. S., and P. J. Makovicky. 2014. Low ecological disparity in Early Cretaceous birds. Proceedings of the Royal Society B: Biological Sciences 281(1787).

Mitchell, J. S., P. D. Roopnarine, and K. D. Angielczyk. 2012. Late Cretaceous restructuring of terrestrial communities facilitated the end-Cretaceous mass extinction in North America. Proceedings of the National Academy of Sciences 109(46):18857-18861.

Mondal, S., and P. J. Harries. 2015. Phanerozoic trends in ecospace utilization: The bivalve perspective. Earth-Science Reviews.

Morris, R. W., and S. H. Felton. 2003. Paleoecologic associations and secondary tiering of *Cornulites* on crinoids and bivalves in the Upper Ordovician (Cincinnatian) of southwestern Ohio, southeastern Indiana, and northern Kentucky. PALAIOS 18(6):546-558.

Mouchet, M. A., S. Villéger, N. W. H. Mason, and D. Mouillot. 2010. Functional diversity measures: an overview of their redundancy and their ability to discriminate community assembly rules. Functional Ecology 24(4):867-876.

Mouillot, D., N. A. J. Graham, S. Villéger, N. W. H. Mason, and D. R. Bellwood. 2013. A functional approach reveals community responses to disturbances. Trends in Ecology and Evolution 28(3):167-177.

Novack-Gottshall, P. 2015. ecospace: Simulating Community Assembly and Ecological Diversification Using Ecospace Frameworks, Version 1.0.1. cran.r-project.org/package=ecospace.

Novack-Gottshall, P. M. 2007. Using a theoretical ecospace to quantify the ecological diversity of Paleozoic and modern marine biotas. Paleobiology 33(2):273-294.

---. 2008a. Ecosystem-wide body size trends in Cambrian-Devonian marine invertebrate lineages. Paleobiology 34(2).

---. 2008b. Using simple body-size metrics to estimate fossil body volume: empirical validation using diverse Paleozoic invertebrates. PALAIOS 23(3):163-173.

---. 2010. Performance of functional diversity metrics applied as measures of disparity. GSA Abstracts with Programs 42(5):140.

---. 2016. General models of ecological diversification. I. Conceptual synthesis. Paleobiology.

Novack-Gottshall, P. M., and A. I. Miller. 2003. Comparative taxonomic richness and abundance of Late Ordovician gastropods and bivalves in mollusc-rich strata of the Cincinnati Arch. PALAIOS 18(6):559-571.

O'Brien, L. J., and J.-B. Caron. 2015. Paleocommunity analysis of the Burgess Shale Tulip Beds, Mount Stephen, British Columbia: comparison with the Walcott Quarry and implications for community variation in the Burgess Shale. Paleobiology FirstView:1-27.

Pakeman, R. J. 2014. Functional trait metrics are sensitive to the completeness of the species' trait data? Methods in Ecology and Evolution 5(1):9-15.

Patzkowsky, M. E., and S. M. Holland. 2003. Lack of community saturation at the beginning of the Paleozoic plateau: the dominance of regional over local processes. Paleobiology 29:545-560.

---. 2007. Diversity partitioning of a Late Ordovician marine biotic invasion: controls on diversity in regional ecosystems. Paleobiology 33(2):295-309.

Pojeta, J., Jr. 1971. Review of Ordovician pelecypods. United States Geological Survey Professional Paper 695.

Powell, M. G., and M. Kowalewski. 2002. Increase in evenness and sampled alpha diversity through the Phanerozoic: comparison of early Paleozoic and Cenozoic marine fossil assemblages. Geology 30(4):331-334.

Prasad, A., L. Iverson, and A. Liaw. 2006. Newer classification and regression tree techniques: bagging and random forests for ecological prediction. Ecosystems 9(2):181-199.

R Development Core Team. 2015. R: A language and environment for statistical computing, Version 3.2.0. R Foundation for Statistical Computing, Vienna, Austria.

Raup, D. M. 1966. Geometric analysis of shell coiling: general problems. Journal of Paleontology 40:1178-1190.

Richards, R. P. 1972. Autecology of Richmondian brachiopods (Late Ordovician of Indiana and Ohio). Journal of Paleontology 46:386-405.

Ros, S., M. D. Renzi, S. E. Damborenea, and A. Marquez-Aliaga. 2012. Early Triassic–Early Jurassic bivalve diversity dynamics Pp. 1-19. Treatise Online. University of Kansas, Lawrence, KS.

Rudkin, D. M., G. A. Young, R. J. Elias, and E. P. Dobrzanski. 2003. The world's biggest trilobite—*Isotelus rex* new species from the Upper Ordovician of Northern Manitoba, Canada. Journal of Paleontology 77(1):99-112.

Sandy, M. R. 1996. Oldest record of peduncular attachment of brachiopods to crinoid stems, Upper Ordovician, Ohio, U.S.A. Journal of Paleontology 70:532-534.

Schumacher, G. A., and D. L. Shrake. 1997. Paleoecology and comparative taphonomy of an *Isotelus* (Trilobita) fossil lagerstätten from the Waynesville Formation (Upper Ordovician, Cincinnatian Series) of southwestern Ohio. Pp. 131-161. *In* C. E. Brett, and G. C. Baird, eds. Paleontological Events: Stratigraphic, Ecological, and Evolutionary Implications. Columbia University Press, New York.

Seilacher, A. 1970. Arbeitskonzept zur Konstruktions-Morphologie. Lethaia 3:393-396.

Sepkoski, J. J., Jr. 1981. A factor analytic description of the Phanerozoic marine fossil record. Paleobiology 7(1):36-53.

Slater, G. J., L. J. Harmon, D. Wegmann, P. Joyce, L. J. Revell, and M. E. Alfaro. 2012. Fitting models of continuous trait evolution to incompletely sampled comparative data using approximate Bayesian computation. Evolution 66(3):752-762.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8(1):25.

Sullivan, M., M. Jones, D. Lee, S. Marsden, A. Fielding, and E. Young. 2006. A comparison of predictive methods in extinction risk studies: contrasts and decision trees. Biodiversity & Conservation 15(6):1977-1991.

Thomas, R. D. K., and W. E. Reif. 1993. The skeleton space: a finite set of organic designs. Evolution 47:341-360.

Tomašových, A., and S. M. Kidwell. 2010. Predicting the effects of increasing temporal scale on species composition, diversity, and rank-abundance distributions. Paleobiology 36(4):672-695.

Van Valen, L. 1974. Multivariate structural statistics in natural history. Journal of Theoretical Biology 45:235-247.

Van Valkenburgh, B. 1988. Trophic diversity in past and present guilds of large predatory mammals. Paleobiology 14:155-173.

Van Valkenburgh, B., and R. E. Molnar. 2002. Dinosaurian and mammalian predators compared. Paleobiology 28:527-543.

Vermeij, G. J. 1987. Evolution and Escalation: An Ecological History of Life. Princeton University Press.

---. 2011. The energetics of modernization: The last one hundred million years of biotic evolution. Paleontological Research 15(2):54-61.

---. 2013. On escalation. Annual Review of Earth and Planetary Sciences 41(1):1-19.

Villéger, S., N. W. H. Mason, and D. Mouillot. 2008. New multidimensional functional diversity indices for a multifaceted framework in functional ecology. Ecology 89(8):2290-2301.

Villéger, S., J. R. Miranda, D. F. Hernández, and D. Mouillot. 2010. Contrasting changes in taxonomic vs. functional diversity of tropical fish communities after habitat degradation. Ecological Applications 20(6):1512-1522.

Villéger, S., P. M. Novack-Gottshall, and D. Mouillot. 2011. The multidimensionality of the niche reveals functional diversity changes in benthic marine biotas across geological time. Ecology Letters 14(6):561-568.

Villier, L., and G. J. Eble. 2009. Assessing the robustness of disparity estimates: the impact of morphometric scheme, temporal scale, and taxonomic level in spatangoid echinoids.

Vogt, R. J., P. R. Peres-Neto, and B. E. Beisner. 2013. Using functional traits to investigate the determinants of crustacean zooplankton community structure. Oikos 122(12):1700-1709.

Wills, M. A. 2001. Morphological disparity: a primer. Pp. 55-143. *In* J. M. Adrain, G. D. Edgecombe, and B. S. Lieberman, eds. Fossils, phylogeny, and form: an analytical approach. Kluwer Academic/Plenum Publishers, New York.

Xiao, S., and M. Laflamme. 2009. On the eve of animal radiation: phylogeny, ecology and evolution of the Ediacara biota. Trends in Ecology & Evolution 24(1):31-40.

Zhang, Z., M. Augustin, and J. L. Payne. 2015. Phanerozoic trends in brachiopod body size from synoptic data. Paleobiology 41(03):491-501.

