Contextual evidence for categorising interactions and guiding their discovery

Kesem Abramov [1], Shir Miryam Nehoray [1], Rami Puzis [2], Peter J. Mucha [3], and Shai Pilosof [1,4]

[1] *Department of Life Sciences, Ben-Gurion University of the Negev, Be'er Sheva, Israel*
[2] *Department of Systems Information and Software Engineering, Ben-Gurion University of the Negev, Be'er Sheva, Israel*
[3] *Department of Mathematics, Dartmouth College, Hanover, NH, USA*
[4] *The Goldman Sonnenfeldt School of Sustainability and Climate Change, Ben-Gurion University of the Negev, Be'er Sheva, Israel*

August 23, 2026

*Corresponding author. Email: pilos@bgu.ac.il

Abstract

Ecological interactions shape community dynamics and stability, yet many go unrecorded. Link prediction methods attempt to tackle incomplete knowledge, but predictions are only conjectures, and validating them all in the field is neither feasible nor desirable. Furthermore, interactions differ in the action needed for their detection. Validation therefore must be guided by ecological and statistical theory. We present a guided-sampling framework that combines predictions with local and regional observations, using within-system variability across replicated networks as contextual evidence. It assigns each link to a fine-resolution, ecologically-aware taxonomy that separates conflated categories (e.g., forbidden versus undersampled interactions). Each category then implies a concrete action (e.g., sample more, change method). A Bayesian formulation quantifies confidence in each assignment and incorporates researchers' knowledge. We include a worked empirical example, and an interactive Bayesian version online (http://lpguide.ecomplab.com). Contextual evidence turns link prediction from a source of untested hypotheses into a plan for fieldwork.

Link prediction is incomplete without validation

Ecological interactions are the backbone of ecosystem function [1]. However, our knowledge of interaction data, typically represented as links in ecological networks, is incomplete, constrained by finite sampling effort, methodological biases, and the cryptic nature of many species [2–5]. These gaps, which are collectively called the "Eltonian Shortfall", bias our representation of ecological networks and hamper our ability to predict how communities will respond to perturbations [6]. Given the central role of interactions for biodiversity and human well-being [7–10], there is a surge of attempts to predict unobserved links using an expanding array of link prediction approaches [11–17]. However, models return candidate interactions, not evidence that they exist. Ranking candidates by their plausibility within a network [18, 19], or interpreting which features drove a prediction [20], might narrow the search, but a reliability score is a model-internal verdict rather than

independent evidence. That is, its plausibility is assessed against the same network and method that produced it. This leaves researchers uncertain whether a prediction is valid, and without operational guidance on how to act on it. Exhaustive validation of all predictions is not feasible neither necessary, because links differ in their plausibility and effect on community stability and function. The operative question is therefore not whether every prediction is correct, but which predictions are worth the fieldwork.

In this paper we propose that validation and interpretation of ecological link prediction should be framed as a process of **evidence accumulation**, in which confidence emerges through repeated, independent lines of evidence rather than from a single evaluation exercise. Evidence accumulation operates along two dimensions: contextual evidence, drawn from replicated networks sharing the same species pool and ecological context [21, 22], and methodological evidence, drawn from independent sampling methods with different detection abilities and biases [23, 24]. Building on these two lines of evidence, we develop a taxonomy for categorising predicted links and offer guidelines for directing sampling efficiently toward the evidence each category requires. We further show how confidence in a link's categorisation can be quantified and how it grows as evidence accumulates, and demonstrate the framework's application to a real ecological system. To facilitate the presentation and intuition we provide an interactive guide at `http://lpguide.ecomplab.com`.

Why current validation schemes fall short

The root of the problem of validating link predictions lies in how predictive performance is evaluated. Validation ideally requires ground truth (Box 1). Such ground truth is almost never available, because confirming an interaction's presence, and even more so its absence, is prohibitively costly. Standard practice therefore evaluates predictions against the observed interactions in the focal network alone, treating them as ground truth — a process we refer to as 'within-network evaluation' (Box 1). This substitution is the crux of the problem. A predicted interaction that is unobserved in the data is classified as a false positive and read as a candidate missing link the model has recovered. However, because many interactions go undetected even under intensive sampling [2, 4], such a prediction is ambiguous: it may be a real interaction the sampling missed, or a model error. The ambiguity runs deeper still: when a model correctly predicts an unobserved interaction not to exist, it flags it as a true negative. However, this single label conflates three ecologically distinct realities: a forbidden interaction that is biologically impossible [25], a feasible interaction unrealised in the focal network due to local conditions [26–28], or a real interaction missed during sampling. Within-network evaluation cannot distinguish between these scenarios.

