Selecting Food Web Models using Normalised Maximum Likelihood

Phillip P.A. Staniczenko[1,2,*], Matthew J. Smith[2] & Stefano Allesina[2,3]

[1]Centre for Biodiversity and Environment Research, University College London, UK (current address)

[2]Department of Ecology & Evolution, University of Chicago, Chicago, IL, USA

[3]Computation Institute, University of Chicago, Chicago, IL, USA

[*]Corresponding author: p.staniczenko@ucl.ac.uk

Running title: Selecting Food Web Models

Number of words: $\sim$ 7000 (66 References, 1 Table, 2 Figures)

Abstract

1. Ecological models link theory and data. They distil processes into a mathematical form that explains the salient features of observed data. Food webs describe the pattern of interactions between species in an ecosystem, and many models have been proposed to explain their structure. When selecting the most appropriate model for data, it is important to penalise against overly complicated models.

2. Here we introduce to ecology the use of Normalised Maximum Likelihood (NML) for model selection and demonstrate its application to models for food web structure. Unlike AIC, which penalises models using the number of parameters, NML normalises the likelihood of data given a model by the sum of likelihoods for all possible food webs with the same number of species. NML favours models that fit observed data well and all other data sets poorly, in contrast with overly flexible models that fit many (unobserved) data sets by the same amount and thus provide little information on the system under investigation. As such, NML represents a natural measure for comparing very different models, and enables ecologists to determine not only whether a particular model is superior to others, but also whether—objectively—the model is a poor description of data.

3. We used NML to compare models from four popular model families (cascade, niche, modular and group) and found that the best models performed much better than random graphs incorporating no ecological principles. However, models specified by empirical characteristics such as species body mass, taxonomic classification or habitat were frequently far-from-optimal, and, in some cases, performed worse than random graphs. This suggests that ecological interactions cannot be explained by a single species trait or coarse-grained environmental factor. The ranking of

empirically-determined models using NML was generally consistent with model selection according to AIC, BIC and Bayes factors. We also show how NML can improve the development of new model families by measuring the effectiveness of incremental changes to existing families or combining families.
4. NML offers ecologists a rigorous and elegant framework for revealing the defining features of data through the systematic formulation, testing and modification of models.

Key words: model selection, minimum description length, normalised maximum likelihood, cascade model, niche model, modularity, group model, AIC, BIC, Bayes factors.

Introduction

Debate in ecology is expected to decrease as the amount of data on a topic increases. A contentious idea will either be accepted or rejected in the face of data collected in the future (Hilborn & Mangel, 1997; Anderson *et al.*, 2000; Burnham & Anderson, 2002; Johnson & Omland, 2004; Ginzburg & Jensen, 2004). However, history has shown that some issues remain controversial long after time and money have been spent—and continue to be spent—collecting large amounts of relevant data. Notable examples include the relationship between diversity and stability (May, 1972; Pimm, 1984; McCann, 2000; Montoya *et al.*, 2006; Allesina & Tang, 2012), niche versus neutral determinants of community composition (Gause, 1934; Hubbell, 2001; Chave, 2004; Alonso *et al.*, 2006; Adler *et al.*, 2007; Levine & HilleRisLambers, 2009), the shape of species-area curves (McGill *et al.*, 2007; Chisholm & Pacala, 2010), the design of natural reserves for fragmented habitats (Simberloff & Abele, 1982), the appropriate choice of functional responses for species interactions (Abrams & Ginzburg, 2000) and the significance of nestedness in mutualistic networks (Bascompte *et al.*, 2003; Bastolla *et al.*, 2009; Thébault & Fontaine, 2010; Staniczenko *et al.*, 2013), to name but a few.

A lack of consensus can arise and persist because additional data often complicates debate. For this reason, models are devised to simplify ecological arguments into an unambiguous mathematical form that can be tested against data (Hilborn & Mangel, 1997; Anderson *et al.*, 2000; Burnham & Anderson, 2002; Ginzburg & Jensen, 2004). Although developing models is a difficult and important challenge in its own right, even with well-designed models and an abundance of high-quality data, the problem of model selection remains.

Model selection is concerned with choosing a statistical model from a set of candidate models (Burnham & Anderson, 2002). Naturally, simple models are preferred to complex ones (Forster & Sober, 1994; Johnson & Omland, 2004; Ginzburg & Jensen, 2004). But formalising what is meant by "simple" and "complex" is not straightforward: How exactly should, say, the number of parameters in a model be balanced against its fit to data? Several popular methods, including AIC (Akaike, 1998; Burnham & Anderson, 2002) and BIC (Ellison, 2004), are based on the concept of likelihood—the probability that a model reproduces observed data (Berger & Wolpert, 1988). But assigning an appropriate penalisation can be problematic when parameters are not easily countable, as is often the case with models for food web structure, which commonly involve structures such as hierarchies (representing feeding dominance, for example) or partitions (representing sets of highly-interacting species, for example). In addition, while most model selection techniques are designed to identify the better models out of a set of candidates, few, if any, are able to determine whether a model is objectively poor for a given data set. Identifying a poorly-performing model is arguably as important as identifying a well-peforming model because it allows the pool of candidate models, which usually increases through time, to be reduced in a rigorous manner (Johnson & Omland, 2004; Ginzburg & Jensen, 2004).

Here we introduce to ecology the use of Normalised Maximum Likelihood (NML) (Barron *et al.*, 1998; Rissanen, 2001; Myung *et al.*, 2006; Grünwald, 2007) to select among models for food web structure. The study of food webs, networks describing feeding interactions among species in an ecosystem, has a long and storied history (Pascual & Dunne, 2006; Bersier, 2007; McCann, 2011). A wide range of models has been proposed to generate synthetic networks with properties similar to those of empirical food webs. But despite rapid progress in model development, there is still no effective way of comparing different models within a single framework.

NML is the latest technique based on the Minimum Description Length (MDL) Principle from Information Theory (Rissanen, 1978, 1989; Barron *et al.*, 1998; Hansen & Yu, 2001; Grünwald, 2000, 2007). In contrast with other methods for model selection, NML enables ecologists to determine not only whether a particular model is superior to others, but also whether, in an absolute sense, the model is a poor description of the data under consideration. To this end, we used NML to compare food web models to two reference points: i) a random graph model that only takes into account the number of species and the number of interactions between those species and ii) the total amount of information,

in bits, contained in food web data.

We analysed six large marine food webs and found that the best-fitting models of four popular model families—cascade (Cohen & Newman, 1985), niche (Williams & Martinez, 2000), modular, and group (Allesina & Pascual, 2009)—always performed better than random graphs. However, versions of these models based on empirical characteristics such as species body mass, taxonomic classification (e.g., phyla or class) or habitat were frequently far-from-optimal, and, in some cases, performed worse than corresponding random graph models. This demonstrates that the best solutions for popular food web models do not map into simple species traits.

We compared the performance of empirically-determined models ranked according to NML to rankings obtained using AIC, BIC and Bayes factors and found consistent results for cascade, MPN and modular model families. The similarity of rankings between the four measures was less clear for group models, but in general, model selection based on NML and Bayes factors gave the most similar results, with AIC providing the most dissimilar ranking of models (favouring models with relatively large number of parameters and possibly leading to over-parameterisation).

NML can also be used to guide the development of new and better models. This process usually takes one of two forms: making small changes to existing models or combining the defining properties of two models (Burnham & Anderson, 2002; Johnson & Omland, 2004; Ginzburg & Jensen, 2004). We considered two variations on the niche model and combined the cascade model with a model inspired by the presence of compartments to illustrate the two forms.

Materials and Methods

Food Webs, Model Selection and Likelihood

For decades, selection among food web models has involved comparing network metrics (Pascual & Dunne, 2006; McCann, 2011). One first builds a synthetic distribution of values for a given network metric using model-generated webs then tests whether the empirical value falls within a predefined confidence interval, for example, the 5th and 95th percentiles of the distribution. If it does, then the model is considered an acceptable model for explaining the empirical network. The more metrics that a model can explain, the better the model (although one must be careful about co-varying metrics, e.g., average

trophic level and fraction of basal species). This approach is simple and economical. However, it suffers from three main problems. First, it has limited discriminatory power: different models will match different sets of metrics well, making it difficult to decide which of two (or more) models better explains the data. Second, it is difficult to account for model complexity: it is not clear how models with many parameters should be penalised, how the contribution of each parameter should be weighted, and under what conditions simpler models should be preferred. Third, it cannot assess whether a model can reproduce data exactly: even if several network metrics are measured, one cannot rule out the possibility that some observed feeding interactions will never be generated by a model.

Allesina *et al.* (2008) introduced the use of likelihood functions in food web analysis to solve these and related problems. A likelihood function measures the probability that a given model, along with its associated parameters, generates exactly the observed data (Berger & Wolpert, 1988). They showed that some popular models for food web structure were in fact incompatible with data (such models had a likelihood of zero) and proposed amendments to make them suitable for analysing empirical food webs. Below, we describe likelihood expressions for five model families: random graph, cascade, minimum potential niche, modular and group.

As with most types of network, a food web can be represented by an adjacency matrix A in which a non-zero element A_{ij} indicates a feeding interaction between a consumer species j and a resource species i. Throughout the manuscript we write the likelihood that a model M with vector of parameters $\boldsymbol{\theta}$ reproduces a given food web as $L(A|M,\boldsymbol{\theta})$. The parameter values that maximise the likelihood function are referred to as the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ and the corresponding likelihood is known as the maximum likelihood $\hat{L}(A|M,\hat{\boldsymbol{\theta}})$. It is often useful to write the maximum log-likelihood $\ln \hat{L} = \hat{\mathcal{L}}_e$; where the subscript indicates the base of the logarithm.

Random graph

The simplest model for food web structure is a directed Erdős-Rényi random graph (Newman *et al.*, 2001). In this model, there is a single probability p of an edge between species i and j, with a corresponding probability $(1-p)$ that two species do not interact (Fig. 1A). As such, the model has a single parameter. The likelihood of reproducing a food web

with S species and U edges is

$$L(A|\text{Rnd}, p) = \prod_i^S \prod_j^S p^{A_{ij}}(1-p)^{1-A_{ij}} = p^U(1-p)^{S^2-U} = p^U(1-p)^Z; \quad (1)$$

where U is the total number of ones in the adjacency matrix and Z is the total number of zeros. Thus, the likelihood is the same for all food webs with the same combination of S and U.

The maximum likelihood estimate of p is simply $\hat{p} = U/S^2$ and the maximum log-likelihood is then

$$\hat{\mathcal{L}}_e(A|\text{Rnd}, \hat{p}) = U \ln \hat{p} + Z \ln(1-\hat{p}); \quad (2)$$

where we assume that $0 \ln 0 = 0$ to avoid infinities.

Cascade model family

Cascade models (Cohen & Newman, 1985) are based on the assumption that species can be ordered along a single dimension, such as body mass or mobility, which defines their probability of interacting with other species. Species are ordered in a hierarchy H, and for a given food web with S species, each unique hierarchy defines a separate cascade model. Because a hierarchy is simply a permutation of species, there are a total of $S!$ possible cascade models for a given food web. For a particular H, each species has a probability p of feeding on species that are below it in the hierarchy and a probability q of being cannibalistic or feeding on higher-ranked species. As such, each cascade model can be seen as two random graphs: one encompassing the upper-triangular part of the matrix when species are ordered by H, with parameter p, and one encompassing the lower-triangular part (including the diagonal), with parameter q (Fig. 1B). The maximum log-likelihood is

$$\hat{\mathcal{L}}_e(A|\text{Casc}_H, \hat{p}, \hat{q}) = U_1 \ln \hat{p} + Z_1 \ln(1-\hat{p}) + U_2 \ln \hat{q} + Z_2 \ln(1-\hat{q}); \quad (3)$$

where H determines how many ones (U_1) and zeros (Z_1) are in the upper-triangular part of the matrix and how many (U_2, Z_2) are in the diagonal or lower-triangular part, which in turn specifies the maximum likelihood estimates $\hat{p} = U_1/(U_1 + Z_1)$ and $\hat{q} = U_2/(U_2 + Z_2)$.

The idea that a cascade model is the combination of two random graph "slices" can be extended to specify more intricate models for food web structure. We continue with this slice-based approach to describe the three remaining model families.

Minimum potential niche model family

The minimum potential niche (MPN) model family (Allesina *et al.*, 2008) is a variation on the niche model (Williams & Martinez, 2000) that focuses on its central idea: intervality. In a MPN model, species are ordered in a hierarchy H and each consumer has a restricted feeding interval of consecutive species which contains all of its prey. Each consumer i feeds on species within its interval with probability p_i and on species outside its interval with probability $q_i = 0$. A MPN model divides an adjacency matrix into $2S$ slices: for each consumer, one slice represents its feeding interval (with associated probability p_i) and the other slice represents its non-feeding interval (with associated probability $q_i = 0$)(Fig. 1C).

The maximum log-likelihood for a MPN model with a given H is

$$\hat{\mathcal{L}}_e(A|\mathrm{MPN}_H, \hat{p}_i, \hat{q}_i = 0) = \sum_i \left(U_{i,1} \ln \hat{p}_i + Z_{i,1} \ln(1 - \hat{p}_i)\right); \tag{4}$$

where $U_{i,1}$ and $Z_{i,1}$ are the number of ones and zeros, respectively, in the slice associated with the feeding interval of consumer i, and the consumer's feeding probability is set to its maximum likelihood estimate $\hat{p}_i = U_{i,1}/(U_{i,1} + Z_{i,1})$.

Modular model family

The modular model family is based on the presence of compartments or modules in ecology (Krause *et al.*, 2003; Allesina & Pascual, 2009; Rezende *et al.*, 2009; Guimerà *et al.*, 2010; Stouffer & Bascompte, 2011). Modules are often associated with different local habitats or seasons, and species within the same module are expected to have a higher probability of interacting with one another compared to two species in different modules. A modular model divides species into a set partition Π (i.e., each species is assigned to only one module and therefore modules are non-overlapping); two species in the same module interact with probability p, while species in different modules interact with probability q. As with cascade models, each partition Π divides an adjacency matrix into two slices: one composed of all the square blocks on the diagonal (within-module interactions) and one composed of all other matrix elements (between-module interactions)(Fig. 1D).

The maximum log-likelihood for a modular model is formally similar to that of a cascade model but is defined by a partition:

$$\hat{\mathcal{L}}_e(A|\mathrm{Mod}_\Pi, \hat{p}, \hat{q}) = U_w \ln \hat{p} + Z_w \ln(1 - \hat{p}) + U_b \ln \hat{q} + Z_b \ln(1 - \hat{q}); \tag{5}$$

where Π determines how many ones (U_w) and zeros (Z_w) are in the matrix slice representing within-module interactions and how many (U_b, Z_b) are in the matrix slice representing between-module interactions, which in turn specifies the maximum likelihood estimates $\hat{p} = U_w/(U_w + Z_w)$ and $\hat{q} = U_b/(U_b + Z_b)$.

Group model family

The group model family extends the concept of compartments introduced with the modular model family. A group model (Allesina & Pascual, 2009) (also known as a stochastic block model (Karrer & Newman, 2011)) is also defined by a partition Π, which specifies to which of γ non-overlapping groups each species belongs. The probability that a consumer j preys on resource i depends exclusively on the corresponding groups of species i and j: $p_{ij} = p_{\Pi_i \Pi_j} = p_{kl}$; where k and l index groups. As such, each partition Π divides the adjacency matrix into γ^2 slices (and therefore there are a total of γ^2 probabilities)(Fig. 1E).