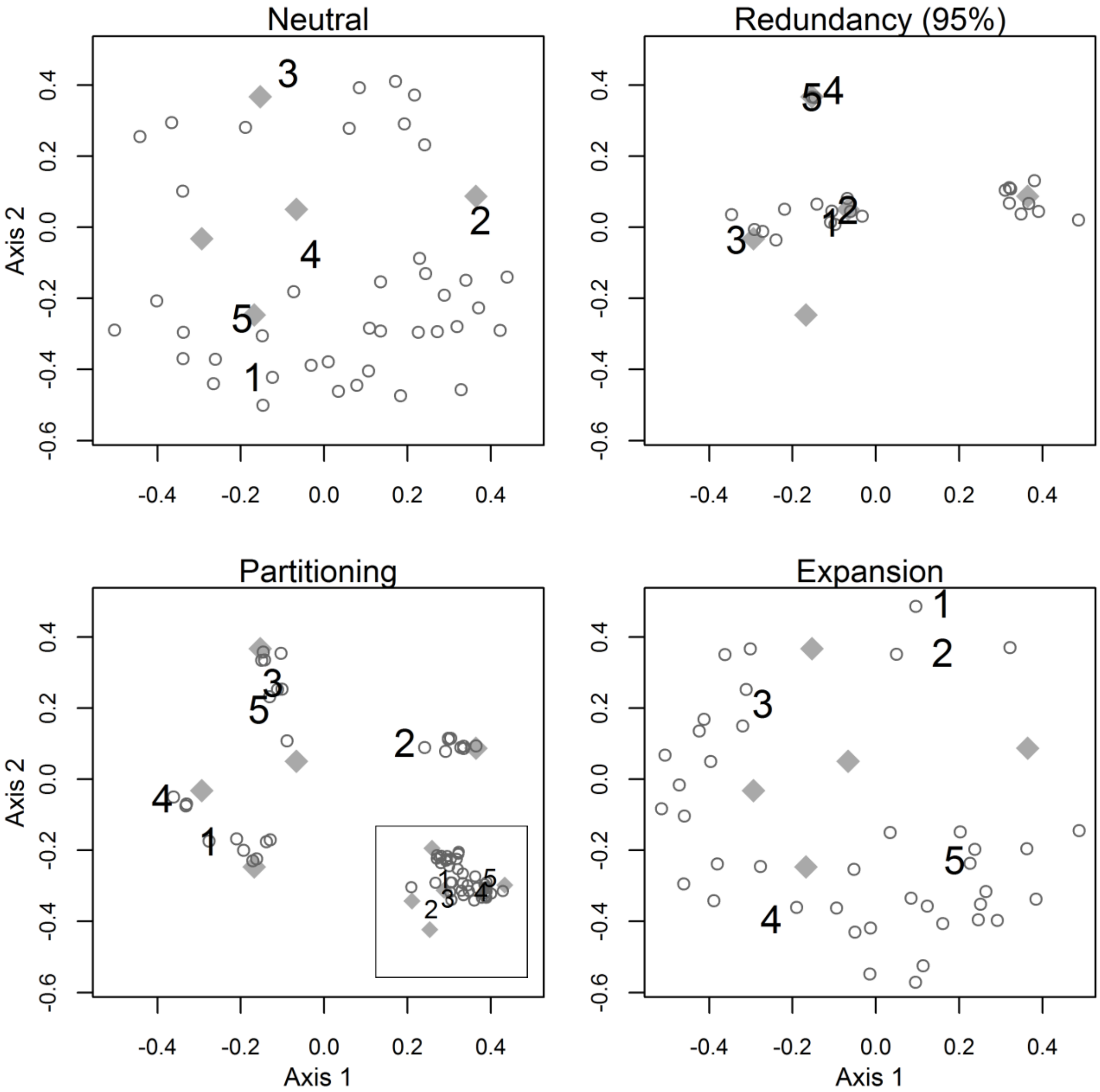


Figure 1. Typical examples of simulated 50-species assemblages produced by the four model rules. Assemblages are plotted on common non-metric multi-dimensional scaling ordination space of functional traits to allow comparative evaluation of model behavior. Five gray diamonds represent common "seed" species whose life habits were assigned stochastically using an 18-character (functional trait space) ecospace framework (modified from Novack-Gottshall 2007), imposing a realistic constraint that each life habit could have at most two character states within a given character. Numbers illustrate the addition of five species to each assemblage (after seed species), with remaining 40 species as hollow circles. All model rules, except redundancy, were

enacted at 100% rule-following for each simulation; redundancy rules were weakened such that all successive species have habits 95% similar to pre-existing ones; at 100% enactment, later life habits are limited to the “seed” species. In the neutral model, functional traits of all species are chosen independently at random, and the entire ecospace becomes inhabited through passive processes. In the redundancy model, new species have life habits similar to pre-existing ones, producing an ecospace with distinct clusters. In the partitioning model, new species inhabit life habits intermediate to pre-existing ones. This model can be enacted in a strict form (larger image) in which new species are restricted to gradients between preexisting species (typically leaving the center empty) and a relaxed form (inset) in which new species progressively fill in empty regions of the space originally defined by the seed species. In the expansion model, new species progressively inhabit novel life habits, producing an ecospace that expands its breadth over the simulation, while leaving the original region uninhabited. Figure 2B shows the average dynamics of eight structural statistics when such simulations were repeated 1000 times.

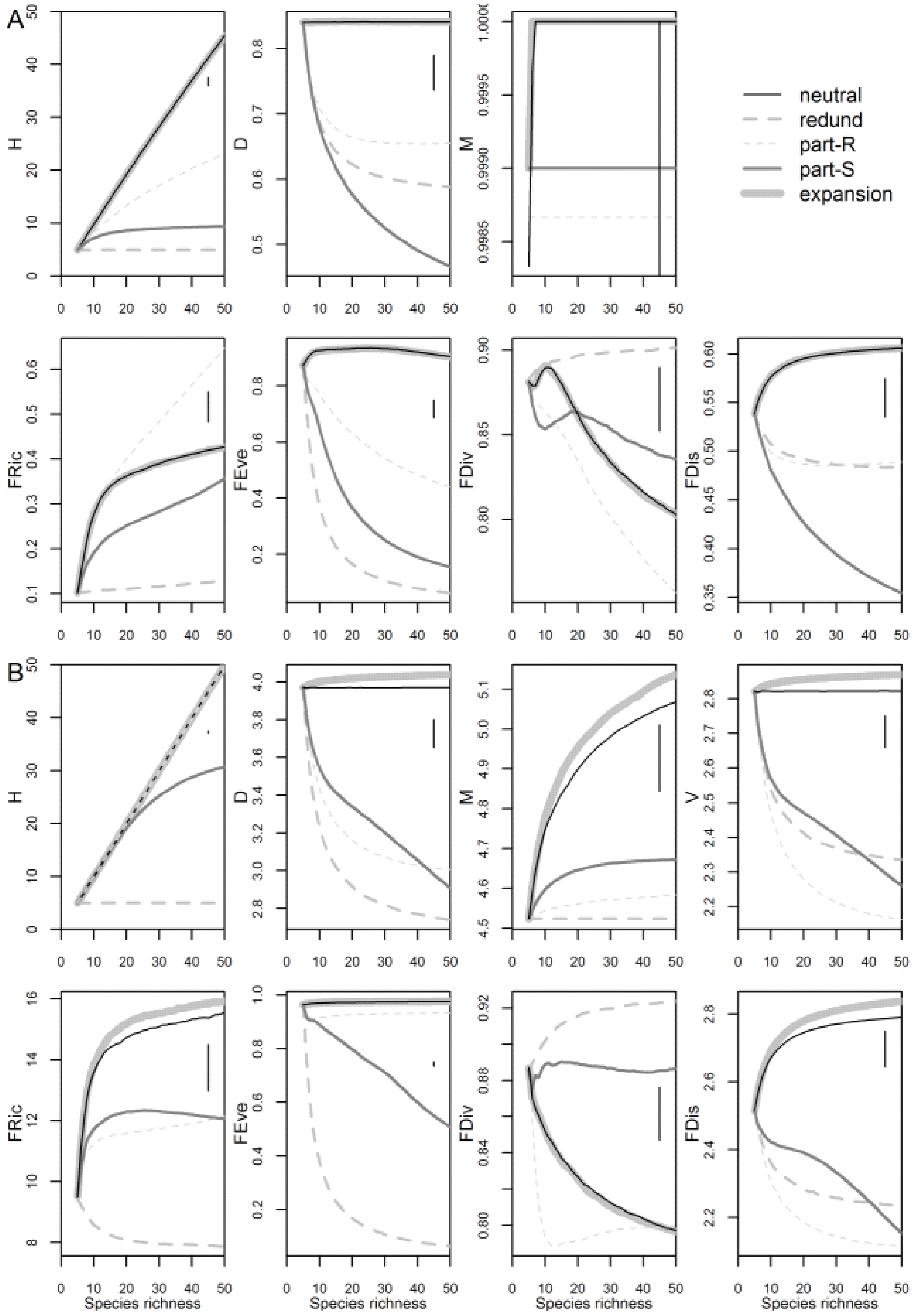


Figure 2. Model dynamics for eight functional-diversity statistics for the ecospace frameworks of (A) Bush and Bambach (Bambach et al. 2007, Bush et al. 2007, Bush and Bambach 2011, Bush et al. 2011) and (B) Novack-Gottshall (2007). In both cases, all models are enacted at 100% rule-following with five seed species. The second case (part B) also stipulates that each life habit can have at most two character states within a given life-habit character (see text for explanation of