For these reasons, when predictions are made, researchers occasionally attempt to validate them by searching for evidence in sources external to the system at hand, querying interaction databases such as GloBI [29] or the Web of Life [30], conducting literature searches [31, 32], or consulting experts [33]. However, such external evidence carries important limitations. Because interaction patterns vary across space, time, and ecological contexts [34–36], external records may confirm predicted links that do not occur in the focal system, simply because the ecological context has been stripped away. Taxonomic inconsistencies between datasets further complicate interpretation, and external evidence is often infeasible for systems with high endemicity, such as island communities [37]. Moreover, interaction similarity between networks decays with both geographical and compositional distance [35], driving a corresponding decay in the transferability of interaction information [37].

Box 1: Basics of within-network evaluation

The most common approach for assessing performance of link prediction is to compare model output to observed interactions, treated as ground truth, to quantify how well the model captures a network's interaction structure using appropriate metrics.

Evaluation therefore begins by splitting the data into training and testing sets. The standard way to do this is link withholding: a subset of observed links is removed, the model is trained on the remainder,

and the withheld links are predicted and compared to their observed values. Both observed links and unobserved pairs can be withheld, so that predictive performance can be scored against both classes (see, e.g., [38]). To use the data efficiently and obtain a stable estimate of performance, withholding is typically organised as k-fold cross-validation: the data are partitioned into k equally-sized folds, each withheld in turn while the model is trained on the remaining $k-1$, and performance is averaged across the k rounds [39]. Its extreme case is the leave-one-out design, where every link is withheld and predicted individually [39]. A specific use of the same machinery is not to estimate performance but to generate candidates by flagging the highest-scoring unobserved pairs as candidate missing links [19].

The predicted links are then evaluated in one of three ways. Threshold-based evaluation converts predicted probabilities into binary outcomes using a cutoff (often 0.5, or chosen by sensitivity analysis [37]): predictions above it are interactions, those below non-interactions. Performance is summarised in a confusion matrix of true positives, true negatives, false positives, and false negatives, from which metrics such as precision are derived [40]. Threshold-free evaluation instead avoids a single cutoff and summarises the model's ability to separate presences from absences across all thresholds, using measures such as the area under the precision–recall curve (PR-AUC) or the receiver operating characteristic curve (ROC-AUC) [17, 40]. When the target is predicting interaction strength (such as visitation frequency) rather than occurrence, observed and predicted values are compared directly through correlation or weighted performance metrics (NSE, NNSE) [37, 41].

Ecological networks are sparse, and this class imbalance complicates evaluation: a model can achieve high accuracy by correctly predicting non-interactions while still failing to recover true interactions [17, 40]. The imbalance can be addressed during the training phase, for example by assigning greater weight to the rarer interaction class, and during the evaluation phase, by choosing metrics that remain informative under imbalance rather than overall accuracy [40].

Critically, all of these approaches treat unobserved interactions as true absences against which predictions are evaluated, and assume observations to be a true realization. These are central assumptions that the evidence accumulation framework is designed to address.

A framework for evidence accumulation and guided sampling

We present a framework rooted in two complementary modes of accumulating additional evidence: obtaining contextual evidence and increasing methodological evidence.

Obtaining contextual evidence

A preferable, more ecologically informed alternative to validation using external sources is obtaining contextual evidence: leveraging replicated networks, sampled across space or time within the same system for guiding interaction discovery. Because replicates share the same species pool and ecological context as the focal network, they come from the same distribution as the focal network and therefore preserve the conditions under which interactions are or are not realised, which is the information that external evidence discards. Therefore, within-system evidence is less likely to produce false confirmations. What constitutes a valid replicate depends on the researcher's knowledge of the system [36, 42], with ecological similarity (e.g., species composition and environmental context) serving as the guiding principle [43]. Crucially, this is not a new practice: many studies already sample interaction networks repeatedly across sites or time points as part of their core research design [34, 36, 44].