The maximum log-likelihood for a group model is

$$\hat{\mathcal{L}}_e(A|\mathrm{G}_\Pi, \hat{p}_{kl}) = \sum_{kl} \left(U_{kl} \ln \hat{p}_{kl} + Z_{kl} \ln(1 - \hat{p}_{kl})\right); \tag{6}$$

where the partition Π determines how many ones (U_{kl}) and zeros (Z_{kl}) are in the matrix slice representing interactions between groups k and l, which in turn specifies the maximum likelihood estimate $\hat{p}_{kl} = U_{kl}/(U_{kl} + Z_{kl})$.

Balancing model fit and complexity

In general, models with many parameters (or more flexible models) yield better likelihoods (Berger & Wolpert, 1988; Burnham & Anderson, 2002; Johnson & Omland, 2004; Ginzburg & Jensen, 2004). This difference in model complexity must be taken into account when comparing likelihoods. The simplest and most popular correction used for model selection is AIC (Akaike Information Criterion) (Akaike, 1998; Burnham & Anderson, 2002), although other methods are also used, including BIC (Schwarz, 1978) and Bayes factors (Kass & Raftery, 1995).

AIC measures the loss of information when a model is used to describe data. Formally, it is an asymptotic approximation to the Kullback-Leibler divergence and is defined in terms of the maximum log-likelihood (Burnham & Anderson, 2002):

$$\mathrm{AIC}(A, M, \hat{\boldsymbol{\theta}}) = 2k - 2\hat{\mathcal{L}}_e(A|M, \hat{\boldsymbol{\theta}}); \tag{7}$$

where k is the number of parameters in the model. Model fit (maximum log-likelihood) is balanced against model complexity by assigning a penalisation of one point of log-likelihood to each parameter. In this way, model complexity in AIC is measured by the number of parameters.

BIC uses a slightly different correction for model complexity (Schwarz, 1978):

$$\mathrm{BIC}(A, M, \hat{\boldsymbol{\theta}}) = k \ln(S^2) - 2\hat{\mathcal{L}}_e(A|M, \hat{\boldsymbol{\theta}}); \tag{8}$$

where the penalisation for each parameter is now proportional to the logarithm of the amount of the data being fit.

For a random graph,

$$\mathrm{AIC}(A, \mathrm{Rnd}, \hat{p}) = 2 - 2(U \ln \hat{p} + Z \ln(1 - \hat{p})); \tag{9}$$

$$\mathrm{BIC}(A, \mathrm{Rnd}, \hat{p}) = \ln(S^2) - 2(U \ln \hat{p} + Z \ln(1 - \hat{p})) \tag{10}$$

while for a cascade model,

$$\mathrm{AIC}(A, \mathrm{Casc}_H, \hat{p}, \hat{q}) = 4 - 2(U_1 \ln \hat{p} + Z_1 \ln(1 - \hat{p}) + U_2 \ln \hat{q} + Z_2 \ln(1 - \hat{q})); \tag{11}$$

$$\mathrm{BIC}(A, \mathrm{Casc}_H, \hat{p}, \hat{q}) = 2\ln(S^2) - 2(U_1 \ln \hat{p} + Z_1 \ln(1 - \hat{p}) + U_2 \ln \hat{q} + Z_2 \ln(1 - \hat{q})). \tag{12}$$

AIC and BIC are simple to compute but have two main drawbacks: they only hold asymptotically (i.e., for large amounts of data), and parameters that have little influence on the likelihood are penalised by exactly the same amount as those that strongly influence the likelihood. Additionally, and perhaps more importantly for food web models, the seemingly straightforward task of counting the number of parameters in a model can, in practice, be very difficult when parameters are not numbers but more complex structures such as partitions or permutations.

With AIC (and a similar argument can be made for BIC), all cascade and modular models are subject to two points of log-likelihood penalisation, despite a cascade model being parameterised by a permutation (species hierarchy) and a modular model by a partition (which species belong to which of up to S modules). So the fact that hierarchies are fundamentally different from partitions is not taken into account, which makes comparing the performance of different model families difficult. Even comparing models within a family is not straightforward: the penalisation is the same whether a modular model involves few or many modules, and makes no consideration of which species belong to

which modules. This limitation extends to group models. Consider the set of models that partition a food web with ten species into two groups. A model that assigns nine species to one group and one species to a second does not have the same degree of freedom, and hence model complexity, as a model that assigns five species to both groups, yet all models involving two groups are penalised by exactly the same amount (four points of log-likelihood). Some of these issues can be addressed using Bayes factors.

Bayes factors are derived from Bayes' theorem (Kass & Raftery, 1995). For a model selection problem in which we have to choose between two models, and with no *a priori* preference for either model, the relative plausibility of the two different models is assessed by the Bayes factor, which is defined as the ratio of each model's marginal likelihood:

$$K(M_1, M_2) = \frac{P(A|M_1)}{P(A|M_2)}. \tag{13}$$

Marginal likelihoods do not depend on any single set of parameters. This is because the expression for marginal likelihood integrates over all parameters in the model. Given the choice of two models, the one with the highest marginal likelihood should be preferred as it offers a better balance between goodness-of-fit and complexity. Bayes factor penalisation for model complexity is not explicit, but is done automatically during the integration over possible parameter values. In fact, the marginal likelihood can be interpreted as the expected likelihood when parameterising the model by randomly sampling parameter values from their priors. Formally, the marginal likelihood is written

$$P(A|M_1) = \int_{\theta_1} \int_{\theta_2} \cdots \int_{\theta_k} L(A|M_1, \boldsymbol{\theta}) P(\boldsymbol{\theta}|M_1)\, d\theta_k \cdots d\theta_2\, d\theta_1; \tag{14}$$

where $P(\boldsymbol{\theta}|M_1)$ is the probability of a given parameterisation when sampling parameters from their prior distributions (see Supporting Information).

The main drawback with using marginal likelihood and Bayes factors to evaluate food web models is the requirement to specify priors. Furthermore, integrating over parameters can be very difficult for complex models, especially those involving discrete parameters or combinatorial structures such as permutations or partitions (Baskerville *et al.*, 2011; Eklöf *et al.*, 2012).

Minimum Description Length Principle

A set of alternative methods for model selection are based on the Minimum Description Length Principle (Rissanen, 1978, 1989; Barron *et al.*, 1998; Hansen & Yu, 2001;

Grünwald, 2000, 2007). In this approach, model selection is considered a problem of data compression. Data have a given length in bits of information, and better models are able to compress data more than worse models. In the simplest application of MDL, if we wanted to transmit a finite-sized amount of data over a channel such as the Internet, we would like to choose a model that minimises the total length $\mathfrak{L}_M(A) = \mathfrak{L}(A|M) + \mathfrak{L}(M)$; where $\mathfrak{L}(A|M)$ is the length in bits of the original data after being encoded by the model and $\mathfrak{L}(M)$ is the length in bits required to describe the model. Given these two pieces of information, the transmitted message can be decoded by the receiver to obtain the original data.

The fraction of realised interactions in a food web is usually quite low (Pascual & Dunne, 2006; McCann, 2011), leading to more zeros than ones in associated adjacency matrices. This regularity can be exploited to compress such matrices: common sequences of symbols (e.g., `00`, two consecutive zeros) can be replaced by short code words, with rarer sequences replaced by longer code words. Consider a four-species food chain that is part of a much larger food web:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{15}$$

Naïve transmission of this matrix involves sending a message of $S^2 = 16$ bits: `0100001000010000`. But we can take advantage of the fact that there are few ones in the message by devising a prefix code—a code that can be uniquely decoded—to compress the data. (Here, the concept of prefix codes can be considered a deterministic analogue of probabilistic food web models.) For example, we can write `0` $\rightarrow$ `10`, `1` $\rightarrow$ `11` and `0000` $\rightarrow$ `0`. Using this code, the number of bits required to send this particular pattern of interactions is reduced from 16 to 11 bits: `0100001000010000` to `10110110110`. Although only a small reduction in this example, especially considering that the prefix code must be transmitted along with the data string, in the case of a large food web, it could very well be beneficial to send a fixed-length prefix code, $\mathfrak{L}(M)$, along with the a compressed data string, $\mathfrak{L}(A'|M)$.

Early implementations of the MDL Principle had limited real-world application because it was difficult to compute the appropriate number of bits required to describe a prefix code or model (see Myung *et al.* (2006) and Supporting Information). However, thirty years of development (Barron *et al.*, 1998) has led to the Normalised Maximum

Likelihood, which does not require the computation of $\mathfrak{L}(M)$, finally making the MDL Principle relevant to the study of food web models and similar problems.

Normalised Maximum Likelihood

NML quantifies how well a model explains a particular data set compared to how well it explains data in general (Barron *et al.*, 1998; Rissanen, 2001; Myung *et al.*, 2006; Grünwald, 2007). The NML distribution of a particular data set A given a model M is

$$\mathrm{NML}(A|M) = \frac{\hat{L}(A|M, \hat{\boldsymbol{\theta}}_A)}{\int \hat{L}(A'|M, \hat{\boldsymbol{\theta}}_{A'})\, dA'}; \tag{16}$$

where the normalisation is over all data sets A' with the same number of data points as the original data set. NML returns values in the range [0,1].

A complex model with many parameters will typically fit many data sets well because of its flexibility. This outcome would result in a large denominator and thus a small value for NML (as would an overly simple model that fits *all* data sets the same amount). On the other hand, a model that fits only observed data well and all other data sets poorly would result in a large value for NML. An analogy is often made to the problem of fitting a polynomial function (model) to a sequence of data consisting of n pairs (x, y), where x and y are real numbers (Grünwald, 2000). The classical solution to this problem is to perform a standard linear regression, which results in a "best-fit" line that captures some of the regularity in the data, but often appears to *underfit* the data. The other extreme solution is to pick a polynomial of degree $n-1$ that goes exactly through all n points being fit. In doing so, there is a large risk of *overfitting*. Instead, we might prefer an intermediate-degree polynomial: one that permits small (but non-zero) error and is still relatively simple (i.e., has few parameters). It is with similar intent that NML quantifies the trade-off between model complexity and goodness-of-fit.

When data are discrete, as with food webs, the integral in Eqn 16 is replaced by a summation. The total length associated with NML is given by

$$\begin{aligned} \mathfrak{L}_M(A) &= -\log_2 \mathrm{NML}(A|M) \\ &= -\log_2 \hat{L}(A|M, \hat{\boldsymbol{\theta}}_A) + \log_2 \sum_{A'} \hat{L}(A'|M, \hat{\boldsymbol{\theta}}_{A'}) \\ &= -\hat{\mathcal{L}}_2(A|M, \hat{\boldsymbol{\theta}}_A) + \log_2 \mathcal{C}(M, A); \end{aligned} \tag{17}$$

where $\mathcal{C}(M, A)$ is a penalisation constant (known as the parametric complexity in the Information Theory literature (Rissanen, 1986, 1987, 1989, 1996)).

The NML for a random graph is

$$\mathrm{NML}(A|\mathrm{Rnd}) = \frac{\hat{L}(A|\mathrm{Rnd},\hat{p})}{\sum_{A'}\hat{L}(A'|\mathrm{Rnd},\hat{p}')}. \tag{18}$$

The numerator is simply the maximum likelihood derived above: $\hat{p}^U(1-\hat{p})^Z$. However, the denominator requires us to sum the maximum likelihoods for each of the 2^{S^2} possible matrices (from completely empty to completely filled including all possible configurations of edges), a computationally-intractable task for large S. (In practice, we would need to calculate fewer than 2^{S^2} values because the maximum likelihood of a random graph depends only on the number of ones and zeros in the matrix, simplifying things slightly and leading to a summation involving S^2+1 terms.) To circumvent this limitation, we prove a new identity that removes the need for the summation altogether.

Suppose that we have a random-graph-like matrix slice of length $X = U+Z$ containing U ones and Z zeros. Each element in the slice is assumed to be the result of a Bernoulli trial in which the probability of obtaining a one is set to its maximum likelihood estimate $\hat{p} = \frac{U}{X}$. The maximum likelihood for a slice is then

$$\hat{L}(U,Z|X) = \hat{p}^U(1-\hat{p})^Z; \tag{19}$$

and the normalised maximum likelihood is

$$\mathrm{NML} = \frac{\hat{L}(U,Z|X)}{\sum_{k=0}^{X}\binom{X}{k}\left(\frac{k}{X}\right)^k\left(\frac{X-k}{X}\right)^{(X-k)}}; \tag{20}$$

where the denominator specifies the sum of maximum likelihoods over all possible numbers of ones in the matrix slice (the first term represents a weighting for the number of configurations involving k ones). In the Supporting Information we show that the denominator can be written in a form that only depends on the length of a matrix slice:

$$\mathcal{C}(X) = 1 + \frac{e^X\Gamma(X,X)}{X^{X-1}} \approx 1 + \left(\sqrt{X\frac{\pi}{2}} - \frac{1}{3} + \frac{\sqrt{2\pi}}{24\sqrt{X}} - \frac{4}{135X} + \frac{\sqrt{2\pi}}{576\sqrt{X^3}} + \frac{8}{2835X^2}\right). \tag{21}$$

This expression can be computed very quickly, and the penalisation for each slice can be multiplied (or added in log-space) to obtain the penalisation constant for models that comprise more than one random-graph-like matrix slice.

The total length associated with NML for a random graph can be written

$$\mathfrak{L}_{\mathrm{Rnd}}(A) = -\hat{\mathcal{L}}_2(A|\mathrm{Rnd},\hat{p}) + \log_2\mathcal{C}(\mathrm{Rnd},A). \tag{22}$$

The maximum likelihood is computed as for AIC, but, instead of a single point of log-likelihood penalisation, $\mathcal{C}(\mathrm{Rnd}, A) = \mathcal{C}(S^2)$, which depends only on the size of the matrix.

As a cascade model is composed of two random-graph-like matrix slices, its total length based on NML can be computed by using the fact that the penalisation constant can be factored:

$$\mathfrak{L}_{\mathrm{Casc}_H}(A) = -\hat{\mathcal{L}}_2(A|\mathrm{Casc}_H, \hat{p}, \hat{q}) + \log_2 \mathcal{C}(\mathrm{Casc}_u, A) + \log_2 \mathcal{C}(\mathrm{Casc}_l, A); \qquad (23)$$

where $\mathcal{C}(\mathrm{Casc}_u, A)$ is the constant associated with the upper triangular part of the matrix and $\mathcal{C}(\mathrm{Casc}_l, A)$ with the diagonal and lower triangular part. Similar expressions for total length can be derived for MPN, modular and group model families (see Supporting Information).