'constraint' parameter). Dynamics are reported as a function of increasing species richness, up to 50 species (i.e., there is a common abscissa in all graphs); the ordinate has been truncated to focus on the trend lines for each statistic. Statistics include life-habit richness (H), mean distance (D), maximum range (M), total variance (V), functional richness (FRic), evenness (FEve), divergence (FDiv), and dispersion (FDis); see text for description of each statistic. Trend lines are the average of 1000 simulations. Vertical bars near top right of each graph illustrate the average standard deviation around each set of trends; the standard deviation for mean distance in the Bush and Bambach framework equals 0.0117 and extends beyond the borders of the graph. PartS and PartR represent the strict and relaxed versions of the partitioning model. The variance trend in part A is omitted because the characters in the Bush and Bambach framework are all factorial, preventing its calculation. The uppermost H trend line in part B is an overlap between the neutral, relaxed partitioning, and expansion trend lines. Despite their rather different structures, the model dynamics for each ecospace framework are quite similar. Notable differences include total overlap between neutral and expansion models in the Bush and Bambach framework, and generally smaller error margins for the Novack-Gottshall framework, which allows more powerful model selection using classification-tree methods (89% of validation models classified correctly versus 73% with the Bush and Bambach framework).

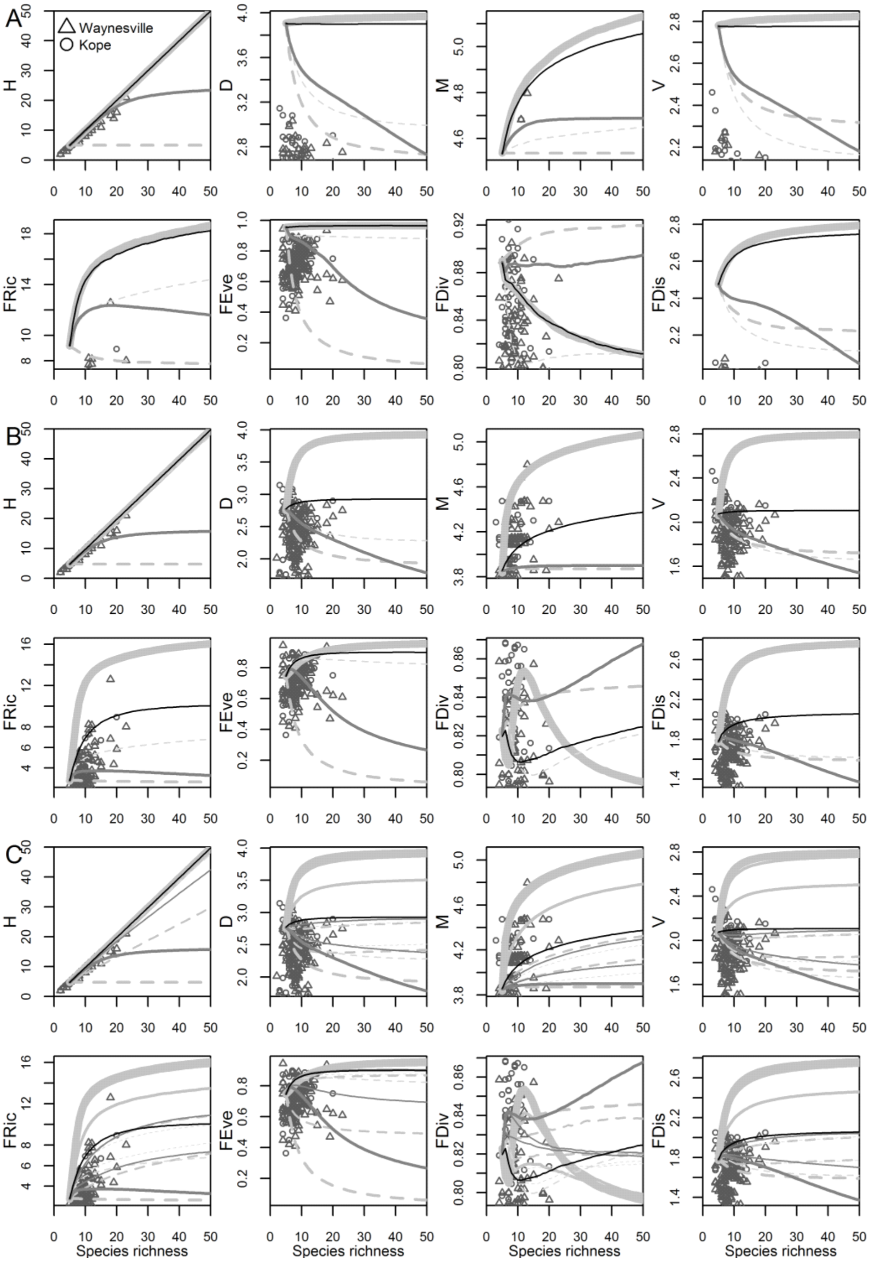


Figure 3. Comparison of Late Ordovician samples to dynamics of different simulation implementations: (A) 100% rule-following implementation with one-state constraint, (B) 100% rule-following implementation with two-state constraint and empirical weighting of states, (C) 100%, 90%, and 50% rule-following with two-state constraint and empirical weighting. The legend and graphical interpretation is the same as Figure 2, and variability around the mean trend lines are of similar magnitude. Statistics for Late Ordovician (type Cincinnatian) Kope and

Waynesville Formation samples are represented by circles and triangles, respectively. In part A, simulations were enacted at 100% rule-following and each life habit was constrained to have just one character state within a given life-habit character; for example, a taxon could be either sexual or asexual, but not both (hermaphroditic). The dynamics are similar to those for the two-state implementation in Figure 2B, although values are slightly diminished, especially for the partitioning models. In all cases, the empirical sample statistics lay well below the mean trend lines, implying a poor fit between these simulations and reality. In part B, the simulations included the two-state constraint, but were modified so that the combined Kope and Waynesville species pool was used to weight how seed species were chosen (i.e., they were chosen at random from the species pool) and the inhabitation of character states of all successive species were weighted by their frequency of occurrence in the species pool. The fit between the empirical samples and the model dynamics is much improved (that is, there is substantially more overlap). In part C, eight additional "weaker" simulations are added, in which the model rules are followed on average 90% and 50% of the time. The average trend lines for these weaker simulations are slightly thinner (but with the same colors and line types) than the 100% implementations to make it easier to distinguish them (although the 90% expansion model trend mostly overlays the 100% implementation). A 0% simulation is always present, the neutral model, in which no model rules are followed. The empirical samples frequently overlay weaker versions of the models, and all models coalesce toward the neutral model as they become weaker and weaker.