Contextual evidence connects to two established lines of work: structure-aware (block) cross-validation [45], which concerns how a model's reliability is tested when the data have natural structure, and metawebs [15, 22, 28], regional catalogues of all interactions possible for a given species pool. From different directions, both show that a system's partition structure carries information: block cross-validation reveals that ignoring it in evaluation makes models look more reliable than they are, while the metaweb literature finds that pooling it

away over-predicts local interactions and discards their variability. Prior to the actual application of a model, its performance is typically assessed by cross-validation, in which the data are split so that the model is trained on certain parts and tested on others (Box 1). Roberts et al. [45] show that because ecological observations are non-independent in space, time, or phylogeny, random splitting leaks correlated observations across the train–test boundary, so predictive error is underestimated. Their solution is to split along the structure itself. In our case, that would be separating the information in the focal network from that in the replicates (see below). Banville et al. [28] formalised the distinction between an interaction's regional feasibility (its metaweb probability) and its local realisation, which additionally requires local co-occurrence and realisation where the partners meet; local probability is thus bounded above by the regional one, and equating the two over-predicts links. Downscaling a metaweb this way yields only an upper bound on local interactions [22, 28]. The variation between local networks reflects the gap between regional feasibility and local realisation, which Banville et al. [28] frame as a primary open challenge that must ultimately be validated in the field. We leverage this gap to generate contextual evidence and to provide operational guidance for targeted sampling.

A taxonomy of predicted links

How, then, can we raise our confidence in local-scale predictions? A predictive model can generate far more candidate links than any field programme can verify. Ranking them by predicted probability is a useful first move, but it orders the candidate set along a single, model-internal axis. It identifies which links the model favours, not what evidence any of them is missing, and therefore not what a field programme should do next. Two links with the same score can require incompatible actions: one has already been recorded in a neighbouring replicate and needs only further sampling locally, whereas the other has been seen nowhere in the system and will not be resolved by any amount of the same method. Researchers' knowledge of the ecosystem and methodology informs the actions that can be taken.

We perform link prediction on each network separately, preserving its local context. We then use between-network variation as evidence. Concretely, we add a third dimension to the classical confusion matrix (Box 1): whether a predicted interaction is observed in at least one replicate within the same system, yielding a taxonomy of eight link categories (Fig. 1). This approach resolves ambiguities invisible to within-network evaluation; for instance, whether an interaction absent from the focal system reflects a forbidden link or is simply locally absent (present in another replicate). It also refines the classical division of unobserved interactions into forbidden and missing links [4, 25, 26], a distinction that predates predictive methods and is too coarse for them. For example, a link classified as a true negative (unobserved and predicted absent) is *possibly forbidden* if it is seen nowhere in the system, but *locally absent* if it appears in another replicate. For simplicity, we first discuss a deterministic (error-free) categorisation; however, each of the three components of evidence carries some error. We then provide a quantitative Bayesian framework that propagates the uncertainty of the evidence into a posterior confidence for each categorisation, and show how confidence grows with the number of replicates in which the prediction is consistently supported, whether through repeated observation or repeated absence.

The taxonomy we propose helps guide sampling and modelling decisions (Fig 2). Among predicted links, *recurrent links* — predicted and observed locally and in replicates — likely require no further action: consistent observation across independent networks indicate compelling evidence. Recurrent links may involve abundant or prominent species, but can also reflect strong partner fidelity in specialists that consistently interact with the same partners across environments [35, 46]. *Locally unique* links are predicted and observed only in the focal network. They may arise through several non-mutually exclusive mechanisms: species rarity reducing encounter probability [4, 47], high endemicity driving species turnover [35], or interaction rewiring under local conditions such as reduced availability of alternative partners [48]. Local uniqueness might be, in many cases, an ecological finding rather than a failure to obtain evidence. *Possibly missing* links are predicted but unobserved locally, yet observed in replicates. This is arguably the most actionable category, as the predicted interaction is demonstrably observed elsewhere, making insufficient local sampling the most parsimonious explanation; however, phenological mismatches or other local biological constraints can also lead to a locally unobserved link