We have focused on NML and total length for models with uniquely defined hierarchies or partitions; as such, hierarchies and partitions can be considered parameters within the context of a model family. In principle, one could also calculate a single NML value for an entire model family. However, such calculations are computationally infeasible, even with the new identity, Eqn 21. For example, computing the NML of the cascade model *family* for a given food web, as opposed to a single model within the cascade model family, would require us to sum maximum likelihood estimates over the 2^{S^2} possible matrices for each of the $S!$ possible hierarchies. Although the expression for NML penalisation (Eqn 21) greatly speeds up the 2^{S^2} sum for individual hierarchies, the complete set of $S!$ hierarchies must still be computed exhaustively. A similar issue arises if we want to compute NML for all possible modular models with, say, four modules, which would require a sum over all possible partitions of S species into four non-overlapping sets—a potentially huge number of possibilities even for small S. (The above issues with NML are also relevant for calculation of Bayes factors.)

Despite the difficulty in computing NML and total lengths for model families, we can still attempt to compare families using the range of total lengths spanned by their constituent models. More specifically, given an empirical food web, we can compare the best-performing models from each model family, with the best of the candidate families said to be the one containing the model with the overall shortest total length.

Although a single total length for a model family is preferable to comparing total length ranges, we are typically interested in model performance when hierarchies and partitions correspond to field-measurable properties, that is, when they are determined

empirically. In this case, a hierarchy or partition need no longer be considered a parameter and we can use NML and total length (Eqn 17) to directly compare food web models. By considering empirically-determined models, we are also able to compare results with AIC, BIC and Bayes factors. Such an exercise is not possible when comparing model families because, with AIC and BIC, it is not clear how many points of log-likelihood to penalise one particular hierarchy or partition in a model family compared to another hierarchy or partition in the same family. It should be stressed that, even though hierarchies and partitions may no longer be considered parameters, the effectiveness of penalising model complexity using the number of parameters remains questionable: with AIC (and similarly with BIC), the same two points of log-likelihood are used to penalise all empirically-determined modular models, regardless the number of species and modules involved and how those species are partitioned into each module.

Results

Model Selection using NML and Total Length

We analysed the performance of cascade, MPN, modular and group models and model families using six marine food webs: Kongsfjorden (Jacob *et al.*, 2011); Lough Hyne (Riede *et al.*, 2010); Reef (Optiz, 1996); St. Marks (Christian & Luczkovich, 1999); Weddell Sea (Jacob, 2005); and Ythan Estuary (Cohen *et al.*, 2009) (Table 1). We compared model families using the range of total lengths spanned by their constituent models. We also used empirical data to determine model hierarchies and partitions, and compared results for total length based on NML to AIC, BIC and Bayes factors.

The total lengths (Eqn 17) of a random graph and uncompressed data represent two helpful points of reference with which to compare more complex food web models. As random graphs only take into account the number of species and the number of interactions between those species, we would expect models incorporating more ecological principles to return shorter total lengths than corresponding random graphs. Models with total lengths longer than random graphs should be treated with caution, and those with total lengths longer than uncompressed food web data are particularly poor descriptions of observed data. When food web connectance ($C = \frac{U}{S^2}$, the proportion of possible trophic interactions between species that are realised) is sufficiently low—as is the case for all currently-published food webs—or high, a random graph will always yield a total length

that is less than the length of the data when uncompressed. Here we used total length of the most basic random graph (see *Materials and Methods*). This type of random graph only retains the connectance of observed data (a nonetheless important property of food webs (Pascual & Dunne, 2006; McCann, 2011)), but other versions could be used, such as one that preserves an empirical food web's in- or out-degree (its distribution of incoming or outgoing trophic interactions, respectively) (Newman *et al.*, 2001).

Model family performance

We obtained total length ranges for model families using a stochastic optimisation algorithm (see Eklöf *et al.* (2013) for details). For each combination of model family and food web, we searched for the hierarchy (cascade and niche) or partition (modular and group) of species that resulted in the shortest (best) and longest (worst) total lengths. For example, with the cascade model family and a given food web, the algorithm would calculate the total length of an initial hierarchy (ordering of species), then adjust the hierarchy and recalculate total length, then accept or reject the new hierarchy based on whether the shortest or longest total length is sought; this process would continue until one was reasonably sure that the hierarchy with the shortest or longest total length had been found. All other models in a family are necessarily contained in this range of total lengths, which enables a quick indicative comparison of model family performance.

The group model family spanned the largest range of total lengths, followed by MPN, then, with similar ranges, modular and cascade families (Fig. 1 and Fig. S1). All model families included a large number of models with total lengths shorter than a corresponding random graph. The best models, those with the shortest total lengths, were from the MPN model family for the four smaller webs and the group model family for the two larger webs. By design, models from the MPN family constrain the number of feeding interactions of each consumer to the observed distribution. As such, this model family consistently explained data well (despite additional model complexity compared to most other families), with even the worst MPN model performing much better than the corresponding random graph. This result confirms that ecological networks are very different from random graphs (Dunne *et al.*, 2002), and, consequently, effective models should, at the very least, retain the degree distribution of empirical food webs. The shift of best-performing model with network size is consistent with earlier results (Allesina & Pascual, 2009), and may be due to large food webs being composed of different sub-systems that

are largely independent of one another, a feature that is well described by group models.

The total length ranges for modular and cascade model families were similar to one another: the best models from each family were much worse than the best MPN or group models, while the worst models had total lengths longer than corresponding random graphs. However, when searching for models resulting in the longest total lengths, only the group model family produced models with total lengths longer than uncompressed data: Models with exactly S groups (where each species is in its own group) always have a likelihood of one, but the penalisation owing to model complexity results in total lengths equal to that of uncompressed data (see Supporting Information). (This also functions as a clear example of the need to penalise for model complexity during model selection.) For all food webs except Lough Hyne and Kongsfjorden, we were able to find group models with fewer than S groups and maximum likelihoods less than one, which, combined with the penalisation owing to model complexity, lead to total lengths longer than uncompressed data.

Empirically-determined models

With cascade and MPN model families, the hierarchy associated with the best model (that with the shortest total length) for a given food web can be considered the optimal ordering of species with respect to the model family. We computed total lengths for models in which hierarchies were defined by body mass and trophic level and compared them to the value for the optimal ordering. Hierarchies based on body mass typically resulted in shorter total lengths than trophic level for both model families. However, both empirical hierarchies were grossly sub-optimal for MPN models, and only slightly less sub-optimal for cascade models (however, total lengths for the empirically-determined MPN models were still much better than comparable total lengths for cascade models with optimal species ordering, Fig. 1 and Fig. S2). This suggests that the optimal feeding hierarchy of species does not map into a single dimension that can be associated with simple species traits.

In a similar vein, we used taxonomic information (Kingdom, Phylum, Class and Order) and habitat to define species partitions in modular and group models. All empirically-determined modular models performed very poorly and often had total lengths comparable to random graphs. With the group model family, partitions based on Phylum and Class fared well (in line with other findings (Eklöf *et al.*, 2012)), while Order and Kingdom

often resulted in much worse models, in some cases producing models with total lengths longer than random graphs (Fig. 1 and Fig. S3). Group models based on habitat tended to sit between models based on Phylum or Class and Order or Kingdom, and only for Ythan food web was it the best performing of the empirically-determined models.

As hierarchies and partitions are no longer considered parameters in empirically-determined models, we are able to compare the assessment of different models within a family according to different selection measures.

Comparison to AIC, BIC and Bayes factors

We compared the ranking of empirically-determined models within each model family according to NML with rankings given by AIC, BIC and Bayes factors (Table S4). (We used uniform priors when calculating Bayes factors (Kass & Raftery, 1995).) Rankings were consistent across the four measures when hierarchies were specified by either body mass or trophic level. For example, all four measures selected body mass over trophic level for explaining the trophic structure of Weddell Sea with a cascade model. Furthermore, for a given food web, the preferred empirical hierarchy (body mass or trophic level) was typically the same for cascade and MPN model families, with Ythan being the only exception to this trend.

Rankings for modular models were also very similar. Only St. Marks food web displayed a difference between the four measures. In this case, total length based on NML and Bayes factors selected the same ranking of the five empirical partitions, while AIC and BIC both selected the same, distinct ranking. The pattern of rankings between measures was less clear for group models. In general, NML and Bayes factors again showed similar rankings (see Lough Hyne, St. Marks, Weddell Sea and Ythan), while AIC often produced rankings that differed most from the other three measures (see Reef and St. Marks).

The similarity between NML and Bayes factors is not unexpected as it has been shown that they can yield similar results under certain mathematical conditions (Grünwald, 2007). The difference seen with AIC is also not unexpected, as it offers the most basic penalisation for model complexity which, unlike even BIC, does not take into account the size of data being fit. As a result, AIC tended to favour models with more parameters compared to those favoured by the other measures, possibly leading to over-parameterisation.

We further compared the four measures using simulations. We tested whether AIC, BIC, Bayes factors and NML could be used to recover information on species partitions

in food webs generated by a known group model (see Supporting Information). We found that AIC tended to favour a larger number of groups than specified in the generating model, whereas BIC, Bayes factors and NML all performed much better at recovering the specific partition or number of groups used to generate synthetic food webs. As with the ranking exercise, we found similar results for Bayes factors and NML; although it is worth noting that calculating Bayes factors incorporates information on synthetic food webs' generating distribution (through priors) not required when calculating NML (see Supporting Information).

Model Development

NML can be used to guide the development of new and better models. This process usually takes one of two forms: making small changes to existing models or combining the defining properties of two models (Burnham & Anderson, 2002; Forster & Sober, 1994; Johnson & Omland, 2004). We considered two variations on the MPN model family and combined the cascade with the modular model family (Fig. 2).

In an MPN model, species are ordered in a hierarchy and each consumer has a restricted feeding interval of consecutive species which contains all of its prey. Each consumer i feeds on species within its interval with probability p_i and on species outside its interval with probability $q_i = 0$. We successively relaxed the constraint on feeding to create two more flexible model families called N2 and N3 (Figs. 2B and 2C). N2 allows each species to feed outside its primary feeding interval with non-zero probability (Stouffer *et al.*, 2006; Allesina *et al.*, 2008), while N3 relaxes feeding constraints further by specifying three feeding intervals for each species: primary, above primary and below primary (see Supporting Information). These three niche models can be seen as more general, random graph slice-based analogues of the probabilistic niche model (Williams *et al.*, 2010), whose treatment under NML would be much more complicated.

We searched for models with the longest and shortest total lengths for each of these two variants. N2 models resulted in the shortest total lengths across all food webs (Fig. 2 and Fig. S4). In some cases, the total length of the N3 model was longer (worse) than the original MPN model, meaning that any increase in likelihood did not overcome the extra penalisation associated with increased model complexity. In light of these results, the additional complexity of the N2 compared to the MPN model appears to be beneficial, although the same cannot be said for the N3 model.

Better-performing models can also be produced by combining two different model families to form a new model family. The hybrid model family is a combination of the cascade and modular families (Fig. 2H). In a hybrid model, species are partitioned into modules, and species within the same module form an independent hierarchy (see Supporting Information). Even with its increased complexity, hybrid models performed much better (shorter total lengths) than models from the two original model families (albeit worse than the niche—MPN, N2 and N3—models, Fig. 2 and Figs. S5).

Discussion

An ecological model should explain observed data well and all other possible data sets of the same size poorly. This property is formalised by NML and can be expressed as a data transmission length: better models result in shorter total lengths because they are better at compressing regularities in observed data. In this way, total length defines a natural scale that is particularly useful for comparing models for food web structure, which are often defined by partitions or permutations that are not readily countable as parameters for complexity penalisation.

NML is not limited in the same way as AIC and BIC because it takes a fundamentally different approach to quantifying model complexity. With NML, the fit of a model to a particular data set is normalised by its fit to all other possible outcomes that could be observed. For a modular model, the normalisation is over all possible combinations of interactions in the within-module and between-module regions (slices) of the adjacency matrix (Fig. 1D); overly-complex models will typically fit many data sets well, leading to large normalising factors and total lengths. Unlike AIC and BIC, NML is not based on asymptotics, and because parameters are not explicitly counted (regardless of whether they are countable or not), models based on very different underlying principles, and with very different complexity, can be effectively compared. And unlike Bayes factors, NML does not require prior distributions for model parameters to be specified (however, it would be interesting to explore ways in which prior knowledge could be incorporated into the NML framework).

Even with the very different approach of NML to penalising model complexity, we found similar results to AIC, BIC and, especially, Bayes factors when ranking empirically-determined models. There was, however, a larger difference between measures in their

ability to recover information on species partitions in food webs generated by a known group model, with AIC performing especially poorly. It will be prudent to further compare these (and other) approaches to model selection, with a view to understanding the advantages and disadvantages of NML for analysing food webs and other ecological data.

Models that can be specified as an arrangement of random-graph-like matrix slices lend themselves especially well to comparison using NML. The conceptual simplicity of slice-based models makes it straightforward to create and assess new models. We showed how small changes to existing model families and combining characteristics of different model families can result in better (and worse) models. In addition to hierarchical or compartment-based feeding, it is worth exploring how other ecological features can be included in slice-based models. It is also worth exploring NML for non-slice-based models, so that all food web models can be compared using the same scale of total length (however, this is unlikely to be straightforward as computation of NML can be very difficult, especially if the identity for parametric complexity, Eqn 21, is not applicable).

As the number of models grows (and not just with regards to models for food web structure), ecologists need a consistent way of focusing their efforts, including discounting models that do not further understanding of the system under investigation. NML and total length enables models based on very different ecological principles to be compared on an equal footing: models with relatively short total lengths offer a good compromise between model fit and complexity—between overfitting and underfitting data. Models with total lengths at least as short as comparable random graphs should be retained and used as the basis for new and better models, while those with longer total lengths, especially those with total lengths longer than uncompressed data, should be the first to be discounted. Given the vast range of ecological systems that can be represented by food webs and other ecological networks, it will be informative to see which models—and which ecological principles—explain many systems, and which models are applicable only to, say, temperate and not tropical systems, or aquatic and not terrestrial systems. NML can help guide the process of model development, which will lead to a better understanding of the differences that define ecological systems.

Acknowledgments

PPAS supported by NSF #1042164 and an AXA postdoctoral research fellowship, MJS and SA by NSF #1148867. We thank S. Pawaar, E. Sander, S. Tang, C. Weinberger and three anonymous referees for comments. Dedicated to Marta E. Allesina, who is a good model of baby girl.

Data Accessibility

The six marine food webs used in this analysis can be found in the references provided in Table 1.

References

Abrams, P.A. & Ginzburg, L.R. (2000). The nature of predation: prey dependent, ratio dependent or neither? *Trends in Ecology & Evolution*, 15, 337–341.

Adler, P.B., HilleRisLambers, J. & Levine, J.M. (2007). A niche for neutrality. *Ecology Letters*, 10, 95–104.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In: *Selected Papers of Hirotugu Akaike*. Springer, pp. 199–213.

Allesina, S., Alonso, D. & Pascual, M. (2008). A general model for food web structure. *Science*, 320, 658–661.

Allesina, S. & Pascual, M. (2009). Food web models: a plea for groups. *Ecology Letters*, 12, 652–662.

Allesina, S. & Tang, S. (2012). Stability criteria for complex ecosystems. *Nature*, 483, 205–208.

Alonso, D., Etienne, R. & McKane, A. (2006). The merits of neutral theory. *Trends in Ecology & Evolution*, 21, 451–457.

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

643 Barron, A., Rissanen, J. & Yu, B. (1998). The minimum description length principle in
644 coding and modeling. In: *Information Theory 50 Years of Discovery*, vol. 44. Wiley
645 USA, pp. 699–716.