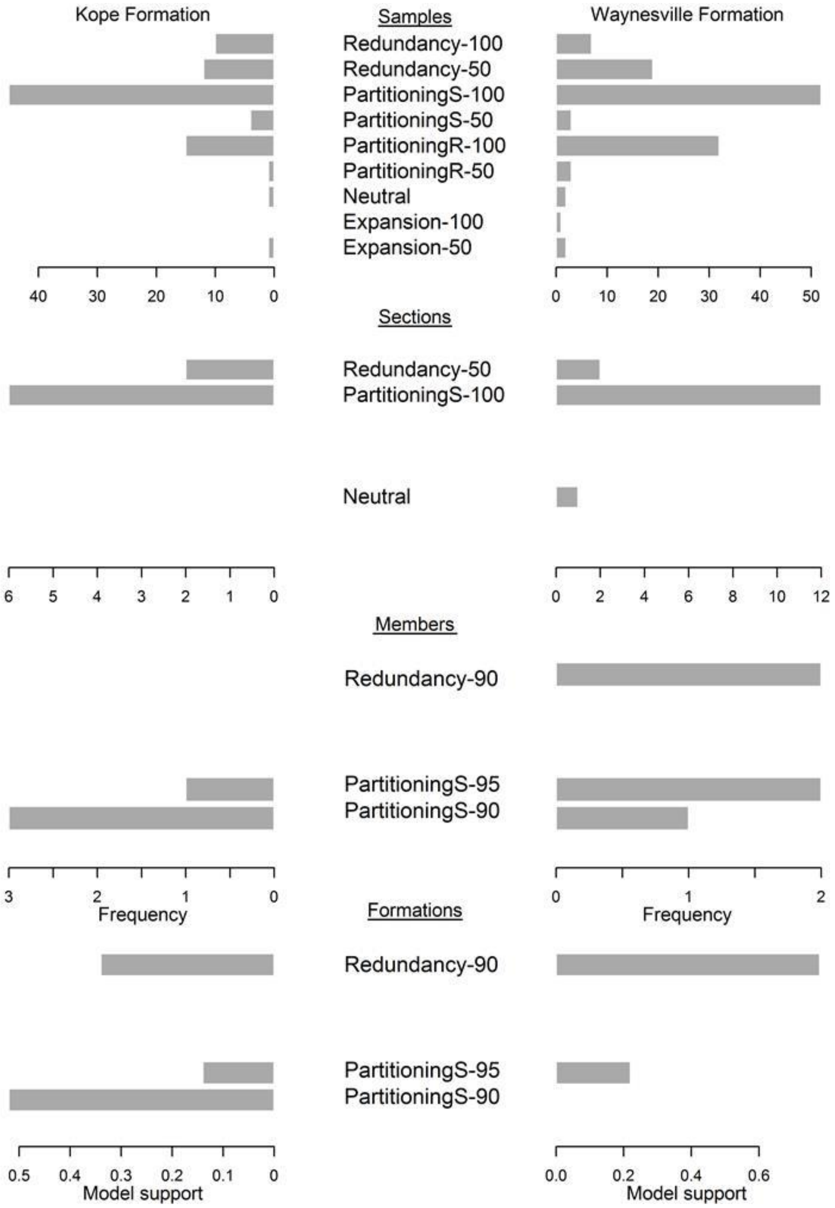


Figure 4. Relative model support for Kope (left column) and Waynesville (right column) samples at scales of individual samples, aggregated stratigraphic sections, members, and entire formation species pool. Model support for samples and sections was calculated using a classification tree trained on the 100% and 50% simulations used in Figure 3C; model support for members and formations used the classification tree trained on 50%, 75%, 90%, 95%, and 100% simulations, which is a more powerful classifier at larger sample sizes. Models are only listed (middle column) when they have support in a particular scale. Support for all but the formation-scale is

the number of samples/sections/members that were classified for each model. Support for the formations (bottom row) is the proportion of votes for each model. The aggregate Kope Formation is overwhelmingly (0.52) classified as the 90% strict partitioning model, although there is also substantial support (0.34) for the 90% redundancy model. The aggregate Waynesville Formation has overwhelming support (0.78) for the 90% redundancy model, with less but substantial support (0.22) for the 95% strict partitioning model.

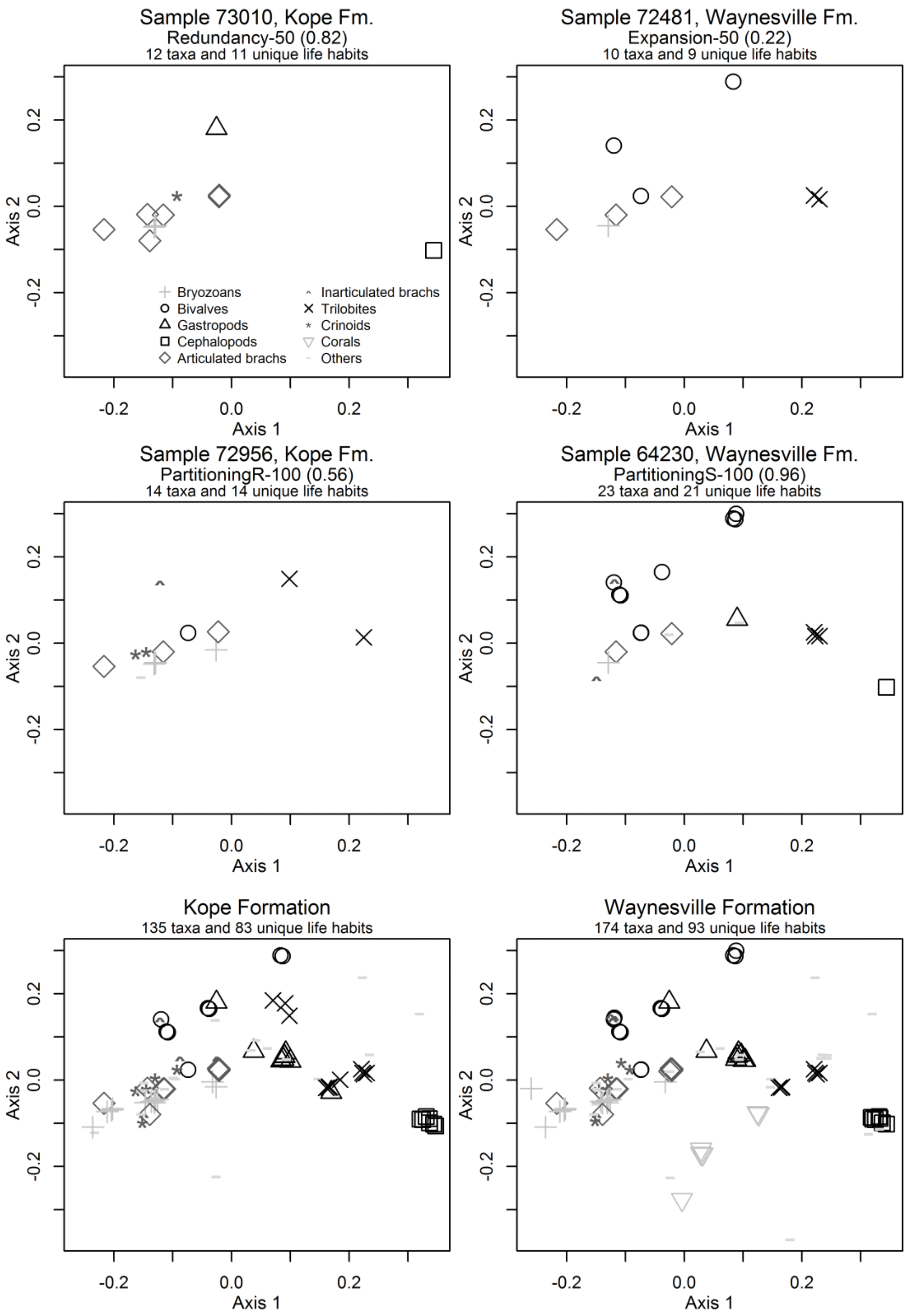
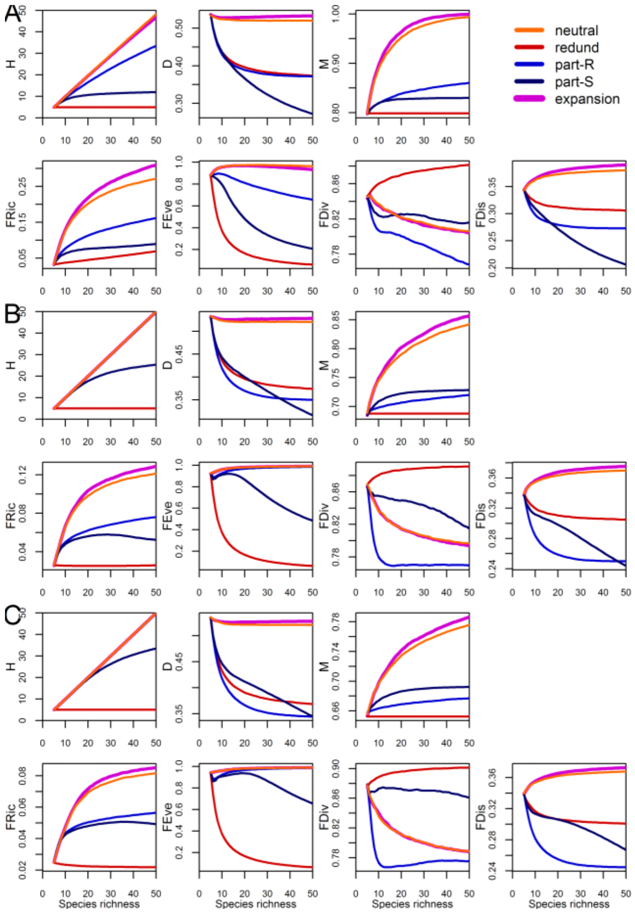
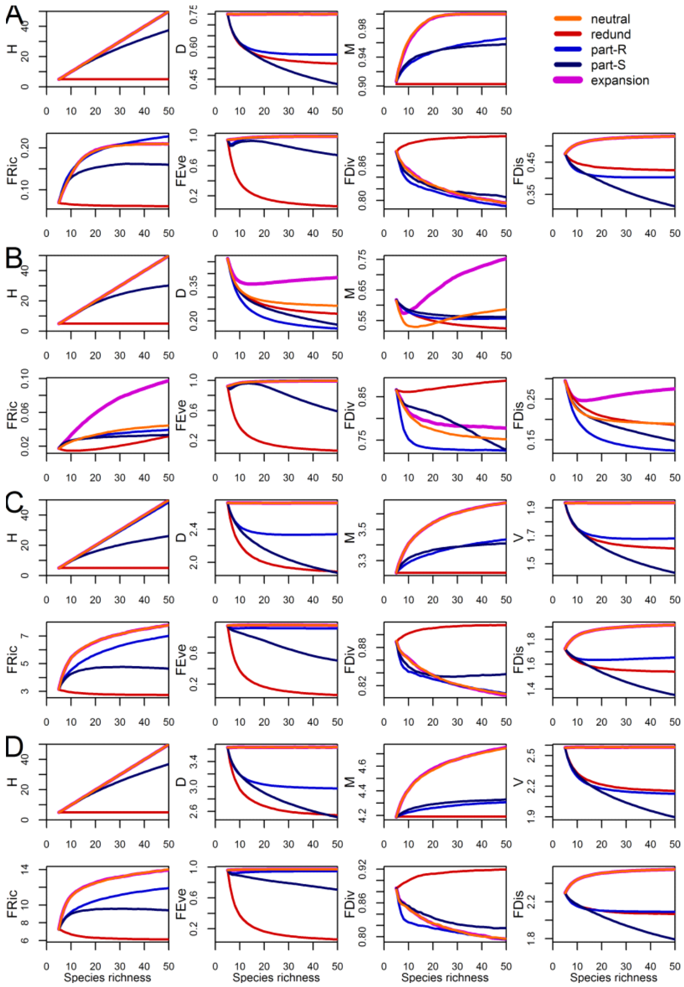