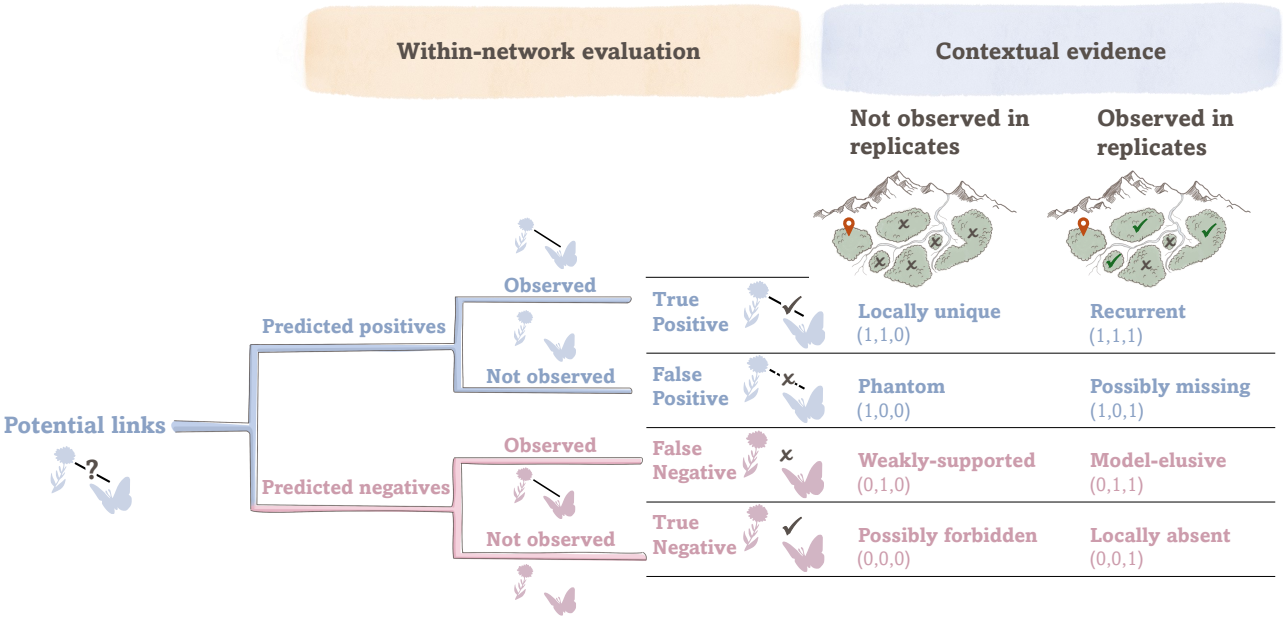

Fig. 1: A taxonomy of predicted links. Predicted ecological interactions are assigned to one of eight link categories using three sources of evidence E: whether a link is predicted by the model (Y), observed in the focal network (O_l), and observed in at least one within-system replicate (O_r). In the simple deterministic version, each source is a single bit, taking the value 1 when it supports the link and 0 when it does not. For example, a *locally unique* link is predicted and observed in the focal network, but recorded in no replicate, giving $E = (Y, O_l, O_r) = (1, 1, 0)$. Within-network evaluation alone (left) uses only Y and O_l, yielding the four classical confusion matrix classes: true positives, false positives, true negatives, and false negatives (Box 1). Adding contextual evidence from within-system replicates (right) resolves these into eight categories, each carrying a distinct ecological interpretation. Section Bayesian Inference attaches a confidence to the assignment by allowing each source of evidence to carry an error. The Supplementary Information states and relaxes the assumptions behind this simple version, grounds the error rates in realisation and detection, and lets confidence accumulate over replicates. We illustrate the framework using a spatial system, though it can also apply to temporal or other forms of replication.

[49]. Targeted detection efforts using the same or complementary sampling methods can discriminate between these alternatives [5, 50]. *Phantom* links are predicted but never observed anywhere, requiring independent methodological evidence to distinguish model artefacts from systematically missed interactions. This distinction is particularly important when the predictive models are based solely on network structure [14, 18, 37], lacking explicit ecological features such as phylogeny or species traits [11, 13, 21].

Among unpredicted links, the taxonomy gives empirical meaning to the theoretical distinction between metawebs and realised networks [28, 51]. *Locally absent* links are unobserved locally but detected in replicates, confirming biological feasibility while attributing local absence to ecological context. This is consistent with documented spatial variability in interaction structure [35, 36, 52]. *Possibly forbidden* links remain absent everywhere and are unsupported by prediction. Links in this category are strong candidates for genuinely impossible interactions constrained by morphological, physiological, or phenological mismatches [25]. However, 'forbiddenness' is better viewed as a continuum than a binary fact [26]: intraspecific trait variability means that an interaction predicted to be forbidden from mean trait values may occur in some individuals, populations, or years. Consistent absence across replicates spanning varied local conditions therefore provides stronger (though never definitive [27]) evidence of true biological constraints than any single-network assessment alone. Nevertheless, absence can still result from a method that cannot detect the interaction, which a complementary method might reveal. *Weakly-supported* links are observed locally but are neither predicted (locally) nor detected in replicates. They may therefore reflect observation or identification errors [53, 54], extremely rare or sporadic interactions, or genuine context-specific realisations that both the model and the broader system fail to capture. Finally, *model-elusive* links are observed locally and across replicates but not predicted. When rare, they reveal specific blind spots. When common, they likely signal a more fundamental model failure requiring revision rather than