646 Bascompte, J., Jordano, P., Melián, C.J. & Olesen, J.M. (2003). The nested assembly of
647 plant-animal mutualistic networks. *Proceedings of the National Academy of Sciences*
648 *USA*, 100, 9383–9387.

649 Baskerville, E.B., Dobson, A.P., Bedford, T., Allesina, S., Anderson, T.M. & Pascual, M.
650 (2011). Spatial guilds in the serengeti food web revealed by a bayesian group model.
651 *PLoS Computational Biology*, 7, e1002321.

652 Bastolla, U., Fortuna, Miguel A., Pascual-Garcia, Alberto, Ferrera, Antonio, Luque, Bar-
653 tolo & Bascompte, Jordi (2009). The architecture of mutualistic networks minimizes
654 competition and increases biodiversity. *Nature*, 458, 1018–1020.

655 Berger, J. O. & Wolpert, R. L. (1988). *The likelihood principle*. 2nd edn. Hayward, CA:
656 Institute of Mathematical Statistics.

657 Bersier, L.-F. (2007). A history of the study of ecological networks. In: *Biological Networks*
658 *(ed. Képès F.)*. World Scientific, pp. 365–421.

659 Burnham, K.P. & Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A*
660 *Practical Information-Theoretic Approach*. 2nd edn. Springer.

661 Chave, J. (2004). Neutral theory and community ecology. *Ecology Letters*, 7, 241–253.

662 Chisholm, R.A. & Pacala, S.W. (2010). Niche and neutral models predict asymptotically
663 equivalent species abundance distributions in high-diversity ecological communities.
664 *Proceedings of the National Academy of Sciences USA*, 107, 15821–15825.

665 Christian, R.R. & Luczkovich, J.J. (1999). Organizing and understanding a winters sea-
666 grass foodweb network through effective trophic levels. *Ecological Modelling*, 117, 99–
667 124.

668 Cohen, J.E. & Newman, C.M. (1985). A stochastic theory of community food webs: I.
669 models and aggregated data. *Proceedings of the Royal Society London B*, 224, 421–448.

Cohen, J.E., Schittler, D.N., Raffaelli, D.G. & Reuman, D.C. (2009). Food webs are more than the sum of their tritrophic parts. *Proceedings of the National Academy of Sciences USA*, 106, 22335–22340.

Dunne, J.A., Williams, R.J. & Martinez, N.D. (2002). Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences USA*, 99, 12917–12922.

Eklöf, A., Helmus, M., Moore, M. & Allesina, S. (2012). Relevance of evolutionary history for food web structure. *Proceedings of the Royal Society London B*, 279, 1588–1596.

Eklöf, A., Jacob, U., Kopp, J., Bosch, J., Castro-Urgal, R., Chacoff, N.P., Dalsgaard, B., de Sassi, C., Galetti, M., Guimarães, P.R., Lomáscolo, S.B., Martín González, A.M., Pizo, M.A., Rader, R., Rodrigo, A., Tylianakis, J.M., Vázquez, D.P. & Allesina, S. (2013). The dimensionality of ecological networks. *Ecology Letters*, 16, 577–583.

Ellison, A.M. (2004). Bayesian inference in ecology. *Ecology Letters*, 7, 509–520.

Forster, M. & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *British Journal for the Philosophy Science*, 45, 1–35.

Gause, G.F. (1934). *The Struggle for Existence*. Hafner Press, New York.

Ginzburg, L.R. & Jensen, C.X.J. (2004). Rules of thumb for judging ecological theories. *Trends in Ecology & Evolution*, 19, 121–126.

Grünwald, P.D. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, 44, 133–152.

Grünwald, P.D. (2007). *The Minimum Description Length Principle*. MIT Press.

Guimerà, R., Stouffer, D.B., Sales-Pardo, M., Leicht, E.A., Newman, M.E.J. & Amaral, L.A.N. (2010). Origin of compartmentalization in food webs. *Ecology*, 91, 2941–2951.

Hansen, A.J. & Yu, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, 96, 746–774.

Hilborn, R. & Mangel, M. (1997). *The Ecological Detective: Confronting Models With Data*. Princeton University Press.

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

Jacob, U. (2005). *Trophic Dynamics of Antarctic Shelf Ecosystems—Food Webs and Energy Flow Budgets*. Ph.D. thesis, University of Bremen, Germany.

Jacob, U., Thierry, A., Brose, U., Arntz, W.E., Berg, S., Brey, T., Fetzer, I., Jonsson, T., Mintenbeck, K. & Mollmann, C. (2011). The role of body size in complex food webs: a cold case. *Advances in Ecological Research*, 45, 181–223.

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

Karrer, B. & Newman, M.E.J. (2011). Stochastic blockmodels and community structure in networks. *Physical Review E*, 83, 016107.

Kass, R.E. & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.

Krause, A.E., Frank, K.A., Mason, D.M., Ulanowicz, R.E. & Taylor, W.W. (2003). Compartments revealed in food-web structure. *Nature*, 426, 282–285.

Levine, J.M. & HilleRisLambers, J. (2009). The importance of niches for the maintenance of species diversity. *Nature*, 461, 254–257.

May, R.M. (1972). Will a large complex system be stable? *Nature*, 238, 413–414.

McCann, K.S. (2000). The diversity-stability debate. *Nature*, 405, 228–233.

McCann, K.S. (2011). *Food Webs*. Princeton University Press.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F. *et al.* (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, 10, 995–1015.

Montoya, J.M., Pimm, S.L. & Solé, R.V. (2006). Ecological networks and their fragility. *Nature*, 442, 259–264.

Myung, J.I., Navarro, D.J. & Pitt, M.A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, 50, 167–179.

Newman, M.E.J., Strogatz, S.H. & Watts, D.J. (2001). Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64, 026118.

Optiz, S. (1996). Trophic interactions in caribbean coral reefs. Tech. Rep. 43, ICLARM, Manila.

Pascual, M. & Dunne, J.A., eds. (2006). *Ecological Networks: Linking Structure to Dynamics in Food Webs*. Oxford University Press USA.

Pimm, S.L. (1984). The complexity and stability of ecosystems. *Nature*, 307, 321–326.

Rezende, E.L., Albert, E.M., Fortuna, M.A. & Bascompte, J. (2009). Compartments in a marine food web associated with phylogeny, body mass, and habitat structure. *Ecology Letters*, 12, 779–788.

Riede, J.O., Rall, B.C., Banasek-Richter, C., Navarrete, S.A., Wieters, E.A., Emmerson, M.C., Jacob, U. & Brose, U. (2010). Scaling of food-web properties with diversity and complexity across ecosystems. *Advances in Ecological Research*, 42, 139–170.

Rissanen, J. (1978). Modeling by the shortest data description. *Automatica*, 14, 465–471.

Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, 14, 1080–1100.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society B*, 49, 223–239.

Rissanen, J. (1989). *Stochastic complexity in statistical inquiry*. Singapore: World Scientic Publishing.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42, 40–47.

748 Rissanen, J. (2001). Strong optimality of the normalized ml models as universal codes
749 and information in data. *IEEE Transactions on Information Theory*, 47, 1712–1717.

750 Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

751 Simberloff, D. & Abele, L.G. (1982). Refuge design and island biogeographic theory:
752 effects of fragmentation. *American Naturalist*, 41–50.

753 Staniczenko, P.P.A., Kopp, J.C. & Allesina, S. (2013). The ghost of nestedness in ecolog-
754 ical networks. *Nature Communications*, 4, 1931 doi: 10.1038/ncomms2422.

755 Stouffer, D.B. & Bascompte, J. (2011). Compartmentalization increases food-web persis-
756 tence. *Proceedings of the National Academy of Sciences USA*, 108, 3648–3652.

757 Stouffer, D.B., Camacho, J. & Amaral, L.A.N. (2006). A robust measure of food web
758 intervality. *Proceedings of the National Academy of Sciences USA*, 103, 19015–19020.

759 Thébault, E. & Fontaine, C. (2010). Stability of ecological communities and the architec-
760 ture of mutualistic and trophic networks. *Science*, 329, 853–856.

761 Williams, R.J., Anandanadesan, A. & Purves, D. (2010). The probabilistic niche model
762 reveals the niche structure and role of body size in a complex food web. *PLoS ONE*, 5,
763 e12092.

764 Williams, R.J. & Martinez, N.D. (2000). Simple rules yield complex food webs. *Nature*,
765 404, 180–183.

Table and Figures

Food web	S	U	C	$\mathfrak{L}_{\text{rnd}}$	$\mathfrak{L}_{\text{raw}}$
Kongsfjorden (Jacob *et al.*, 2011)	252	1124	0.017	8157	63504
Lough Hyne (Riede *et al.*, 2010)	326	4262	0.040	25808	106276
Reef (Optiz, 1996)	210	2065	0.046	12036	44100
St. Marks (Christian & Luczkovich, 1999)	116	1128	0.083	5598	13456
Weddell Sea (Jacob, 2005)	381	10182	0.070	53204	145161
Ythan Estuary (Cohen *et al.*, 2009)	77	307	0.051	1749	5929

Table 1: Properties of the six marine food webs used in this analysis. Number of species (S), number of trophic interactions *or* directed edges *or* 1s in adjacency matrix (U) and connectance ($C = \frac{U}{S^2}$); total length $\mathfrak{L}_M(A) = -\log_2 \text{NML}(A|M)$ (Eqn 17) for random graph model, $\mathfrak{L}_{\text{rnd}}$, and uncompressed adjacency matrix, $\mathfrak{L}_{\text{raw}}$.

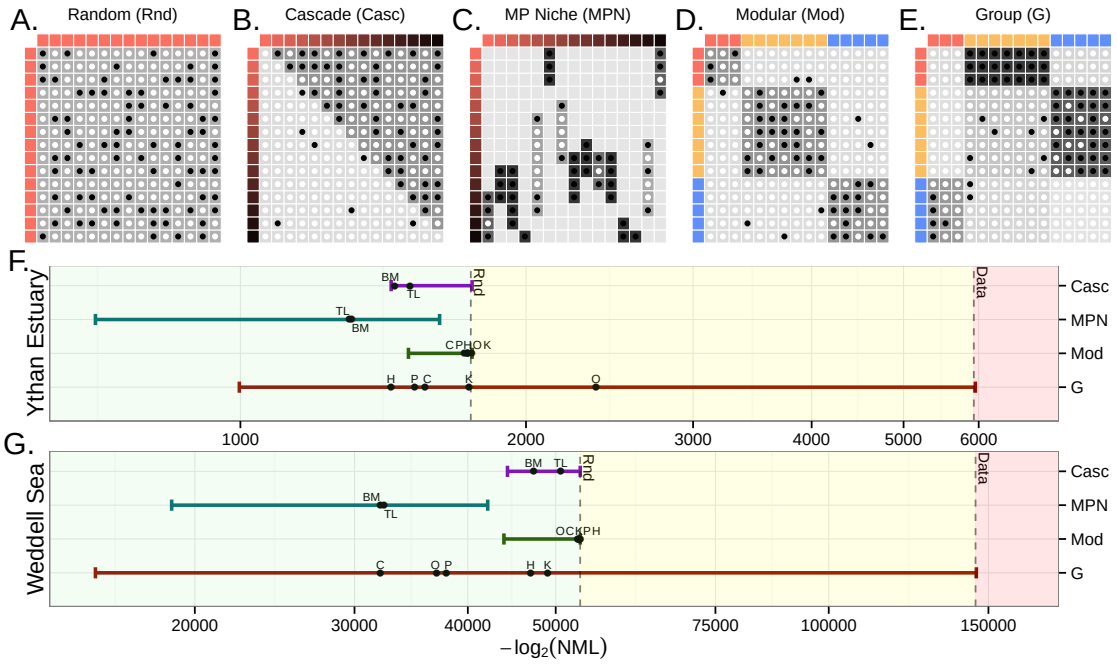

Figure 1: Food web models and total length. **Panels A to E:** Examples of five model families constructed in terms of random-graph-like matrix slices. Black dots represent realised trophic interactions between species (rows and columns), and dark grey regions delineate slices, with darker greys indicating greater density of interactions within slices. Erdős-Rényi random graph (panel A) has a single probability of a feeding interaction, and therefore the entire adjacency matrix represents a single slice. Cascade model (panel B) is defined by a hierarchy (ordering of species) and two slices: the upper-triangular slice represents feeding on lower-ranked species and the lower-triangular slice represents cannibalism and feeding on higher-ranked species. Minimum potential niche model (panel C) is defined by a species hierarchy and $2S$ slices: for each species, one slice represents its feeding interval (which contains all of its prey) and the other represents its non-feeding interval (with associated probability of feeding equal to zero). Modular model (panel D) is defined by a partition of species into modules and two slices: all square blocks on the diagonal (within-module interactions) and the remaining matrix elements (between-module interactions). Group model (panel E) is defined by a partition of species into γ groups and γ^2 slices; the probability that a species feeds on another species depends exclusively on the groups to which each species belongs (for a total of γ^2 distinct probabilities). **Panels F and G:** Model fit to two empirical food webs using total length based on NML. For each combination of model family and food web, we searched for the hierarchy or partition of species that resulted in the shortest (best fit) and longest (worst fit) total length (all other models are necessarily contained in this range). Vertical dashed lines mark two reference points: the total length of a random graph model and the uncompressed adjacency matrix (Table 1). Dots show the total length of models in which empirical data are used to define hierarchies (BM: body mass; TL: trophic level) and partitions (H: habitat; and taxonomic information, K: Kingdom; P: Phylum; C: Class; O: Order).