Figure 5. Ordinations of four representative Kope and Waynesville samples best fit by various models, and the corresponding ordination of each formation species pool. Ordinations were conducted as two-dimensional non-metric multidimensional scaling on a common Kope/Waynesville species pool to allow comparison across graphs. Text above each graph notes which model was best fit (i.e., had the most votes in the classification tree) for each sample, as

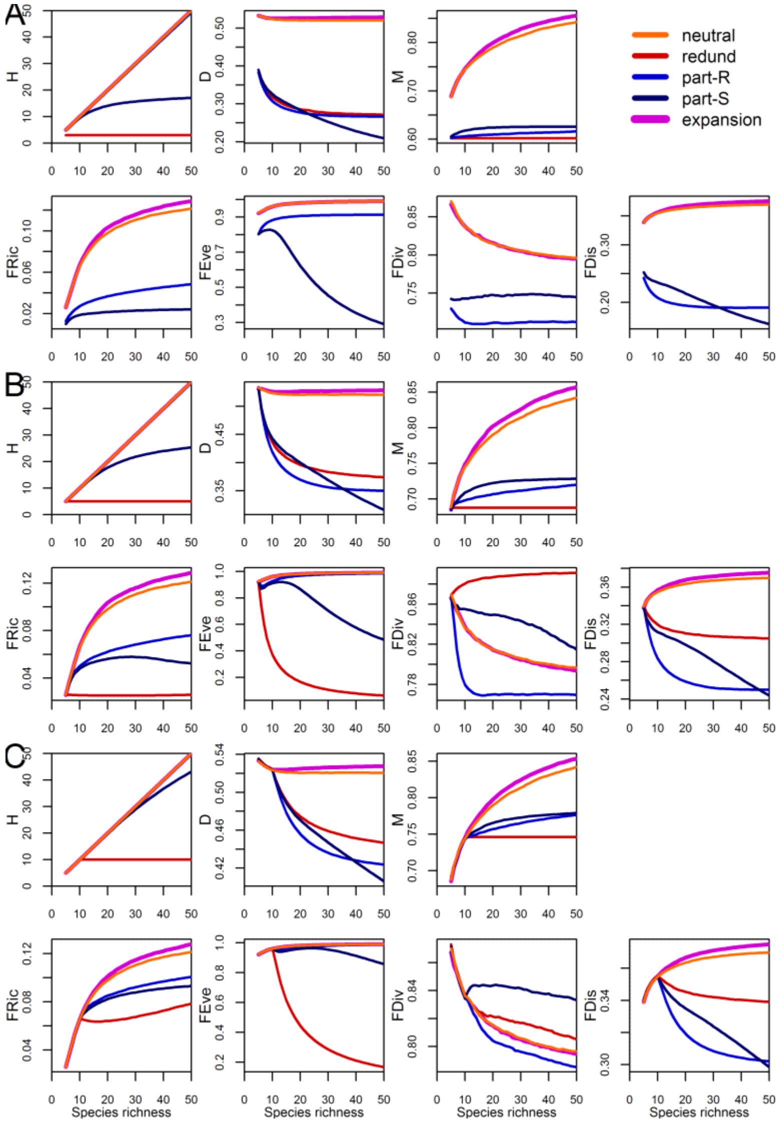
well as the relative support for that "vote-winning" model. Also listed are the Paleobiology Database sample collection number, its taxonomic richness and number of unique life habits (H), and symbols representing major taxonomic groups. Note that jittering (moving overlapping points by tiny amounts) was not used in this figure; points that overlap incompletely represent distinct life habits. Aside from at the formation-scale, there is relatively little absolute redundancy among life habits.



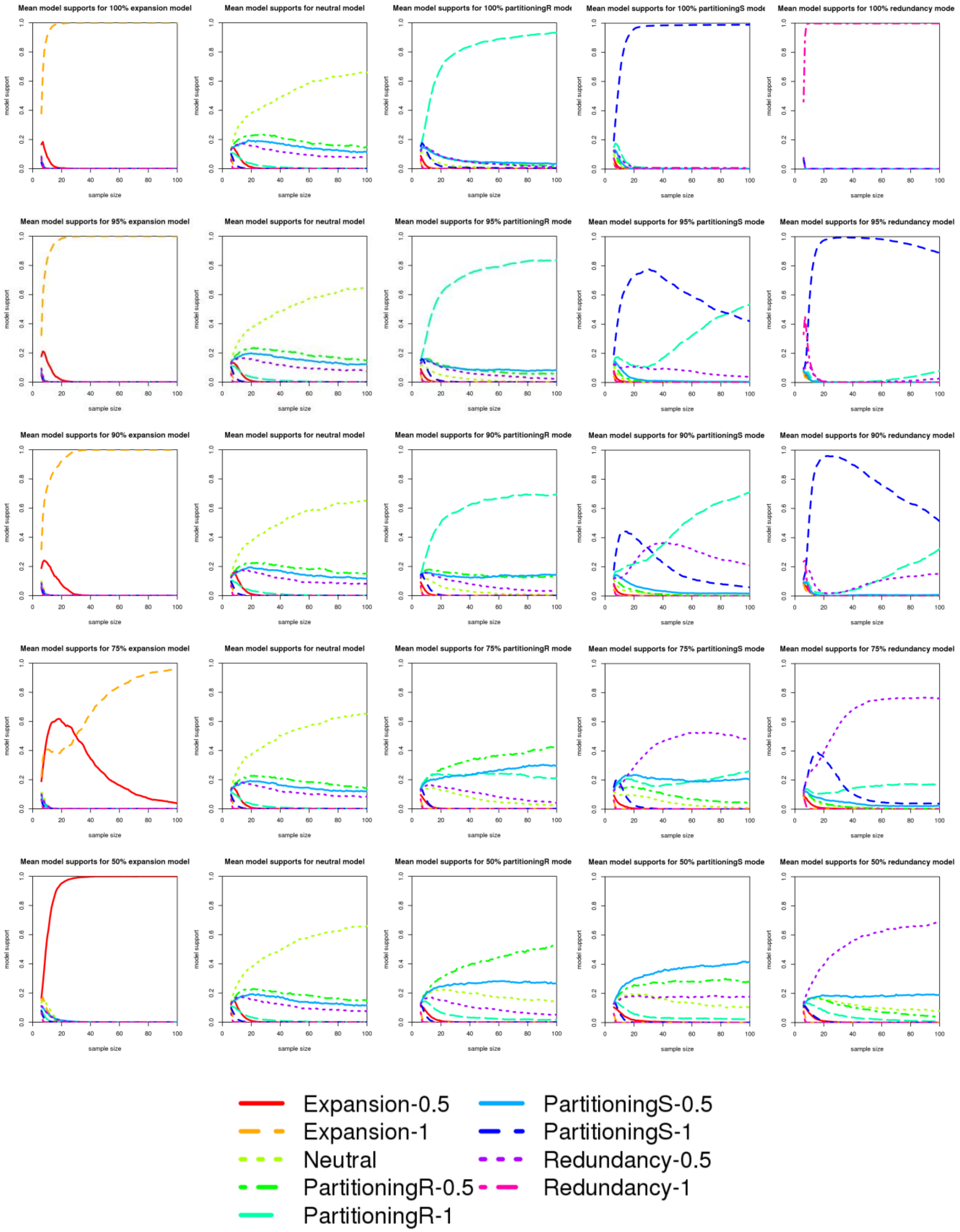
Supplementary Figure 6. Comparing statistical dynamics for different ecospace framework structures: varying number of characters, (A) 5 characters, (B) 15 characters, and (C) 25 characters. Each framework had mixed character types, in identical proportions (40% binary, 20% three-state factor, 20% five-state factor, and, 20% five-state ordered numeric character types). 5 "seed" species were chosen at random to begin each simulation. Other simulation details and graphical interpretation are the same as is Figure 2. Trends in total variance were excluded because the inclusion of factors prevented their calculation. The dynamics are generally similar, although larger frameworks allow modestly more powerful model selection using classification-tree methods (83%, 85%, and 86% of training models, respectively, classified correctly using classification-tree methods). See Supplementary Appendix 2 for additional details.