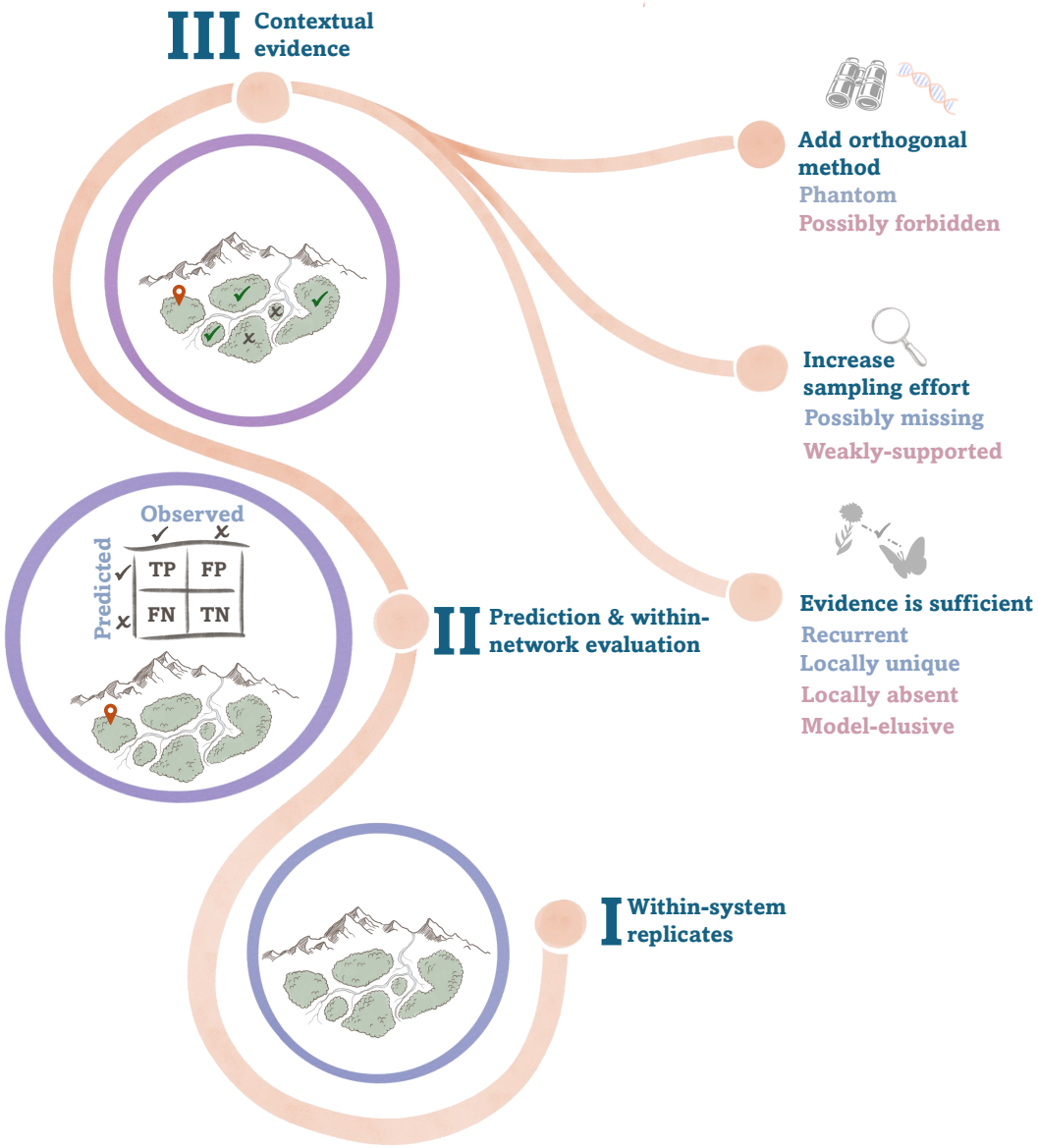

Fig. 2: A framework for evidence accumulation. Evidence accumulation is a sequential process. First, contextual evidence is drawn from within-system replicates, categorising each link. Then, methodological evidence can be added, either by increasing sampling effort with the same method or by adding methods with different detection biases. For each link category, the framework suggests an action: whether no further action is needed, sampling effort should be increased, or independent sampling methods should be used. Link categories are colour-coded consistently across this figure and Fig. 1.

further sampling.

