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).

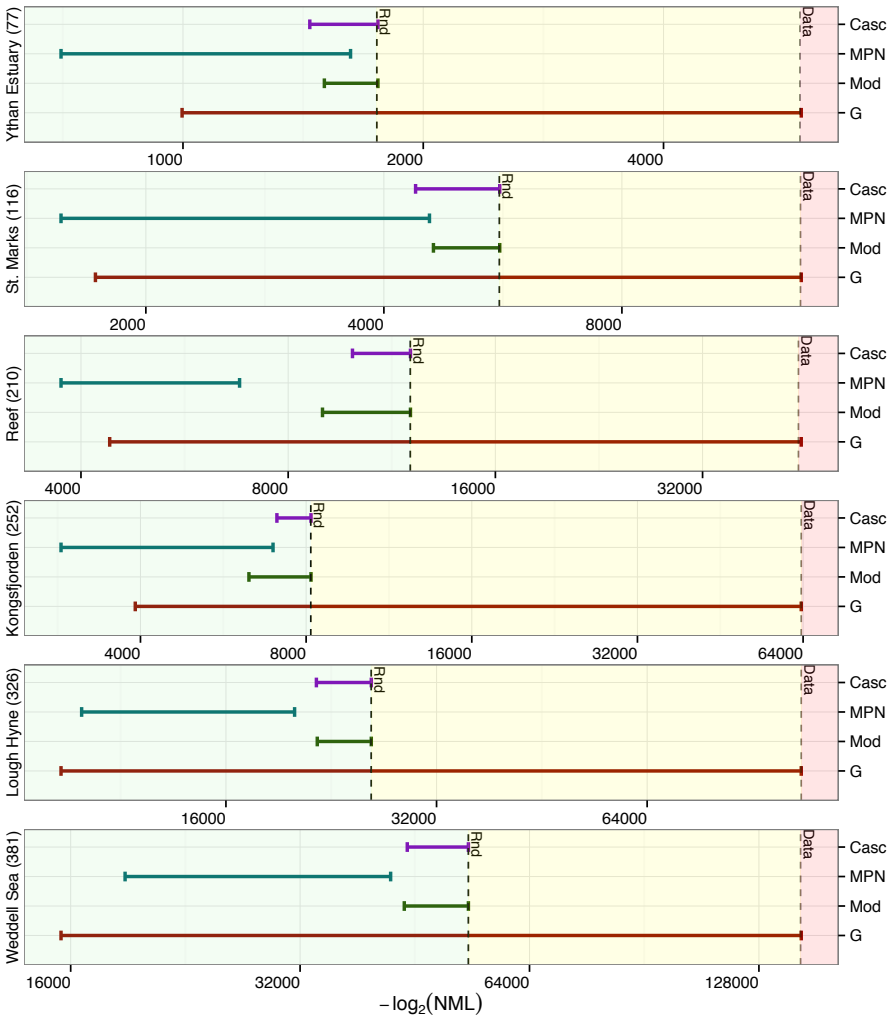

Figure S1: Total length for cascade, MPN, modular and group model families. For each combination of model family and food web, we searched for the hierarchy or partition of species that resulted in the shortest (best fit) and longest (worst fit) total length (all other models are necessarily contained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