Supplementary Figure 7. Comparing statistical dynamics for different ecospace framework structures: varying character types, (A) factor, (B) ordered factor, (C) ordered numeric, and (D) binary. Each framework had 15 characters, four states per character (except for binary, which had two binary states per character), and five seed species. Trends in total variance were excluded in parts A and B because the inclusion of factors prevented their calculation. Other simulation details and graphical interpretation are the same as in Supplementary Figure 6. Dynamics are generally similar, but frameworks built with ordered factors performed substantially better (94% of trained models classified correctly) than the others (78% for unordered factors, 79% for ordered numerics, and 81% for binaries). See Supplementary Appendix 2 for additional details.



Supplementary Figure 8. Comparing statistical dynamics for different ecospace framework structures: varying number of "seed" species chosen at random at start of simulation, (A) 3 species, (B) 5 species, and (C) 10 species. Each framework had 15 mixed character types, such that part B is identical to Supplementary Figure 6B. Trends are only plotted starting at 5 species for comparative purposes, which explains idiosyncratic behaviors at low sample sizes (i.e., missing trend lines in part A and overlapping models in part C). Other simulation details and graphical interpretation are the same as is Supplementary Figure 6. Note that the functional-diversity statistics could not be calculated for the redundancy model in part A because their calculation requires a minimum of four unique life habits; however, all statistics (except V) were included as potential predictor variables in the classification tree algorithm. Dynamics are generally similar across models, but simulations with fewer seed species provide the most powerful model selection using classification-tree methods (85% of models classified correctly for both 3-seed and 5-seed simulations). Starting with larger numbers of seed species impedes enacting distinct model rules, and results in 75% of models classified correctly. See Supplementary Appendix 2 for additional details.


Mean model supports for 100% expansion model
Mean model supports for neutral model
Mean model supports for 100% partitioningR mode
Mean model supports for 100% partitioningS mode
Mean model supports for 100% redundancy mode
Mean model supports for 95% expansion model
Mean model supports for neutral model
Mean model supports for 95% partitioningR mode
Mean model supports for 95% partitioningS mode
Mean model supports for 95% redundancy model
Mean model supports for 90% expansion model
Mean model supports for neutral model
Mean model supports for 90% partitioningR mode
Mean model supports for 90% partitioningS mode
Mean model supports for 90% redundancy model
Mean model supports for 75% expansion model
Mean model supports for neutral model
Mean model supports for 75% partitioningR mode
Mean model supports for 75% partitioningS mode
Mean model supports for 75% redundancy model
Mean model supports for 50% expansion model
Mean model supports for neutral model
Mean model supports for 50% partitioningR mode
Mean model supports for 50% partitioningS mode
Mean model supports for 50% redundancy model
model support
0.0
0.2
0.4
0.6
0.8
1.0
sample size
0
20
40
60
80
100
Expansion-0.5
Expansion-1
Neutral
PartitioningR-0.5
PartitioningR-1
PartitioningS-0.5
PartitioningS-1
Redundancy-0.5
Redundancy-1

Supplementary Figure 9. Performance of classification tree used on validation samples as a function of sample size. This is the tree used to classify empirical samples in analyses for Figures 4 and 5. The classification tree was trained on 5,265,500 simulated samples spanning nine models (identified in figure legend, with mean trends for 100% and 50% training data sets visualized in Figure 3C): the neutral model; the four models that implement the redundancy, two versions of partitioning, and expansion rules; and four "weakened" versions of these four models, in which rules were followed at random 50% of the time. Numbers in model names are abbreviated, with "-1" referring to the 100%-rule-following implementation and "-0.50" referring to the 50% implementation. Performance of the tree was evaluated against 2,375,000 validation samples (i.e., new samples not used when training the tree), that included simulated data not only from the 50% and 100% rule-following simulations, but also those from 75%, 90%, and 95% implementations. In other words, there were 1000 simulations per model-sample size * 95 sample sizes [5–100 species per sample] * 25 models [with independent samples from the neutral model included in each set]).

Each row corresponds to a validation data set, with 100% at the top, 95% below that, continuing to 50% at the bottom. Each column corresponds to a model, with expansion at left, followed rightward by neutral, relaxed implementation partitioning, strict partitioning, and redundancy. Trend lines illustrate the proportion of votes for each model in the classification tree, as a function of sample size. For example, the top-left graph (illustrating how "known" 100%-expansion validation samples were classified), shows that nearly all such validation samples were correctly classified as simulated from the 100%-expansion model, even at low sample sizes, and with essentially all such samples classified perfectly at sample sizes as low as 15. That nearly 40% of samples with just 6 species are correctly classified is especially promising because this is the smallest sample size when expansion rules were implemented (recall that that five species were assigned at random to seed the simulation). Similar performance occurs for most of the 100% implementations. Classification performance for the 50% implementations is less accurate, although the correct model always receives majority support, nearly always at or above support values of 0.40.

Patterns of support for the intermediate (and novel) validation samples, those produced by implementing model rules 75%, 90%, and 95% of the time, demonstrate that the use of classification trees produces powerful and intuitively understandable performance as a model-selection method. In general, model support at the 95% level mimics closely that at the 100% level, with somewhat weaker (but still majority) support for the more similar model. This makes intuitive sense; the 95% samples are produced by the 100% expansion model 95% of the simulation, on average. As one looks down a column, relative support transitions across classified models as expected: support for the 100% model diminishes and gives way to support for the 50% model. The switch from one model to another occurs generally at the half-way point (75%), and this point is also where most confusion in model classification occurs. This pattern further suggests that support for different weights of the same model (e.g., the 75% expansion samples are classified as both 100% and 50% expansion models), can provide evidence for a weakened model (i.e., intermediate between the 100% and 50% implementations), especially for smaller sample sizes (below 30–40 taxa). This mutual support for alternative models is lessened at larger sample sizes, where the simulations used to train the classification tree generally have reduced variability around the asymptotes at these sample sizes, and which leads to increased classification confusion as the tree is more powerful at rejecting the trained implementations. Trends down the "neutral" column are largely identical, reflecting the fact that the neutral model is always the absence of any assembly rule (either always 100% neutral or 0% an alternative model), and it is reassuring that the independent "neutral" simulations are consistently classified across iterations.

Other patterns are worth noting when evaluating classification trees as a method of model selection. Across sample sizes, the correct model consistently achieves minimum support levels greater than 0.2, from which it is possible to conclude that this support value be recommended as evidence of "substantial" support for the correct model, and values above 0.40 tend to be associated with "unambiguously strong" support for the correct model. These recommended support levels are also found for more complicated trees (i.e., those trained on 50%, 75%, 90%, 95%, and 100%-strength implementations), especially for sample sizes greater than 40 species. (It is worth noting that these "critical" 0.2 and 0.4 support values are the result of validation tests of these models, and they may not be generalizable to other implementations; however, their consistency when evaluated across 9-, 13-, and 21-training set implementations is promising.)