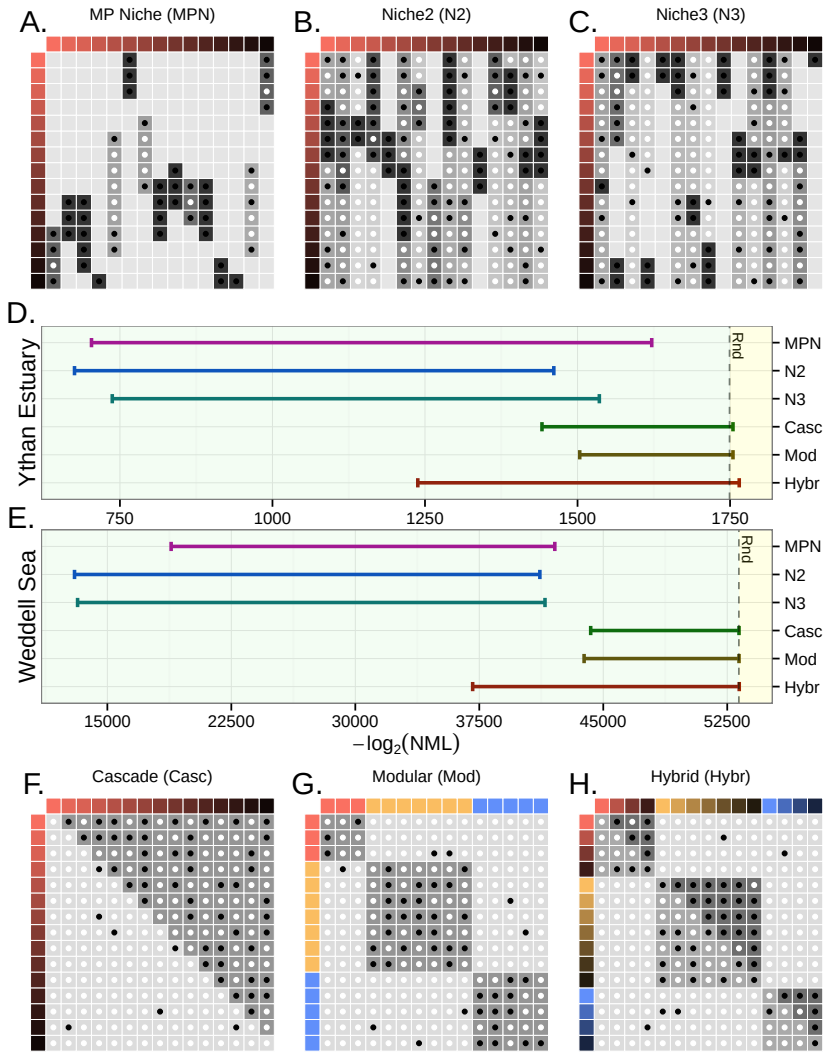

Figure 2: Developing new food web models. We relaxed the feeding constraint in the MPN model family (panel A, description in Fig. 1 caption) to design a more flexible N2 model family (panel B) that allows each species to feed outside its primary feeding interval with non-zero probability. N3 model family (panel C) relaxes feeding constraints further by specifying three feeding intervals for each species: primary, above primary and below primary. N2 models consistently performed better (shorter total lengths) than MPN and N3 models (panels D and E), while the improved fit of N3 models often could not overcome their additional model complexity (panel E). We combined cascade (panel F) and modular (panel G) model families to form the hybrid model family (panel H). In a hybrid model, species are partitioned into modules, and species within the same module form an independent hierarchy. Hybrid models performed much better than models from the two original model families (albeit worse than the better niche models, panels D and E).

Supporting Information

Selecting Food Web Models using Normalised Maximum Likelihood

Phillip P.A. Staniczenko[1,2,*], Matthew J. Smith[2] & Stefano Allesina[2,3]

[1]Centre for Biodiversity and Environment Research, University College London, UK (current address)

[2]Department of Ecology & Evolution, University of Chicago, Chicago, IL, USA

[3]Computation Institute, University of Chicago, Chicago, IL, USA

[*]Corresponding author: p.staniczenko@ucl.ac.uk

Abstract

We begin with a brief overview of AIC, BIC and Bayes factors for model selection. This is followed by a short introduction to the Minimum Description Length Principle and Normalised Maximum Likelihood (NML), including a derivation of the identity for the NML penalisation constant that permits its rapid computation. We add to the expressions of maximum likelihood and total length for random graph and cascade models in the main text with expressions for six additional model families: minimum potential niche, the two niche model variations introduced in the Model Development section of the main text, modular, group and hybrid (combination of modular and cascade model families). Then we describe simulation methods and results comparing AIC, BIC, Bayes factors and NML as means to recover information on species partitions in food webs generated by a known model. Finally, we present results for all models for the complete set of six marine food webs, as well as the table referenced in the main text showing the ranking of empirically-determined models within each model family according to AIC, BIC and Bayes factors and NML.

Contents

For the convenience of the reader, some parts of the main text have been included verbatim in this document.

1 Model selection

Models with many parameters (or more flexible models) yield better likelihoods. AIC is usually used to balance differences in fit and complexity among models. Here we present a slightly extended description of AIC, BIC and Bayes factors.

We continue with the example of food webs and models for food web structure, and use the same notation as in the main text: A food web can be represented by an adjacency matrix A in which a non-zero element A_{ij} indicates a feeding interaction between a consumer species j and a resource species i. We write the likelihood that a model M with vector of parameters $\boldsymbol{\theta}$ reproduces a given food web as $L(A|M, \boldsymbol{\theta})$. The parameter values that maximise the likelihood function are referred to as the maximum likelihood estimates $\hat{\boldsymbol{\theta}}$ and the corresponding likelihood is known as the maximum likelihood $\hat{L}(A|M, \hat{\boldsymbol{\theta}})$. We write the maximum log-likelihood $\ln \hat{L} = \hat{\mathcal{L}}_e$; where the subscript indicates the base of the logarithm.

1.1 AIC and BIC

AIC measures the asymptotic loss of information when a model is used to describe data, formally, it is an asymptotic approximation to the Kullback-Leibler divergence [1]. It is defined in terms of the maximum log-likelihood:

$$\mathrm{AIC}(A, M, \hat{\boldsymbol{\theta}}) = 2k - 2\hat{\mathcal{L}}_e(A|M, \hat{\boldsymbol{\theta}}); \tag{S1}$$

where k is the number of parameters in the model. Model fit (maximum log-likelihood) is balanced against model complexity by assigning a penalisation of one point of log-likelihood to each parameter. In this way, model complexity in AIC is measured by the number of parameters.

BIC [2] is similar to AIC but uses a different correction for model complexity:

$$\mathrm{BIC}(A, M, \hat{\boldsymbol{\theta}}) = k \ln(S^2) - 2\hat{\mathcal{L}}_e(A|M, \hat{\boldsymbol{\theta}}); \tag{S2}$$

where the penalisation for each parameter is now proportional to the logarithm of the amount of the data being fit. For the random graph,

$$\mathrm{AIC}(A, \mathrm{Rnd}, \hat{p}) = 2 - 2(U \ln \hat{p} + Z \ln(1 - \hat{p})) \tag{S3}$$

and

$$\mathrm{BIC}(A, \mathrm{Rnd}, \hat{p}) = \ln(S^2) - 2(U \ln \hat{p} + Z \ln(1 - \hat{p})). \tag{S4}$$

AIC and BIC are simple to compute but have two main drawbacks: they only hold asymptotically (i.e., for large amounts of data), and parameters that have little influence on the likelihood are penalised by exactly the same amount as those that strongly influence the likelihood [1]. Additionally, the seemingly straightforward task of counting the number of parameters in a model can, in practice, be very difficult when parameters are not numbers but more complex structures such as partitions or permutations.

1.2 Bayes factors

Bayes factors [3] are derived from Bayes' theorem. They measure the probability of a model given data. Following Bayes' theorem, the posterior probability of a model given data is

$$P(M_1|A) = \frac{P(A|M_1)P(M_1)}{P(A)}; \tag{S5}$$

where $P(M_1)$ is a prior for the model, and $P(A)$ is the probability of the data. For a model selection problem in which we have to choose between two models, and with no *a priori* preference for either model (i.e., $P(M_1) = P(M_2)$), the plausibility of the two different models is assessed by the Bayes factor, which is defined as the ratio of the two posterior probabilities:

$$K(M_1, M_2) = \frac{P(M_1|A)}{P(M_2|A)} = \frac{P(A|M_1)}{P(A|M_2)}. \tag{S6}$$

The term $P(A|M_1)$ is a marginal likelihood and does not depend on any single set of parameters. This is because the expression for marginal likelihood integrates over all parameters in the model. Given the choice of two models, the one with the highest marginal likelihood should be preferred as it offers a better balance between goodness-of-fit and complexity. Bayes factor penalisation for model complexity is not explicit, but is done automatically during the integration over possible parameter values. In fact, the marginal likelihood can be interpreted as the expected likelihood when parameterising the model by randomly sampling parameter values from their priors. Formally, the marginal likelihood is written

$$P(A|M_1) = \int_{\theta_1} \int_{\theta_2} \cdots \int_{\theta_k} L(A|M_1, \boldsymbol{\theta}) P(\boldsymbol{\theta}|M_1)\, d\theta_k \cdots d\theta_2\, d\theta_1; \tag{S7}$$

where $P(\boldsymbol{\theta}|M_1)$ is the probability of a given parameterisation when sampling parameters from their prior distributions.

It is straightforward to integrate over parameters if a suitable prior is chosen. For example, consider the random graph model introduced in the main text and choose a beta distribution, $B(\alpha, \beta)$, with hyper-parameters α and β for the prior distribution of p. The marginal likelihood in this case is

$$P(A|\text{Rnd}) = \int_0^1 \left(p^U (1-p)^Z\right) \left(\frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha,\beta)}\right) dp = \frac{B(U+\alpha, Z+\beta)}{B(\alpha,\beta)}; \tag{S8}$$

where the first term in the integral is the likelihood and the second term is the prior distribution for p.

Marginal likelihood and Bayes factors have been used to evaluate food web models [4, 5]. The main drawback is the requirement to specify priors. Furthermore, integrating over parameters can be very difficult for complex models, especially those involving discrete parameters or combinatorial structures such as permutations or partitions (as is the case with models for food web structure).

2 Minimum Description Length Principle

We begin this section with a brief introduction to the MDL Principle, followed by a recap of NML, and end by deriving the identity needed to rapidly compute NML penalisation for model complexity.

2.1 Background

Model selection is considered a problem of data compression in the MDL approach [6–11]. Data has a given length in bits of information, and better models are able to compress data more than worse models. In the simplest application of MDL, if we wanted to transmit a finite-sized amount of data over a channel such as the Internet, we would like to choose a model that minimises the total length $\mathfrak{L}_M(A) = \mathfrak{L}(A|M) + \mathfrak{L}(M)$; where $\mathfrak{L}(A|M)$ is the length of the original data after being encoded by the model and $\mathfrak{L}(M)$ is the length required to describe the model. Given these two pieces of information, the transmitted message can be decoded by the receiver to obtain the original data.

We first determine the length of uncompressed data to see the potential benefit offered by compression. MDL is particularly suited for discrete structures such as the adjacency matrix of a food web. This is because an adjacency matrix can be seen as a message written in an alphabet composed of just two symbols: 1 and 0, the presence and absence, respectively, of an interaction. For a food web comprising S species, the message to be transmitted is S^2 symbols long. Because the size of a food-web alphabet is two and we only need one bit of information for each symbol, the length of the uncompressed data is $\mathfrak{L}(A) = S^2$ bits.

The very idea of specifying a length for data may seem strange when dealing with continuous variables. For example, encoding $\pi = 3.14159\ldots$ would require an infinitely long message. However, all measurements in science are done with limited precision (especially in the computer age), such

that one can treat any value as a discrete quantity, and therefore encode it in a finite-sized message. For example, real numbers are typically encoded in computers as binary strings of 32 or 64 bits. In the case of food webs, the length of data is intuitive, and we will not discuss the topic any further.

A model can be used to encode, compress and transmit data as a shorter message compared to the uncompressed length. The Kraft inequality specifies how long the compressed message should be given the best possible encoding (model) [11]. For a probability distribution $\mathcal{P}(x)$ that describes the probability of a given symbol in the original data, the Kraft inequality states that there exists a prefix code (a code that can be uniquely decoded, see example in main text) that encodes the (compressed) message as a string of $-\lceil \log_2 \mathcal{P}(x) \rceil$ bits; where the symbol $\lceil a \rceil$ means the smallest integer greater than or equal to a. Henceforth, we assume the existence of fractionary bits (a common assumption in Information Theory) and simply write $-\log_2 \mathcal{P}(x)$. The Kraft inequality provides a connection between the length of the compressed message and likelihoods. It implies that we can use a model M to compress an adjacency matrix from $\mathfrak{L}(A) = S^2$ bits to $\mathfrak{L}(A|M) = -\log_2 \hat{L}(A|M, \hat{\boldsymbol{\theta}}) = -\hat{\mathcal{L}}_2(A|M, \hat{\boldsymbol{\theta}})$ bits.

But to correctly decode the compressed message, the receiver requires a description of the model $\mathfrak{L}(M)$ in addition to $\mathfrak{L}(A|M)$. Unfortunately, $\mathfrak{L}(M)$ is often difficult to compute, which has limited the applicability of MDL to real-world problems [12].

2.2 Normalised Maximum Likelihood

As introduced in the main text, NML quantifies how well a model explains a particular data set compared to how well it explains data in general [8,11–13]. The NML distribution of a particular data set A given a model M is

$$\mathrm{NML}(A|M) = \frac{\hat{L}(A|M, \hat{\boldsymbol{\theta}}_A)}{\int \hat{L}(A'|M, \hat{\boldsymbol{\theta}}_{A'})\, dA'}; \tag{S9}$$

where the normalisation is over all data sets A' with the same number of data-points as the original data set. NML returns values in the range [0,1].

A complex model with many parameters will typically fit many data sets well because of its flexibility. This outcome would result in a large denominator and thus a small value for NML (as would an overly simple model that fits *all* data sets the same amount). On the other hand, a model that fits only observed data well and all other data sets poorly would result in a large value for NML. An analogy is often made to the problem of fitting a polynomial function (model) to a sequence of data consisting of n pairs $(x,\ y)$, where x and y are real numbers [10]. The classical solution to this problem is to perform a standard linear regression, which results in a "best-fit" line that captures *some* of the regularity in the data, but often appears to *underfit* the data. The other extreme solution is to pick a polynomial of degree $n-1$ that goes *exactly* through all the n points being fit. In doing so, there is a large risk of *overfitting*. Instead, we might prefer an intermediate-degree polynomial: one that permits small (but non-zero) error and is still relatively simple (i.e., has few parameters). It is with similar intent that NML quantifies the trade-off between model complexity and goodness-of-fit.

When data are discrete, as with food webs, the integral is replaced by a summation. The total length associated with NML is given by

$$\begin{aligned}\mathfrak{L}_M(A) &= -\log_2 \mathrm{NML}(A|M) \\ &= -\log_2 \hat{L}(A|M,\hat{\boldsymbol{\theta}}_A) + \log_2 \sum_{A'} \hat{L}(A'|M,\hat{\boldsymbol{\theta}}_{A'}) \\ &= -\hat{\mathcal{L}}_2(A|M,\hat{\boldsymbol{\theta}}_A) + \log_2 \mathcal{C}(M,A);\end{aligned} \tag{S10}$$

where $\mathcal{C}(M,A)$ is a penalisation constant (known as the parametric complexity in the Information Theory literature [7, 14–16]).

2.3 Derivation of NML penalisation constant

The penalisation constant for models built from random-graph-like matrix slices can be computed very quickly using a new identity that we now derive.

Suppose that we have a slice of length $X = U+Z$ containing U ones and Z zeros. Each element in the slice is assumed to be the result of a Bernoulli trial in which the probability of obtaining a one is set to its maximum likelihood estimate

$$\hat{p} = \frac{U}{X}. \tag{S11}$$

The maximum likelihood for a slice is then

$$\hat{L}(U, Z|X) = \hat{p}^U (1-\hat{p})^Z; \tag{S12}$$

and the normalised maximum likelihood is

$$\text{NML} = \frac{\hat{L}(U, Z|X)}{\sum_{k=0}^{X} \binom{X}{k} \left(\frac{k}{X}\right)^k \left(\frac{X-k}{X}\right)^{(X-k)}}; \tag{S13}$$

where the denominator specifies the sum of maximum likelihoods over all possible numbers of ones in the matrix slice (the first term represents a weighting for the number of configurations involving k ones).