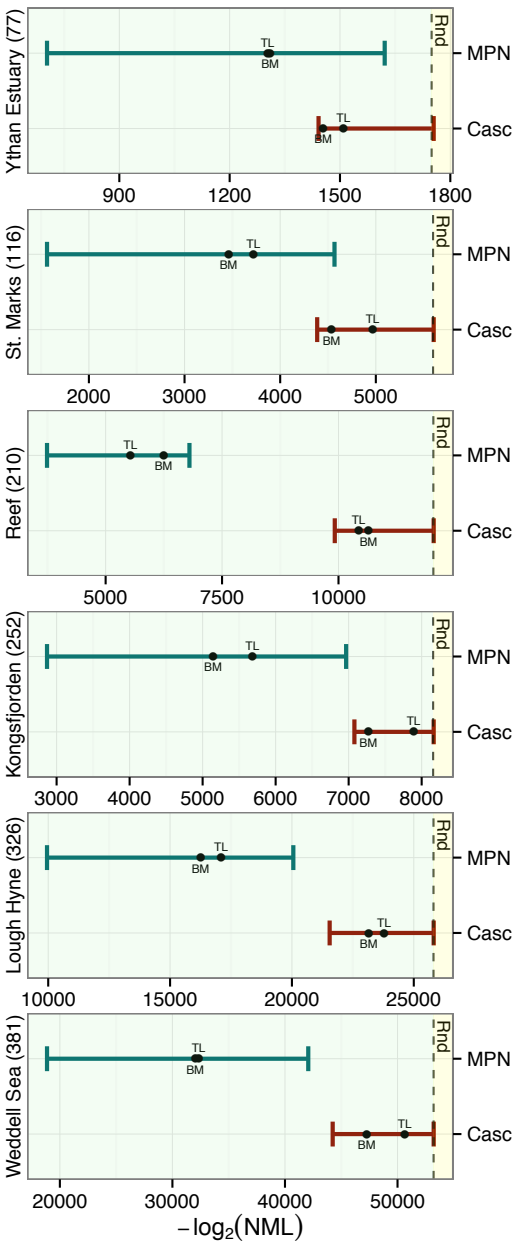

Figure S2: Total length for cascade and MPN model families when empirical data on body mass (BM) and trophic level (TL) were used to define model hierarchies.

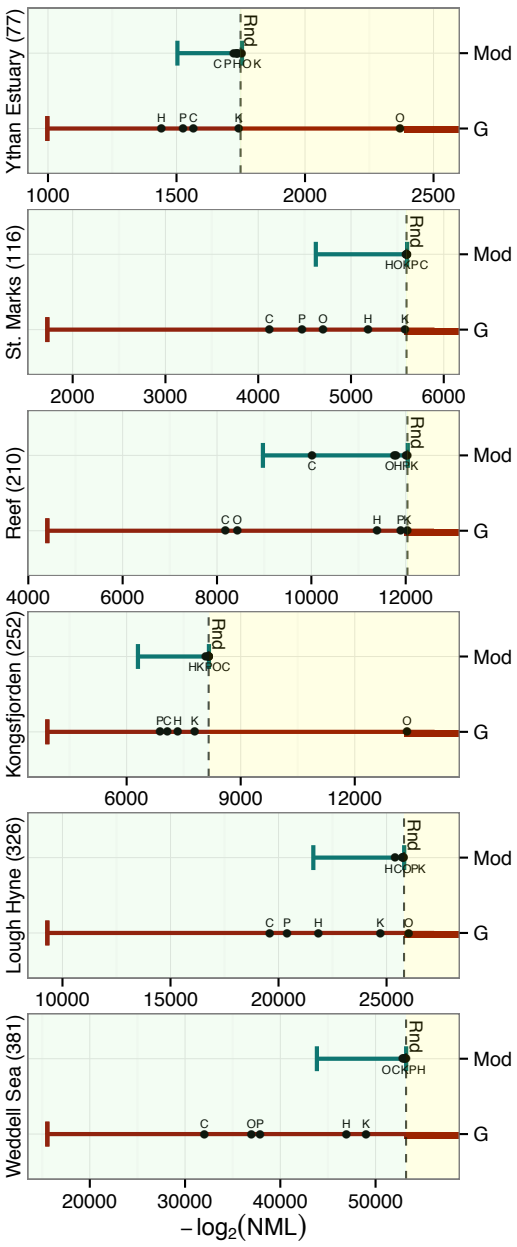

Figure S3: Total length for modular and group model families when empirical data on habitat (H) and taxonomic information—Kingdom (K), Phylum (P), Class (C) and Order (O)—were used to define model partitions.