Model support for simulations created with weakened (<90% implementation) models tend to be confused with other models (i.e., they have greater numbers of misclassifications). Weakened implementations of the partitioning and redundancy models (redundancy-0.5, partitioningR-0.5, and partitioningS-0.5), in particular, get confused for one another; see Supplementary Appendix 2 for additional confirmation of this claim. This is not a weakness of the classification tree but a confirmation of its power; in their weakened form, these three models are truly dynamically similar. The classification tree is still sufficiently powerful to distinguish these models at least 40% of the time, even at small sample sizes.

Supplementary Appendix 1. Example of how life-habit character states were inferred and coded. See Supplementary Table 1 for raw character-state codings.

Mature asaphid trilobite *Isotelus maximus* specimens have a shell volume of ca. 270 cm3, using caliper measurements of major axes (Feldmann 1996) and volumetric allometries from Novack-Gottshall (2008b). It is coded as a carnivorous predator given its large, forked hypostome with a rasp-like texture, likely supplemented with limbs to dispatch prey (Fortey and Owens 1999, Hegna 2010). *Rusophycus* trace fossils further support this foraging habit, suggesting evidence of raptorial attacks on *Palaeophycus*-producing shallow-burrowing worms using large limbs (Brandt et al. 1995, Rudkin et al. 2003, English and Babcock 2007). Raised holochroal eyes well suited to deep water (Fordyce and Cronin 1993) and absence of other trace fossils suggests a hunting strategy where the trilobite swam habitually above the seafloor searching for shallow-infaunal prey at some distance. Swimming height is less well constrained, but was most likely nektobenthic given flattened exoskeleton profile and previously discussed trace fossil evidence. Given there are no asexual arthropods, it likely reproduced gonochoristically.

Supplementary Table 1. 18 characters are in bold, with 37 character states numbered. First six characters are coded as ordered numeric factors and remainder are coded as binary states. (Such binary codings are used instead of discrete factors to allow for logically possible state combinations. For example, a hermaphrodite would be coded as reproductive code [1,1] for being both asexual and sexual, and eating food from on top of and within the seafloor would be coded as primary feeding microhabitat code [1,1]. See Novack-Gottshall (2007) for definitions and discussion of characters and states.

1) **Skeletal body volume**: $\boxed{1}$ (100–1000 cm3)
2) **Mobility**: $\boxed{1}$ (Habitually mobile)
3) **Primary stratification**: $\boxed{0.5}$ (1-10 cm from seafloor)
4) **Immediate stratification**: $\boxed{0.5}$ (1-10 cm from seafloor)
5) **Primary food stratification**: $\boxed{0.25}$ (0.1-1.0 cm from seafloor)
6) **Immediate food stratification**: $\boxed{0.75}$ (10-100 cm from seafloor)

Reproduction:
7. Asexual: $\boxed{0}$ (No)
8. Sexual: $\boxed{1}$ (Yes)

Substrate/medium composition:
9. Biotic: $\boxed{0}$ (No)
10. Lithic: $\boxed{1}$ (Yes)
11. Fluidic: $\boxed{0}$ (No)

Substrate consistency:
12. Hard: $\boxed{0}$ (No)
13. Soft: $\boxed{1}$ (Yes)
14. Insubstantial: $\boxed{0}$ (No)

Substrate relationship:
15. Attached: $\boxed{0}$ (No)
16. Free-living: $\boxed{1}$ (Yes)

Primary microhabitat:
17. Above primary substrate: $\boxed{1}$ (Yes)
18. Within primary substrate: $\boxed{0}$ (No)

Immediate microhabitat:
19. Above immediate substrate: $\boxed{1}$ (Yes)
20. Within immediate substrate: $\boxed{0}$ (No)

Support:
21. Supported: $\boxed{0}$ (No)
22. Self-supported: $\boxed{1}$ (Yes)

Primary feeding microhabitat:
23. Above primary substrate: $\boxed{1}$ (Yes)
24. Within primary substrate: $\boxed{1}$ (Yes)

Immediate feeding microhabitat:
25. Above immediate substrate: $\boxed{1}$ (Yes)
26. Within immediate substrate: $\boxed{1}$ (Yes)

Diet:
27. Autotroph: $\boxed{0}$ (No)
28. Microbivore: $\boxed{0}$ (No)
29. Carnivore: $\boxed{1}$ (Yes)

Physical condition of food:
30. Incorporeal feeder: $\boxed{0}$ (No)
31. Particle feeder: $\boxed{0}$ (No)
32. Bulk feeder: $\boxed{1}$ (Yes)

Feeding habit:
33. Ambient feeder: $\boxed{0}$ (No)
34. Filter feeder: $\boxed{0}$ (No)
35. Attachment feeder: $\boxed{0}$ (No)
36. Mass feeder: $\boxed{0}$ (No)
37. Raptorial feeder: $\boxed{1}$ (Yes)

Supplementary Appendix 2. Technical details on classification tree methods, and confusion-matrix results compared across various simulations.

Classification trees are a powerful and flexible tool for classifying complex datasets that have complex, often non-linear and localized relationships among variables (Breiman et al. 1984, Cutler et al. 2007) and are being increasingly used by ecologists (De'ath and Fabricius 2000, De'ath 2002, Sullivan et al. 2006, Cutler et al. 2007, Boyer 2010, Durst and Roth 2012) and paleontologists (Finnegan et al. 2012, Finnegan et al. 2015). Except where noted in text, classification-tree analyses were run in the following manner: classification trees were trained on 1000-replicate data sets (per model, each implemented 100% rule-following) with sample sizes of 6 to 50 species. (In other words, the training set for each tree included 225,000 samples, equal to 1000 simulations * 45 sample sizes [6–50 species per sample], or 45,000 simulated samples for each model * 5 models.) Random-forest classification was conducted in the R randomForest package (Breiman 2006), building ensembles of 50 bootstrapped replicate trees ('ntrees=50')—in which some samples and variables are excluded at random during tree building to enhance statistical power and reduce overfitting—to terminal node sizes of 1 ('nodesize=1'). These algorithmic parameters were justified by sensitivity analyses.

The basic structure was to predict (classify) each model as a function of species richness and eight functional diversity/disparity statistics:

$$\text{Model} \sim S + H + D + M + V + FRic + FEve + FDiv + FDis,$$

with variables defined in text. In the following confusion matrices, the total variance (V) statistic was excluded for consistency, as this statistic could not be calculated for ecospace -framework simulations with factors present. (Although the classification tree considers V a relatively important statistic for model selection, its high correlation with equally important mean distance

(D) yields relatively little effect, less than 1%, when excluding it.) The four functional diversity statistics were included in all analyses, even in the one case (3 seed species, Supplementary Fig. 8) where they could not be calculated (excluding them had a negligible effect on model classification).

Classification rates and confusion matrices provide measures of the ability of classification trees to classify model trends. Validation tests were not run for most of the following trials, so classification rates may be elevated slightly (<10%) because of potential model overfitting, but relative rankings for comparative purposes should remain accurate. Validation tests (with data sets of the same size) were run on several simulations (e.g., those in Fig. 2 and Fig. 3C), with validation classification rates reported in brackets.

The first column in the confusion matrices is the known ("true") model from which samples were simulated. The next five columns show the models the tree classified these samples. In an ideal classification tree (i.e., in which all models are predicted perfectly without error), for example, all "Expansion-1" samples (those samples simulated following the expansion model 100% of the time) would be classified into this same model. In other words, the diagonals in the confusion matrix show correctly classified models. The final column ("class.error") tallies the proportion of each known model that was incorrectly classified.

Comparing number of characters in ecospace frameworks (Supplementary Fig. 6 and Supplementary Fig. 7B)

5 mixed characters: 83.38% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           29920    13958              751              348            23       0.3351
Neutral               11626    32145              804              402            23       0.2857
PartitioningR-1         767      899            40073             3147           114       0.1095
PartitioningS-1         337      454             3537            40493           176       0.1001
Redundancy-1              1        1                7                9         44982       0.0004
```

15 mixed characters (and 5 seed species): 84.65% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           28839    15867              140              154             0       0.3591
Neutral               14364    30344              163              129             0       0.3257
PartitioningR-1         117      128            43404             1351             0       0.0355
PartitioningS-1          70       52             2001            42877             0       0.0472
Redundancy-1              0        0                0                0         45000       0.0000
```

25 mixed characters: 85.94% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           29658    15265               43               34             0       0.3409
Neutral               13955    30985               32               28             0       0.3114
PartitioningR-1          23       18            44045              913             1       0.0212
PartitioningS-1          10       15             1295            43678             2       0.0294
Redundancy-1              0        0                0                0         45000       0.0000
```

Comparing number of "seed" species used at start of simulations (Supplementary Fig. 8)

3 seed species: 84.53% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           29230    15744                6               20             0       0.3504
Neutral               14069    30911                7               13             0       0.3131
PartitioningR-1           4       10            41158             3815            13       0.0854
PartitioningS-1           7       16             1070            43903             4       0.0244
Redundancy-1              0        0                0                7         44993       0.0002
```