We now show how the denominator can be written in a form that only depends on the length of a matrix slice:

$$\mathcal{C}(X) = \sum_{k=0}^{X} \binom{X}{k} \left(\frac{k}{X}\right)^k \left(\frac{X-k}{X}\right)^{X-k} \tag{S14}$$

$$= 2 + \sum_{k=1}^{X-1} \binom{X}{k} \left(\frac{k}{X}\right)^k \left(\frac{X-k}{X}\right)^{X-k} \tag{S15}$$

$$= 2 + \frac{1}{X^X} \sum_{k=1}^{X-1} \binom{X}{k} k^k (X-k)^{X-k}. \tag{S16}$$

Now define

$$\mathcal{A}(X) = \sum_{k=1}^{X-1} \binom{X}{k} k^k (X-k)^{X-k}.$$

Riordan & Sloane [17] showed that

$$\mathcal{B}(X) = \sum_{k=0}^{X-2} \frac{X^k}{k!} = \frac{\mathcal{A}(X)}{X!}. \tag{S17}$$

Hence,

$$\mathcal{C}(X) = 2 + \frac{(X-1)! \sum_{k=0}^{X-2} \frac{X^k}{k!}}{X^{X-1}}; \tag{S18}$$

which, given that $\Gamma(X-1, X) = (X-2)! e^{-X} \sum_{k=0}^{X-2} \frac{X^k}{k!}$, becomes

$$\mathcal{C}(X) = 2 + \frac{e^X X(X-1)\Gamma(X-1, X)}{X^X}. \tag{S19}$$

Finally, because $\Gamma(X, X) = (X-1)\Gamma(X-1, X) + e^{-X} X^{X-1}$,

$$\boxed{\mathcal{C}(X) = 1 + \frac{e^X \Gamma(X, X)}{X^{X-1}}.} \tag{S20}$$

This expression can be efficiently computed using the expansion

$$\Gamma(X, X) \approx X^{X-1} e^{-X} \left(\sqrt{X \frac{\pi}{2}} - \frac{1}{3} + \frac{\sqrt{2\pi}}{24\sqrt{X}} - \frac{4}{135X} + \frac{\sqrt{2\pi}}{576\sqrt{X^3}} + \frac{8}{2835X^2} + \dots \right); \tag{S21}$$

to yield

$$\mathcal{C}(X) \approx 1 + \left(\sqrt{X \frac{\pi}{2}} - \frac{1}{3} + \frac{\sqrt{2\pi}}{24\sqrt{X}} - \frac{4}{135X} + \frac{\sqrt{2\pi}}{576\sqrt{X^3}} + \frac{8}{2835X^2} \right). \tag{S22}$$

A table of the first few exact and approximate values shows how good the approximation is:

X	Exact, Eq.(S15)	Exact, Eq.(S20)	Approx.	Approx., Eq.(S22)
1	2	2	2	1.999146
2	$\frac{5}{2}$	$\frac{5}{2}$	2.5	2.499696
3	$\frac{26}{9}$	$\frac{26}{9}$	2.88889	2.888731
4	$\frac{103}{32}$	$\frac{103}{32}$	3.21875	3.218656
5	$\frac{2194}{625}$	$\frac{2194}{625}$	3.5104	3.510334
6	$\frac{1223}{324}$	$\frac{1223}{324}$	3.774691	3.774643
7	$\frac{472730}{117649}$	$\frac{472730}{117649}$	4.018139	4.018102
8	$\frac{556403}{131072}$	$\frac{556403}{131072}$	4.245018	4.244989

The penalisation for each slice can be multiplied (or added in log-space) to obtain the penalisation constant for models that comprise more than one random-graph-like matrix slice.

In the main text, we wrote the total length associated with NML for a random graph as

$$\mathfrak{L}_{\mathrm{Rnd}}(A) = -\hat{\mathcal{L}}_2(A|\mathrm{Rnd}, \hat{p}) + \log_2 \mathcal{C}(\mathrm{Rnd}, A); \tag{S23}$$

but as the random graph model is analogous to the entire adjacency matrix being a single slice, the penalisation constant can now be simplified such that the total length becomes

$$\mathfrak{L}_{\mathrm{Rnd}}(A) = -\hat{\mathcal{L}}_2(A|\mathrm{Rnd}, \hat{p}) + \log_2 \mathcal{C}(S^2); \tag{S24}$$

where $\mathcal{C}(S^2)$ indicates that the penalisation depends only on the size of the matrix.

As a cascade model is essentially the composition of two half-random graphs, its total length based on NML can therefore be written

$$\mathfrak{L}_{\mathrm{Casc}_H}(A) = -\hat{\mathcal{L}}_2(A|\mathrm{Casc}_H, \hat{p}, \hat{q}) + \log_2 \mathcal{C}(X_u) + \log_2 \mathcal{C}(X_l); \tag{S25}$$

where the constant $\mathcal{C}(X_u)$ depends only on the size of the upper triangular part of the matrix and $\mathcal{C}(X_l)$ on the diagonal and lower triangular part. (For food webs, $X_u = \frac{S(S-1)}{2}$ and $X_l = S + \frac{S(S-1)}{2} = \frac{S(S+1)}{2}$.)

3 Food web models

In the main text, we described the random graph and cascade model family, and showed how to calculate maximum likelihood and total length based on NML for these models. Here we describe six additional model families: minimum potential niche, the two niche model variations introduced in the *Model Development* section of the main text, modular, group and hybrid (combination of modular and cascade model families). Each family is motivated by ecological features that specify how random-graph-like matrix slices are defined and combined in the model. The use of slices makes the calculation of NML a simple extension of the methods given for the random graph and cascade model examples.

3.1 Niche Model

3.1.1 Minimum Potential Niche Model

The minimum potential niche (MPN) model family [18] is a variation on the niche model [19] that focuses on its central idea: intervality. In a MPN model, species are ordered in a hierarchy H and each consumer has a restricted feeding interval of consecutive species which contains all of its prey. Each consumer i feeds on species within its interval with probability p_i and on

species outside its interval with probability $q_i = 0$. A MPN model divides an adjacency matrix into $2S$ slices: for each consumer, one slice represents its feeding interval (with associated probability p_i) and the other slice represents its non-feeding interval (with associated probability $q_i = 0$).

The maximum likelihood for a MPN model with a given H is

$$\hat{L}(A|\mathrm{MPN}_H, \hat{p}_i, \hat{q}_i = 0) = \prod_i \hat{p}_i^{U_{i,1}} (1 - \hat{p}_i)^{Z_{i,1}}, \tag{S26}$$

and the maximum log-likelihood is

$$\hat{\mathcal{L}}_e(A|\mathrm{MPN}_H, \hat{p}_i, \hat{q}_i = 0) = \sum_i \left(U_{i,1} \ln \hat{p}_i + Z_{i,1} \ln(1 - \hat{p}_i) \right); \tag{S27}$$

where $U_{i,1}$ and $Z_{i,1}$ are the number of ones and zeros, respectively, in the slice associated with the feeding interval of consumer i, and the consumer's feeding probability is set to its maximum likelihood estimate $\hat{p}_i = U_{i,1}/(U_{i,1} + Z_{i,1})$.

Total length based on NML can be computed by factoring the penalisation constant into the contribution from each slice (as with cascade models):

$$\mathfrak{L}_{\mathrm{MPN}_H}(A) = -\hat{\mathcal{L}}_2(A|\mathrm{MPN}_H, \hat{p}_i, \hat{q}_i = 0) + \sum_i \log_2 \mathcal{C}(X_{i,1}); \tag{S28}$$

where the penalisation constant $\mathcal{C}(X_{i,1})$ depends only on the size $X_{i,1} = U_{i,1} + Z_{i,1}$ of the slice associated with the feeding interval of consumer i.

3.1.2 Niche2

We can relax the constraint on feeding in the MPN model to design a more flexible model family that is inspired by the probabilistic niche model [20] which we call Niche2 (N2). A consumer's feeding interval no longer has to contain all of its prey items: each consumer preys on species within its interval with probability p_i and on species outside of its interval with probability q_i. The N2 model family includes all MPN models. The maximum likelihood for a given H is

$$\hat{L}(A|\mathrm{N2}_H, \hat{p}_i, \hat{q}_i) = \prod_i \hat{p}_i^{U_{i,1}} (1 - \hat{p}_i)^{Z_{i,1}} \hat{q}_i^{U_{i,2}} (1 - \hat{q}_i)^{Z_{i,2}}, \tag{S29}$$

and the maximum log-likelihood is

$$\hat{\mathcal{L}}_e(A|\mathrm{N2}_H, \hat{p}_i, \hat{q}_i) = \sum_i \left(U_{i,1} \ln \hat{p}_i + Z_{i,1} \ln(1-\hat{p}_i) + U_{i,2} \ln \hat{q}_i + Z_{i,2} \ln(1-\hat{q}_i)\right); \tag{S30}$$

where $U_{i,1}$ and $Z_{i,1}$ are the number of ones and zeros, respectively, in the slice associated with the feeding interval of consumer i, $U_{i,2}$ and $Z_{i,2}$ with the consumer's non-feeding interval, and the two feeding probabilities (for each consumer) are set to their maximum likelihood estimates: $\hat{p}_i = U_{i,1}/(U_{i,1} + Z_{i,1})$ and $\hat{q}_i = U_{i,2}/(U_{i,2} + Z_{i,2})$.

Total length based on NML for N2 requires an extra term compared to an MPN model to take into account the size of the non-feeding interval for each consumer:

$$\mathfrak{L}_{\mathrm{N2}_H}(A) = -\hat{\mathcal{L}}_2(A|\mathrm{N2}_H, \hat{p}_i, \hat{q}_i) + \sum_i \log_2 \mathcal{C}(X_{i,1}) + \sum_i \log_2 \mathcal{C}(X_{i,2}); \tag{S31}$$

where the penalisation constant $\mathcal{C}(X_{i,1})$ depends only on the size $X_{i,1} = U_{i,1} + Z_{i,1}$ of the slice associated with the feeding interval of consumer i and $C(X_{i,2})$ on the size $X_{i,2} = U_{i,2} + Z_{i,2}$ of the slice associated with the consumer's non-feeding interval.

3.1.3 Niche3

A third model family, which we call Niche3 (N3), is inspired by the generalised niche model [18, 21] and relaxes feeding constraints even further compared to MPN and N2. Each consumer feeds on species within its interval with probability p_i, on species above its interval with probability q_i and on species below its interval with probability r_i. N3 expressions for maximum likelihood and total length are trivial extensions to those for N2 (only additional terms for r_i must be included), so are not provided. Although the N3 model family includes all N2 and MPN models, the penalisation owing to model complexity will always be higher for N3 models because of the additional probability r_i required for each species.

3.2 Modular Model

The modular model family is based on the presence of compartments or modules in ecology [22–26]. Modules are often associated with different local habitats or seasons, and species within the same module are expected to have a higher probability of interacting with one another compared to two species in different modules. A modular model divides species into a set partition Π (i.e., each species is assigned to only one module and therefore modules are non-overlapping); two species in the same module interact with probability p, while species in different modules interact with probability q. As with cascade models, each partition Π divides an adjacency matrix into two slices: one composed of all the square blocks on the diagonal (within-module interactions) and one composed of all other matrix elements (between-module interactions).

The maximum likelihood for a modular model is formally similar to that of a cascade model but is defined by a partition:

$$\hat{L}(A|\mathrm{Mod}_\Pi, \hat{p}, \hat{q}) = \hat{p}^{U_w}(1-\hat{p})^{Z_w}\hat{q}^{U_b}(1-\hat{q})^{Z_b}, \tag{S32}$$

and the maximum log-likelihood is

$$\hat{\mathcal{L}}_e(A|\mathrm{Mod}_\Pi, \hat{p}, \hat{q}) = U_w \ln \hat{p} + Z_w \ln(1-\hat{p}) + U_b \ln \hat{q} + Z_b \ln(1-\hat{q}); \tag{S33}$$

where Π determines how many ones (U_w) and zeros (Z_w) are in the matrix slice representing within-module interactions and how many (U_b, Z_b) are in the matrix slice representing between-module interactions, which in turn specifies the maximum likelihood estimates $\hat{p} = U_w/(U_w + Z_w)$ and $\hat{q} = U_b/(U_b + Z_b)$.

Total length based on NML can be computed by factoring the penalisation constant with respect to the contribution from each slice:

$$\mathfrak{L}_{\mathrm{Mod}_\Pi}(A) = -\hat{\mathcal{L}}_2(A|\mathrm{Mod}_\Pi, \hat{p}, \hat{q}) + \log_2 \mathcal{C}(X_w) + \log_2 \mathcal{C}(X_b); \tag{S34}$$

where the penalisation constant $\mathcal{C}(X_w)$ depends only on the size $X_w = U_w + Z_w$ of the within-module slice and $\mathcal{C}(X_b)$ on the size $X_b = U_w + Z_b$ of the between-module slice.

3.3 Group Model

The group model family extends the concept of compartments introduced with the modular model family. A group model [23] (also known as a stochastic block model [27]) is also defined by a partition Π, which specifies to which of γ non-overlapping groups each species belongs. The probability that a consumer j preys on resource i depends exclusively on the corresponding groups of species i and j: $p_{ij} = p_{\Pi_i \Pi_j} = p_{kl}$; where k and l index groups. As such, each partition Π divides the adjacency matrix into γ^2 slices (and therefore there are a total of γ^2 probabilities).

The maximum likelihood for a group model is

$$\hat{L}(A|\mathrm{G}_\Pi, \hat{p}_{kl}) = \prod_{kl} \hat{p}_{kl}^{U_{kl}} (1 - \hat{p}_{kl})^{Z_{kl}}, \tag{S35}$$

and the maximum log-likelihood is

$$\hat{\mathcal{L}}_e(A|\mathrm{G}_\Pi, \hat{p}_{kl}) = \sum_{kl} \left(U_{kl} \ln \hat{p}_{kl} + Z_{kl} \ln(1 - \hat{p}_{kl}) \right); \tag{S36}$$

where the partition Π determines how many ones (U_{kl}) and zeros (Z_{kl}) are in the matrix slice representing interactions between groups k and l, which in turn specifies the maximum likelihood estimate $\hat{p}_{kl} = U_{kl}/(U_{kl} + Z_{kl})$.

Total length based on NML is

$$\mathfrak{L}_{\mathrm{G}_\Pi}(A) = -\hat{\mathcal{L}}_2(A|\mathrm{G}_\Pi, \hat{p}_{kl}) + \sum_{kl} \log_2 \mathcal{C}(X_{kl}); \tag{S37}$$

where the penalisation constant $\mathcal{C}(X_{kl})$ depends only on the size $X_{kl} = U_{kl} + Z_{kl}$ of the slice associated with interactions between groups k and l.

An interesting set of group models are those with exactly S groups (where each species is in its own group), which result in total lengths that are equal

to that of the uncompressed adjacency matrix. Because each species belongs to a group of its own, each feeding probability (of which there are $\gamma^2 = S^2$) is equal to either 1 or 0 (depending on whether there is an interaction or not, respectively), leading to a maximum likelihood that is always equal to 1. Each matrix slice is only one matrix element in size, so the total length based on NML is

$$\mathfrak{L}_{\mathrm{G}_{\Pi_S}}(A) = -S^2 \log_2 1 + S^2 \log_2 \mathcal{C}(X = 1) = 0 + S^2 \log_2 2 = S^2; \qquad \text{(S38)}$$

which is the the same total length as naïvely transmitting the adjacency matrix.