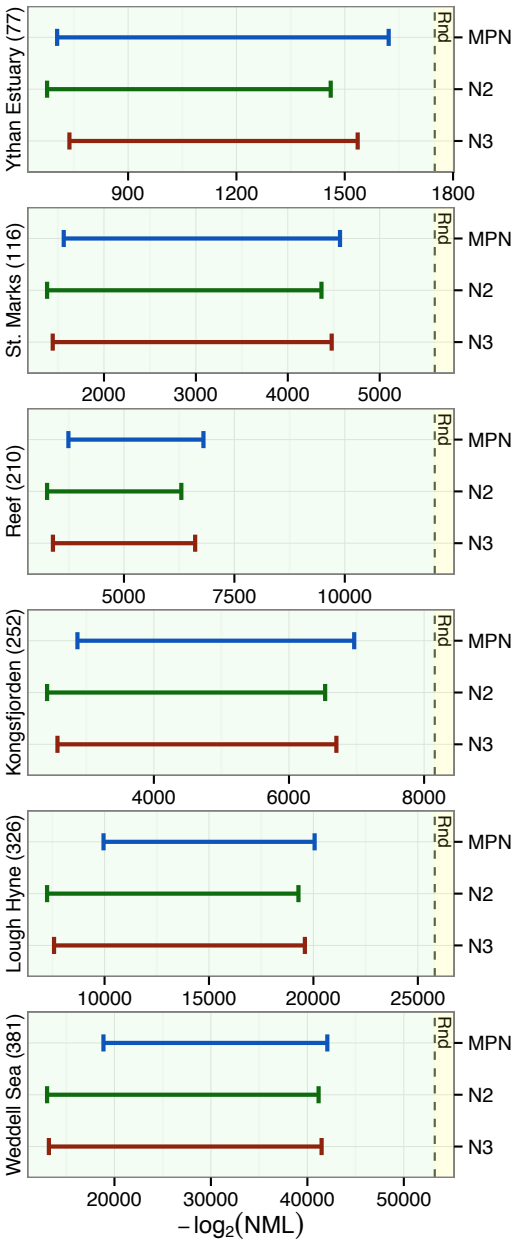

Figure S4: Total length for MPN model family and N2 and N3 variants formed by successively relaxing model feeding constraints.

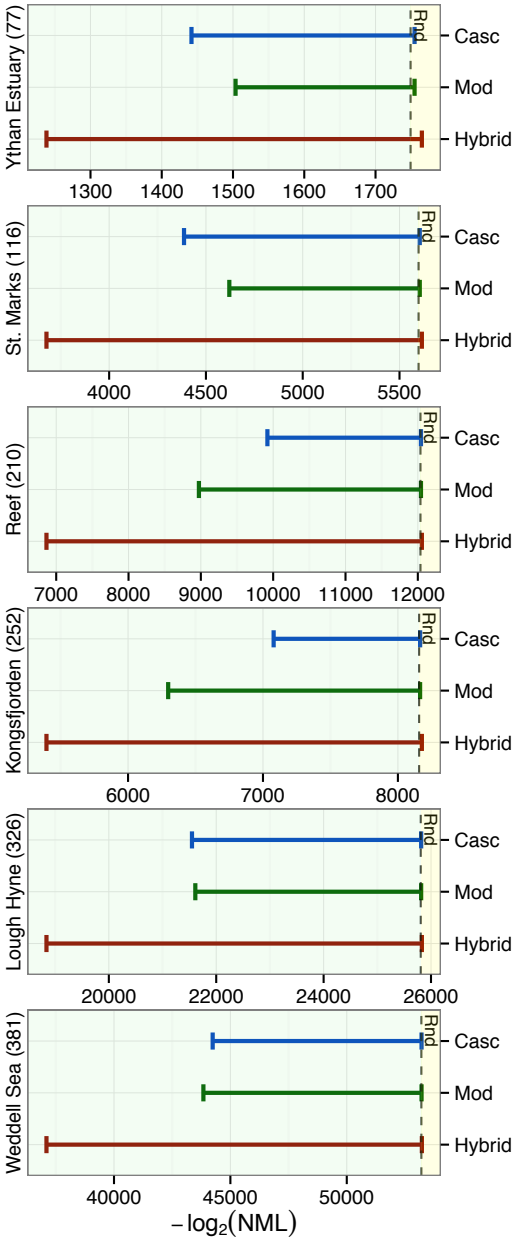

Figure S5: Total length for Hybrid model family, which is a combination of cascade and modular model families.