10 seed species: 74.68% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           26956    14223             1487             1277          1057       0.4010
Neutral               14034    27106             1520             1299          1041       0.3976
PartitioningR-1        1415     1442            37161             3882          1100       0.1742
PartitioningS-1        1152     1269             5414            36075          1090       0.1983
Redundancy-1           1022     1065             1069             1109         40735       0.0948
```

Comparing character types used in ecospace framework (Supplementary Fig. 7)

Binary character types: 81.08% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           26723    18042              187               48             0       0.4062
Neutral               17896    26847             , 198               59             0       0.4034
PartitioningR-1         167      137            42086             2610             0       0.0648
PartitioningS-1          29       23             3172            41774             2       0.0717
Redundancy-1              0        0                0                0         45000       0.0000
```

Factor character types: 78.15% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           24391   , 20520               72               17             0       0.4580
Neutral              , 20608    24318               56               18             0       0.4596
PartitioningR-1          35       24            41589             3351             1       0.0758
PartitioningS-1           2       10             4435            40551             2       0.0989
Redundancy-1              0        0                0                1         44999       0.0000
```

Ordered factor character types: 94.06% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           41526     2816              306              352             0       0.0772
Neutral                2501    41301              718              480             0       0.0822
PartitioningR-1         357      911            42126             1606             0       0.0639
PartitioningS-1         273      475             2569            41683             0       0.0737
Redundancy-1              0        0                0                0         45000       0.0000
```

Ordered numeric character types: 79.23% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           25708    18686              432              173             1       0.4287
Neutral               18913    25492              422              173             0       0.4335
PartitioningR-1         291      375            41042             3284             8       0.0880
PartitioningS-1          70       94             3790            41024            22       0.0884
Redundancy-1              0        0                0                0         45000       0.0000
```

<u>Other ecospace frameworks:</u>

Bush and Bambach ecospace framework (Fig. 2A): 74.85% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           24198   , 20127              448              182            45       0.4623
Neutral             , 20157    24170              452              167            54       0.4629
PartitioningR-1         357      267            37250             6702           424       0.1722
PartitioningS-1         170      140             5278            37939          1473       0.1569
Redundancy-1              0        4               24              121         44851       0.0033
```

[Same, but using a validation data set: 72.47% for validation data set]

```
     True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
     Expansion-1           21102    23300              376              169            53      0.5311
     Neutral             , 20814    23493              471              162            60      0.4779
     PartitioningR-1         425      351            36388             7363           473      0.1914
     PartitioningS-1         266      242             5206            37361        , 1925      0.1698
     Redundancy-1              3        0               32              244         44721      0.0062
```

Manuscript ecospace framework, limiting binary characters to two mutually "present" states ['constraint=2'] (Fig. 2B): 90.79% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           34808    10091               81             , 20             0       0.2265
Neutral                8095    36758              113               34             0       0.1832
PartitioningR-1          53       70            44060              817             0       0.0209
PartitioningS-1          10       23             1062            43905             0       0.0243
Redundancy-1              0        0                0                0         45000       0.0000
```

[Same, but using a validation data set: 89.10% for validation data set]

```
     True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
     Expansion-1           32566    12335               77               22             0      0.2763
     Neutral               10004    34863               96               37             0      0.2253
     PartitioningR-1          54       69            44135              742             0      0.0192
     PartitioningS-1          14       29             1054            43903             0      0.0244
     Redundancy-1              0        0                0                0         45000      0.0000
```

Manuscript ecospace framework, limiting binary characters to single "present" state ['constraint=1'] (Fig. 3A): 91.08% correct

```
True model      Expansion-1   Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           34898      9891              135               76             0       0.2245
Neutral                7040     37686            , 191               83             0       0.1625
PartitioningR-1         121       140            43638             1101             0       0.0303
PartitioningS-1          47        50             1194            43709             0       0.0287
Redundancy-1              0         0                0                0         45000       0.0000
```

Manuscript ecospace framework, limiting binary characters to two mutually "present" states and weighing state occupation by frequency within empirical Kope/Waynesville-formation species pool ['constraint=2'] (Fig. 3B): 95.51% correct

```
True model      Expansion-1  Neutral  PartitioningR-1  PartitioningS-1  Redundancy-1  class.error
Expansion-1           44489      259               63               72           117       0.0114
Neutral                 304    42343             1697              502           154       0.0590
PartitioningR-1         129     2849            40394             1491           137       0.1024
PartitioningS-1         112      514             1444            42750           180       0.0500
Redundancy-1             33       11               11               13         44932       0.0015
```

Manuscript ecospace framework, limiting binary characters to two mutually "present" states ['constraint=2'], weighing state occupation by frequency within empirical Kope/Waynesville-formation species pool, and including weakened (50%-rule-following) models (Fig. 3C in part)

This trial illustrates the confusion matrix for the classification tree used in the manuscript to classify empirical Kope and Waynesville formation samples. The tree was trained on simulated data spanning nine models: the neutral model; the four models that implement the redundancy, two versions of partitioning, and expansion rules; and four "weakened" versions of these four models, in which rules were followed at random 50% of the time (i.e., the 100% and 50% training data sets visualized in Figure 3C). Model names are abbreviated, with "-1" referring to the 100%-rule-following implementation and "-0.5" referring to the 50% implementation. The classification tree included the total variance (V) statistic (unlike those above).

The confusion matrix was produced from a validation data set, i.e., samples excluded from the training data set used to build the classification tree. The training data set included 5,265,000 samples (9 models * 3000 simulations * 195 sample sizes [5–200 species per sample]) and the validation data set (whose confusion matrix is illustrated below) included 855,000 samples (1000 simulations * 95 sample sizes [5–100 species per sample] = 95,000 samples per model * 9 models). The classification rate on the training data set was 91.7% (79.9% when restricted to 1000 samples, sample sizes below 50, and excluding V from the training data set, for comparative purposes with above confusion matrices) and the overall classification rate on the validation data set was 78.1% (75.2% when restricted to sample sizes below 50 and excluding V, for comparative purposes with above confusion matrices). See Supplementary Figure 9 for an alternate version of the confusion matrix, illustrating performance as a function of sample size and including additional validation samples.

```
Truth      Exp-0.5  Exp-1 Neutral PartR-0.5 PartR-1 PartS-0.5 PartS-1 Redund-0.5 Redund-1  class.error
Exp-0.5      92049    816     937       285      82       327      80        288      136       0.0311
Exp-1          713  93969      68        35      19        27      12         49      108       0.0109
Neutral       1275    112   65234     12889     581      7598     208       6974      129       0.3133
PartR-0.5      895    102   17747     47997    2468     19126     403       6120      142       0.4948
PartR-1        265     54    1116      2615   83205      3416    1253       2939      137       0.1242
PartS-0.5     1164     83   13221     26207    2442     36869     731      14157      126       0.6119
PartS-1        151     51     158       218    1261       378   91424        425      934       0.0376
Redund-0.5     810    151    9706      5842    2555     12988     725      62071      152       0.3466
Redund-1        27     31       9        16      21        11      30         16    94839       0.0017
```

Supplementary Appendix 3. Spreadsheet with life-habit/functional-trait codings for the Kope and Waynesville Formation species pool.

KWTraits.csv is a comma-separated value (.csv) format file listing the aggregate species pool for the Kope and Waynesville Formation used in empirical analyses. (The file is also included as a data file within the ecospace R package.) The first three columns list taxonomic information. The remaining columns list ecospace character states (functional traits). See Supplementary Appendix 1 and Novack-Gottshall (2007) for information on characters and states. See text for explanation of how multistate characters were rescaled.

Supplementary Appendix 4. Model-selection support data files for Kope and Waynesville Formation samples, stratigraphic section, member, and formation aggregates.

Files are in comma-separated value (.csv) format. The first five columns describe the Paleobiology Database collection identification number, scale (hand sample, stratigraphic section, etc.) of the sample, and stratigraphic/section names. Columns 6–14 list sample size (S, species richness) and values for eight disparity statistics (with NA designating when a statistic could not be calculated, because there were fewer than four unique life habits in the sample); see text for descriptions and abbreviations of statistics. The last column identifies which model has the best support among those candidates considered. The remaining columns list the classification-tree support each sample has for each candidate model considered.

emp2-modelfits.csv lists model support using the classification tree trained on the 50% and 100%-strength training data sets. emp3-modelfits.csv lists model support for the tree trained on 50%, 90%, and 100% training data. emp5-modelfits.csv lists model support for the tree trained on 50%, 75%, 90%, 95%, and 100% training data.