3.4 Hybrid Model

The hybrid model family is a combination of the cascade and modular families. In a hybrid model, species are partitioned into modules, and species within the same module form an independent hierarchy. A partition Π divides species into k modules, and for each module, a hierarchy H_k dictates two feeding probabilities as in a cascade model: each species has a probability p_k of feeding on species that are below it in the hierarchy and a probability q_k of being cannibalistic or feeding on higher-ranked species. Feeding between modules takes place with a single probability r (as in a modular model). The total number of matrix slices therefore equals $2k + 1$.

The maximum likelihood for a hybrid model is

$$\hat{L}(A|\mathrm{Hybr}_{\Pi,H_k}, \hat{p}_k, \hat{q}_k, \hat{r}) = \hat{r}^{U_b}(1-\hat{r})^{Z_b} \prod_k \hat{p}_k^{U_{k,1}}(1-\hat{p}_k)^{Z_{k,1}} \hat{q}_k^{U_{k,2}}(1-\hat{q}_k)^{Z_{k,2}}, \qquad \text{(S39)}$$

and the maximum log-likelihood is

$$\begin{aligned}\hat{\mathcal{L}}_e(A|\mathrm{Hybr}_{\Pi,H_k}, \hat{p}_k, \hat{q}_k, \hat{r}) &= U_b \ln \hat{r} + Z_b \ln(1-\hat{r}) \\ &+ \sum_k \left(U_{k,1} \ln \hat{p}_k + Z_{k,1} \ln(1-\hat{p}_k) + U_{k,2} \ln \hat{q}_k + Z_{k,2} \ln(1-\hat{q}_k)\right); \end{aligned} \qquad \text{(S40)}$$

where Π determines how many ones (U_b) and zeros (Z_b) are in the matrix slice representing between-module interactions (with associated maximum likelihood estimate $\hat{r} = U_b/(U_b + Z_b)$) and how many are in the upper-triangular ($U_{k,1}$, $Z_{k,1}$) and lower-triangular ($U_{k,2}$, $Z_{k,2}$) parts of each module k (with associated maximum likelihood estimates $\hat{p}_k = U_{k,1}/(U_{k,1} + Z_{k,1})$ and $\hat{q}_k = U_{k,2}/(U_{k,2} + Z_{k,2})$, respectively).

Total length based on NML is

$$\begin{aligned}\mathfrak{L}_{\mathrm{Hybr}_{\Pi,H_k}}(A) = &- \hat{\mathcal{L}}_2(A|\mathrm{Hybr}_{\Pi,H_k}, \hat{p}_k, \hat{q}_k, \hat{r}) \\ &+ \log_2 \mathcal{C}(X_b) + \sum_k \left(\log_2 \mathcal{C}(X_{k,1}) + \log_2 \mathcal{C}(X_{k,2})\right); \end{aligned} \tag{S41}$$

where the penalisation constant $\mathcal{C}(X_b)$ depends only on the size $X_b = U_b + Z_b$ of the slice associated with between-module interactions, $\mathcal{C}(X_{k,1})$ with the size $X_{k,1} = U_{k,1} + Z_{k,1}$ and $\mathcal{C}(X_{k,2})$ with the size $X_{k,2} = U_{k,2} + Z_{k,2}$ of slices associated with upper-triangular and lower-triangular within-module interactions, respectively. Even with its increased complexity, hybrid models performed much better (shorter total lengths) than models from the two original model families (Fig. 2 and Fig. S5).

4 Simulation Methods and Results

In this section we test whether AIC, BIC, Bayes factors and NML can be used to recover information on species partitions in food webs generated by a known model.

4.1 Methods

We generated a set of 100 random adjacency matrices from the group model family. Each matrix had dimension 100 x 100 and we partitioned the 100 species into five groups of different size by randomly sampling four break-points. The $5^2 = 25$ probabilities of connection between groups in each matrix, p_{kl} (see section on the group model, above), were drawn from a beta

distribution $B(\alpha = 0.5, \beta = 0.5)$. (We also repeated analysis with beta distribution $B(\alpha = 0.7, \beta = 0.7)$.) Each element of an adjacency matrix (whether there is an interaction between species i and j) was then filled by sampling from a Bernouilli distribution using the appropriate p_{kl}. The procedure results in a particular configuration of edges generated by a group model and a known partition into five groups (specified by the breakpoints described above). We refer to this partition as the *true* partition.

Presented with one of these generated adjacency matrices, we used a stochastic optimisation algorithm (see [28] for details) to search for the partition of species that maximised the log-likelihood (Eqn S36); we searched for the best partition into one group, two groups, and so on up to ten groups. These ten partitions, along with the true partition, are analogous to empirical partitions described in the main text, insofar as they are known before calculating AIC, BIC, Bayes factors and NML (rather than defining a partition by, say, taxonomy, we obtained our set of partitions to be assessed by the four measures by maximising log-likelihood).

For each matrix and its 11 best partitions (one to ten groups and the true partition), we calculated the corresponding values for AIC, BIC, Bayes factors and NML. When calculating the value for Bayes factors we used hyperpriors matching the generating distribution, i.e., matching either $B(\alpha = 0.5, \beta = 0.5)$ or $B(\alpha = 0.7, \beta = 0.7)$, as appropriate. We recorded which of the 11 partitions provided the best score for each measure (and log-likelihood) across the 100 matrices (Tables S2 and S3). This enabled us to compare how well each measure recovered information on the generating structure of a food web. Measures perform well if they return a large fraction of best scores for true partitions and partitions into around five groups (recall that true partitions are composed of five groups).

4.2 Results

Log-likelihood favoured partitions into ten groups for all 100 matrices generated with $B(\alpha = 0.5, \beta = 0.5)$, despite true partitions being composed of only five groups (Table S2). AIC also overwhelmingly favoured ten groups for explaining the structure of the 100 generated food webs. BIC performed much better and favoured the true partition for 64 matrices, the best partition into six groups for 33 matrices and seven groups for three matrices. Bayes factors and NML also performed well: favouring true partitions and the best partitions into six and seven groups. Results for Bayes factors and NML were comparable. Indeed, we expect results for Bayes factors (with hyper-priors matching the generating beta distribution) and NML to converge in the limit of infinite data [11]. We observed qualitatively similar results across measures when using $B(\alpha = 0.7, \beta = 0.7)$ to determine interaction probabilities (Table S3).

4.3 Discussion

Model selection using log-likelihood unsurprisingly favoured the maximum number of groups considered in this exercise. AIC, with its limited penalisation for model complexity, also favoured a large number of groups (supporting the observation of its tendency to overfit). BIC, Bayes factors and NML all performed well at recovering information on species partitions in food webs generated by a known model.

The true partition for a given matrix typically had the highest log-likelihood out of the set of partitions into five groups. With BIC, the large penalisation for model complexity was such that the true partition was often favoured over partitions into six (or seven) groups, despite their higher likelihood. (We expect the large penalisation associated with BIC to result in much worse performance if the real number of groups is much larger than five.)

Unlike BIC, where the penalisation term is the same for all partitions with the same number of groups, complexity penalisation for Bayes factors

and NML can vary between partitions with the same number of groups. This resulted in Bayes factors and NML often favouring partitions other than the true partition (although with number of groups similar to that of the true partition). The results for Bayes factors and NML are very similar, yet it is worth noting that NML appears to reflect information on the matrices' generating distribution without the need for explicit statement (in the form of (hyper-)priors in the case of Bayes factors).

5 Additional Results

Here we present complete results for six marine food webs (Table S1) and seven models for food-web structure: cascade, MPN, N2, N3, modular, group and hybrid. We obtained total length ranges for model families using a stochastic optimisation algorithm (see [28] for details). For each combination of model family and food web, we searched for the hierarchy (cascade and niche) or partition (modular and group) of species that resulted in the shortest (best) and longest (worst) total lengths (Fig. S1). All other models in a family are necessarily contained in this range, which enables a quick indicative comparison of model family performance. The search involved trialling different species hierarchies (permutations) for cascade, MPN, N2 and N3 model families (Figs. S2 and S4); different species partitions and number of modules/groups for modular and group model families (Fig. S3); and different combinations of partitions and hierarchies for the hybrid model family (Fig. S5).

We also used empirical data on body mass and trophic level to determine species hierarchies in cascade and MPN models (Fig. S2), and data on taxonomic information (Kingdom, Phylum, Class and Order) and habitat to determine species partitions in modular and group models (Fig. S3).

The total lengths of a random graph and uncompressed data represent two helpful points of reference with which to compare more complex food

web models. As random graphs only take into account the number of species and the number of interactions between those species, we would expect models incorporating more ecological principles to return shorter total lengths than corresponding random graphs. Models with total lengths longer than random graphs should be treated with caution, and those with total lengths longer than uncompressed food web data are particularly poor descriptions of observed data.

In all cases, we assessed the performance of models and model families using the total length: $\mathfrak{L}_M(A) = -\log_2 \mathrm{NML}(A|M)$ (Eqn S10).

Also included below is Table S4, which shows the ranking of empirically-determined models within each model family according to AIC, BIC and Bayes factors and NML.

References

[1] Burnham K, Anderson D (2002) Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer, second edition.

[2] Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6: 461–464.

[3] Kass R, Raftery A (1995) Bayes factors. J Am Stat Assoc 90: 773–795.

[4] Baskerville E, Dobson A, Bedford T, Allesina S, Anderson T, et al. (2011) Spatial guilds in the serengeti food web revealed by a bayesian group model. PLoS Comput Biol 7: e1002321.

[5] Eklöf A, Helmus M, Moore M, Allesina S (2012) Relevance of evolutionary history for food web structure. P Roy Soc Lond B Bio 279: 1588–1596.

[6] Rissanen J (1978) Modeling by the shortest data description. Automatica 14: 465–471.

[7] Rissanen J (1989) Stochastic complexity in statistical inquiry. Singapore: World Scientic Publishing.

[8] Barron A, Rissanen J, Yu B (1998) The minimum description length principle in coding and modeling. In: Information Theory 50 Years of Discovery, Wiley USA, volume 44. pp. 699–716.

[9] Hansen A, Yu B (2001) Model selection and the principle of minimum description length. J Am Stat Assoc 96: 746–774.

[10] Grünwald P (2000) Model selection based on minimum description length. J Math Psychol 44: 133–152.

[11] Grünwald P (2007) The Minimum Description Length Principle. MIT Press.

[12] Myung J, Navarro D, Pitt M (2006) Model selection by normalized maximum likelihood. J Math Psychol 50: 167–179.

[13] Rissanen J (2001) Strong optimality of the normalized ml models as universal codes and information in data. IEEE T Inform Theory 47: 1712–1717.

[14] Rissanen J (1986) Stochastic complexity and modeling. Ann Stat 14: 1080–1100.

[15] Rissanen J (1987) Stochastic complexity. J Roy Stat Soc B 49: 223–239.

[16] Rissanen J (1996) Fisher information and stochastic complexity. IEEE T Inform Theory 42: 40–47.

[17] Riordan J, Sloane N (1969) The enumeration of rooted trees by total height. J Australian Math Soc 10: 278–282.

[18] Allesina S, Alonso D, Pascual M (2008) A general model for food web structure. Science 320: 658–661.

[19] Williams R, Martinez N (2000) Simple rules yield complex food webs. Nature 404: 180–183.

[20] Williams R, Anandanadesan A, Purves D (2010) The probabilistic niche model reveals the niche structure and role of body size in a complex food web. PLoS ONE 5: e12092.

[21] Stouffer D, Camacho J, Amaral L (2006) A robust measure of food web intervality. Proc Natl Acad Sci USA 103: 19015–19020.

[22] Krause A, Frank K, Mason D, Ulanowicz R, Taylor W (2003) Compartments revealed in food-web structure. Nature 426: 282–285.

[23] Allesina S, Pascual M (2009) Food web models: a plea for groups. Ecol Lett 12: 652–662.

[24] Rezende E, Albert E, Fortuna M, Bascompte J (2009) Compartments in a marine food web associated with phylogeny, body mass, and habitat structure. Ecol Lett 12: 779–788.

[25] Guimerà R, Stouffer D, Sales-Pardo M, Leicht E, Newman M, et al. (2010) Origin of compartmentalization in food webs. Ecology 91: 2941–2951.

[26] Stouffer D, Bascompte J (2011) Compartmentalization increases food-web persistence. Proc Natl Acad Sci USA 108: 3648–3652.

[27] Karrer B, Newman M (2011) Stochastic blockmodels and community structure in networks. Phys Rev E 83: 016107.

[28] Eklöf A, Jacob U, Kopp J, Bosch J, Castro-Urgal R, et al. (2013) The dimensionality of ecological networks. Ecol Lett 16: 577–583.

[29] Jacob U, Thierry A, Brose U, Arntz W, Berg S, et al. (2011) The role of body size in complex food webs: a cold case. Adv Ecol Res 45: 181–223.

[30] Riede J, Rall B, Banasek-Richter C, Navarrete S, Wieters E, et al. (2010) Scaling of food-web properties with diversity and complexity across ecosystems. Adv Ecol Res 42: 139–170.

[31] Optiz S (1996) Trophic interactions in caribbean coral reefs. Technical Report 43, ICLARM, Manila.

[32] Christian R, Luczkovich J (1999) Organizing and understanding a winters seagrass foodweb network through effective trophic levels. Ecol Model 117: 99–124.

[33] Jacob U (2005) Trophic Dynamics of Antarctic Shelf Ecosystems—Food Webs and Energy Flow Budgets. Ph.D. thesis, University of Bremen, Germany.

[34] Cohen J, Schittler D, Raffaelli D, Reuman D (2009) Food webs are more than the sum of their tritrophic parts. Proc Natl Acad Sci USA 106: 22335–22340.

Table and Figures

Food web	S	U	C	$\mathfrak{L}_{\text{rnd}}$	$\mathfrak{L}_{\text{raw}}$
Kongsfjorden [29]	252	1124	0.017	8157	63504
Lough Hyne [30]	326	4262	0.040	25808	106276
Reef [31]	210	2065	0.046	12036	44100
St. Marks [32]	116	1128	0.083	5598	13456
Weddell Sea [33]	381	10182	0.070	53204	145161
Ythan Estuary [34]	77	307	0.051	1749	5929

Table S1: Properties of the six marine food webs used in this analysis. Number of species (S), number of trophic interactions *or* directed edges *or* 1s in adjacency matrix (U) and connectance ($C = \frac{U}{S^2}$); total length $\mathfrak{L}_M(A) = -\log_2 \text{NML}(A|M)$ (Eqn S10) for random graph model, $\mathfrak{L}_{\text{rnd}}$, and uncompressed adjacency matrix, $\mathfrak{L}_{\text{raw}}$.

Groups	LL	AIC	BIC	BF	NML
True	0	0	64	32	29
1	0	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	0	0	0	0	0
6	0	0	33	52	54
7	0	0	3	14	15
8	0	0	0	1	1
9	0	1	0	1	1
10	100	99	0	0	0

Table S2: Group size resulting in the best score for 100 matrices generated using a known model and partition according to log-likelihood (LL), AIC, BIC, Bayes factors with hyper-priors matching the generating distribution (BF), and NML. Each matrix was generated using a group model with a different but known partition into five groups (True), with interaction probabilities drawn from a beta distribution $B(\alpha = 0.5, \beta = 0.5)$.

Groups	LL	AIC	BIC	BF (unif)	BF (match)	NML
True	0	0	73	35	22	21
1	0	0	0	0	0	0
2	0	0	0	0	0	0
3	0	0	0	0	0	0
4	0	0	0	0	0	0
5	0	0	1	0	0	0
6	0	0	26	55	57	61
7	0	0	0	7	18	15
8	0	0	0	2	2	2
9	0	0	0	1	1	1
10	100	100	0	0	0	0

Table S3: Group size resulting in the best score for 100 matrices generated using a known model and partition according to log-likelihood (LL), AIC, BIC, Bayes factors with hyper-priors matching the generating distribution (BF), and NML. Each matrix was generated using a group model with a different but known partition into five groups (True), with interaction probabilities drawn from a beta distribution $B(\alpha = 0.7, \beta = 0.7)$.

		Casc		MPN		Modular					Group				
		BM	TL	BM	TL	H	K	P	C	O	H	K	P	C	O
Kongsfjorden	AIC	1	2	1	2	1	2	3	5	4	3	4	2	1	5
	BIC	1	2	1	2	1	2	3	5	4	1	3	2	4	5
	BF	1	2	1	2	1	2	3	5	4	2	3	1	4	5
	NML	1	2	1	2	1	2	3	5	4	3	4	1	2	5
Lough Hyne	AIC	1	2	1	2	1	5	4	2	3	4	5	3	1	2
	BIC	1	2	1	2	1	5	4	2	3	2	4	1	3	5
	BF	1	2	1	2	1	5	4	2	3	3	4	1	2	5
	NML	1	2	1	2	1	5	4	2	3	3	4	2	1	5
Reef	AIC	2	1	2	1	3	5	4	1	2	3	5	4	2	1
	BIC	2	1	2	1	3	5	4	1	2	3	5	4	1	2
	BF	2	1	2	1	3	5	4	1	2	3	5	4	1	2
	NML	2	1	2	1	3	5	4	1	2	3	5	4	1	2
St. Marks	AIC	1	2	1	2	1	5	3	4	2	4	5	3	2	1
	BIC	1	2	1	2	1	5	3	4	2	3	4	2	1	5
	BF	1	2	1	2	1	3	4	5	2	3	5	2	1	4
	NML	1	2	1	2	1	3	4	5	2	4	5	2	1	3
Weddell	AIC	1	2	1	2	5	3	4	2	1	4	5	3	2	1
	BIC	1	2	1	2	5	3	4	2	1	3	4	2	1	5
	BF	1	2	1	2	5	3	4	2	1	4	5	2	1	3
	NML	1	2	1	2	5	3	4	2	1	4	5	3	1	2
Ythan	AIC	1	2	2	1	3	5	2	1	4	2	4	3	1	5
	BIC	1	2	2	1	3	5	2	1	4	1	3	2	4	5
	BF	1	2	2	1	3	5	2	1	4	1	4	2	3	5
	NML	1	2	2	1	3	5	2	1	4	1	4	2	3	5

Table S4: Model selection ranking (1: best to 5: worst) of empirically-determined models within each model family according to AIC, BIC, Bayes factors (BF) and total length based on NML. Cascade and MPN hierarchies specified by body mass (BM) and trophic level (TL). Modular and group partitions specified by habitat (H), kingdom (K), Phylum (P), class (C) and order (O).

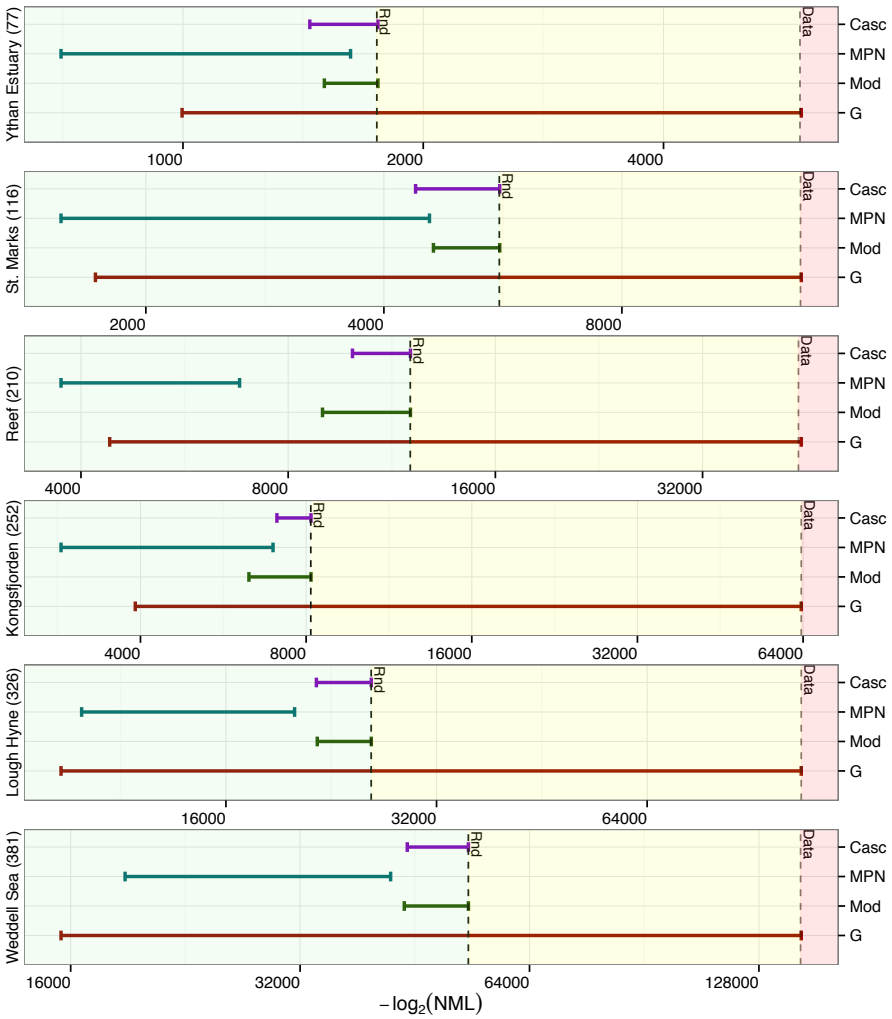

Figure S1: Total length for cascade, MPN, modular and group model families. For each combination of model family and food web, we searched for the hierarchy or partition of species that resulted in the shortest (best fit) and longest (worst fit) total length (all other models are necessarily contained in this range). Vertical dashed lines mark two reference points: the total length of a random graph model and the uncompressed adjacency matrix.

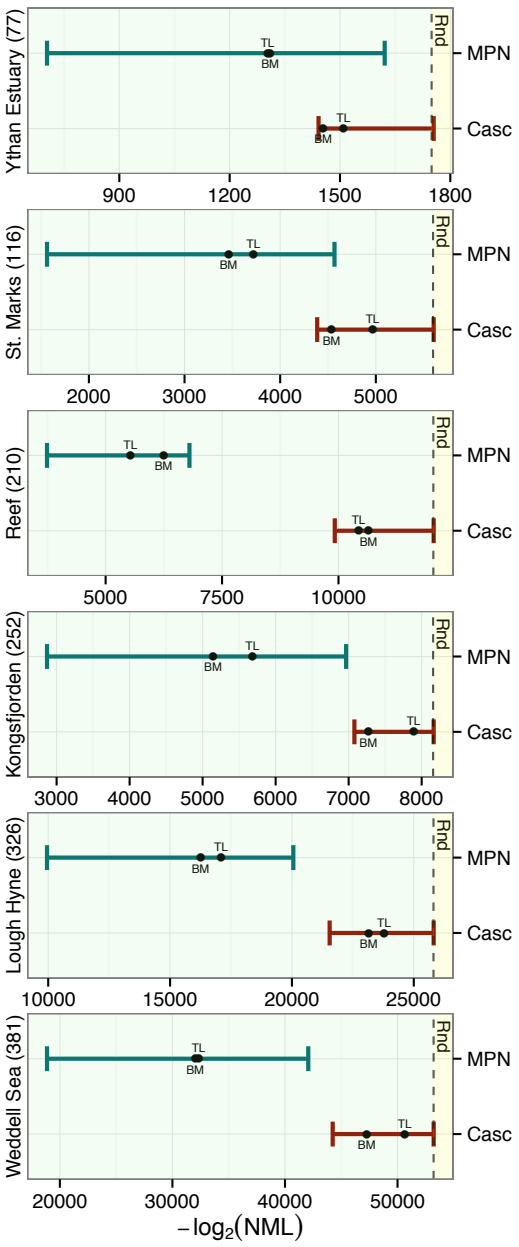

Figure S2: Total length for cascade and MPN model families when empirical data on body mass (BM) and trophic level (TL) were used to define model hierarchies.

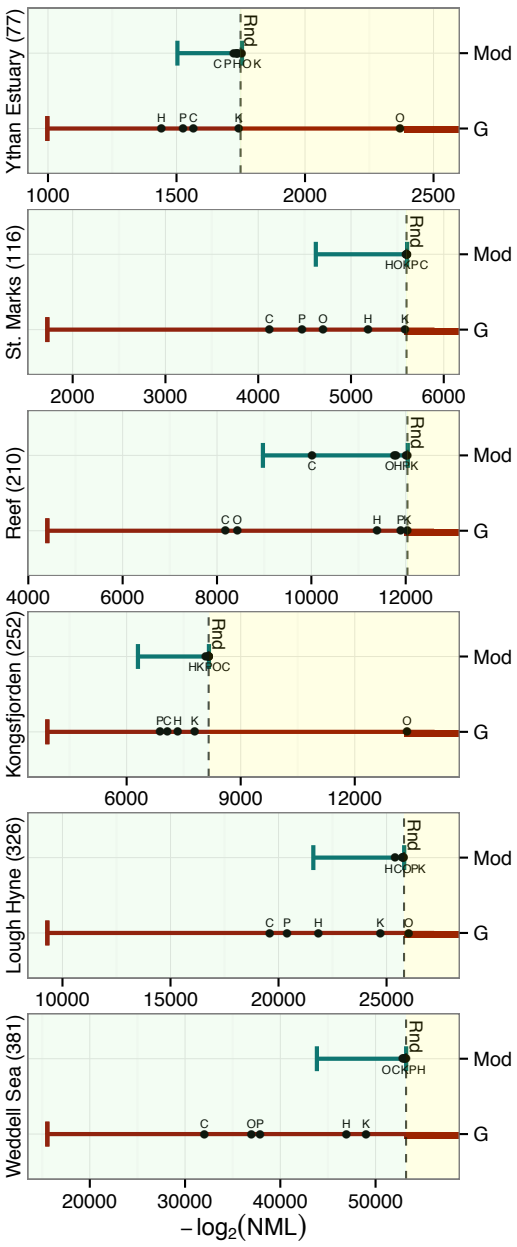

Figure S3: Total length for modular and group model families when empirical data on habitat (H) and taxonomic information—Kingdom (K), Phylum (P), Class (C) and Order (O)—were used to define model partitions.

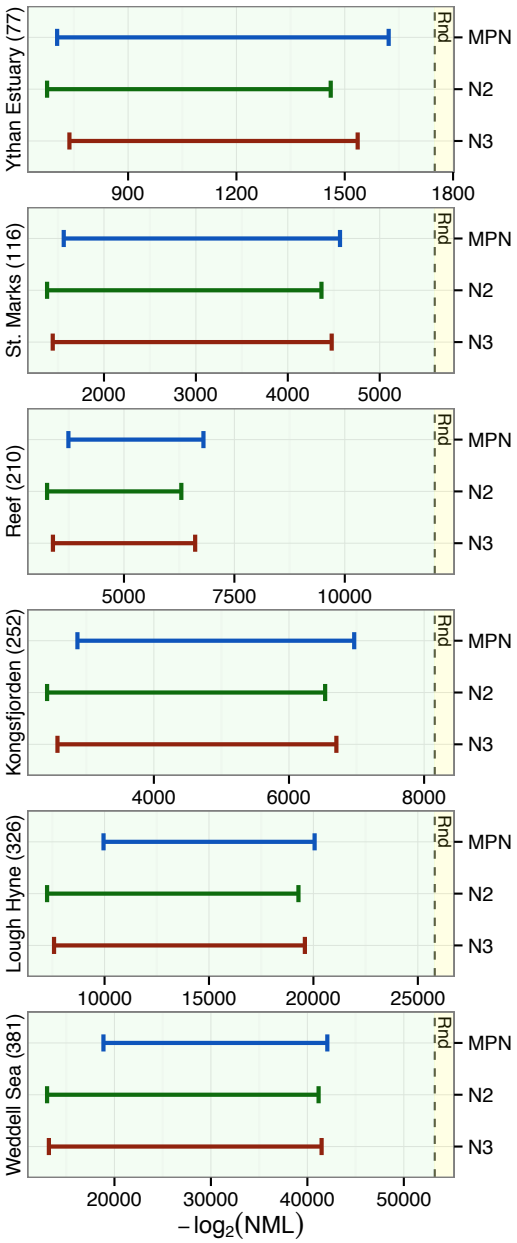

Figure S4: Total length for MPN model family and N2 and N3 variants formed by successively relaxing model feeding constraints.

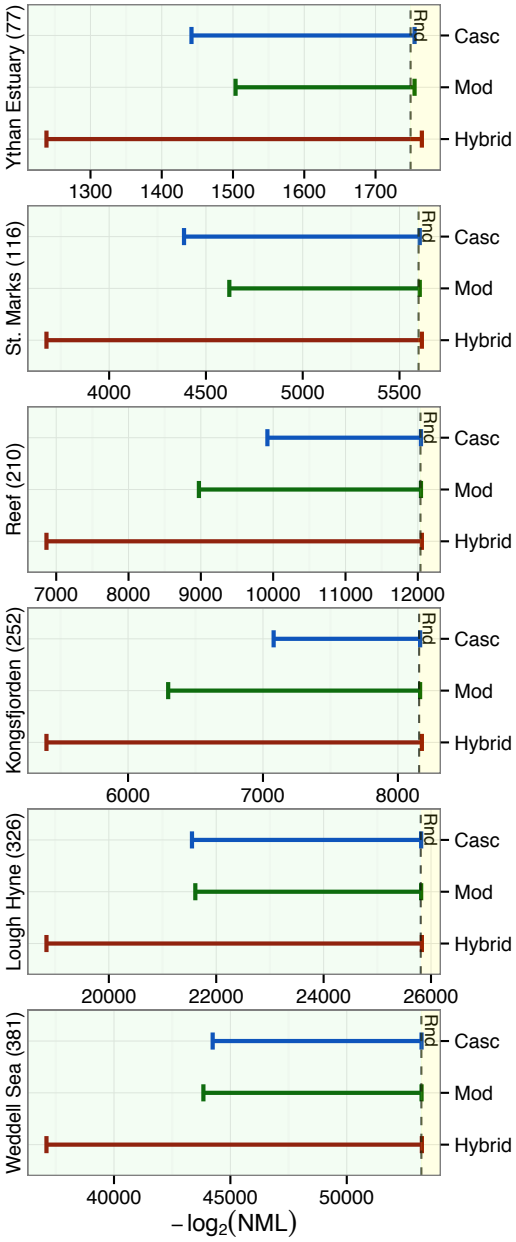

Figure S5: Total length for Hybrid model family, which is a combination of cascade and modular model families.