Accumulating methodological evidence

Field ecologists must make the most of limited resources, which means narrowing the set of interactions worth sampling. Link prediction is a first step: it produces a set of candidate links that can be ranked by their probabilities [19]. Contextual evidence narrows that set further, setting specific, ecologically interpretable detection goals — for example, targeting locally absent or possibly forbidden links. Contextual evidence can be applied without predictions, but the ecological interpretation is then less resolved. For instance, without a prediction, possibly missing and locally absent links become indistinguishable, as both are unobserved locally yet present in replicates.

Contextual evidence, though necessary, is not always sufficient. Replicates sampled with the same method share its biases: e.g., an interaction consistently missed by visual observation goes undetected across every replicate, and a prediction model trained on structurally similar networks reproduces the same blind spots [23, 24, 55]. Evidence accumulation therefore requires a second stage: improving the methodological evidence (Fig. 2), comprising two complementary possible actions. The first is to increase sampling effort with the same method, appropriate mainly when that method has already detected the interaction in replicates. The second is to add orthogonal methods with complementary detection biases [24, 56], to overcome the limits of any single method [23, 57]. The two kinds of evidence are applied in sequence, considering contextual evidence first because only by categorising a link through within-system replication does the appropriate methodological action become clear (Fig. 1). We demonstrate the application of the framework to a real plant–pollinator system below (Section Empirical demonstration).

Together, these two kinds of evidence make the taxonomy actionable rather than merely descriptive. Each category points to a specific follow-up action, or, for well-supported links, to accepting the current evidence as sufficient (Fig. 2). As a general guideline, we recommend the following logic. For *recurrent*, *locally unique*, *locally absent*, and *model-elusive* links, no further sampling is likely needed. In those cases, the categorisation is already interpretable (though frequent model-elusive links point to a need for model revision rather than more data). For *possibly missing* and *weakly-supported* links, increased sampling effort at the focal site is likely sufficient and targeted. For *phantom* and *possibly forbidden* links, independent methods are necessary, because intensifying the same method will likely not help. Nevertheless, the course of action rests ultimately with the researchers, who know the boundaries of their system and can weigh the cost of adding methods against that of intensifying sampling. Familiarity with a system also lets researchers translate contextual evidence into mechanism, understanding for instance why a predicted interaction is unique to one network yet absent from others.

Three additional considerations inform how the framework is applied in practice. First, the logic of contextual evidence can in principle be transposed to the methodological dimension, yielding an analogous taxonomy. However, this is rarely equivalent, and is often less feasible. Researchers typically have access to far fewer independent methods than replicates, and methods are not ecologically exchangeable in the way replicates are. A link detected by DNA barcoding but not by visual observation carries a different ecological meaning than a link observed in one location but not another. In our view, contextual evidence therefore remains the more accessible foundation for this taxonomic framework. Second, within the methodological dimension, the framework prescribes which action is appropriate for each link category, but not which is feasible.

Third, changing the predictive model is another lever available to researchers, but it operates differently from the other two: model revision changes the predicted dimension of the categorisation, which cascades into reassigning link categories across the contextual dimension. It is therefore not a validation action but a reset of the categorisation itself, and its effects on link category distributions can also be diagnostic, revealing which categories shrink or grow when model assumptions change. This is particularly informative for model-elusive links: when a link is consistently observed but never predicted, a conceptual change in model architecture may

be necessary to resolve the (apparent) model failures, such as using a different kind of model (e.g., tree-based vs. neural networks) or changing the set of features used for prediction [11, 13, 16, 58].

A Bayesian reading of the link taxonomy

Until now, we deterministically assigned each link to one of the eight categories from Fig. 1, based on three pieces of evidence: the model's prediction (Y), whether it was observed locally (O_l), and whether it was observed in the replicates (O_r). For example, a *possibly missing* link is predicted as positive ($Y = 1$), not observed locally ($O_l = 0$), and observed in replicates ($O_r = 1$). We collect these evidence as $E = (Y, O_l, O_r)$.

However, every evidence source $d \in \{Y, l, r\}$ is error-prone. The prediction can be wrong about the truth and not merely about the data it was fitted to; local sampling may miss interactions; and a realisable link may go unrecorded across the replicates. Therefore, the deterministic assignment, which we use for simplicity, effectively contains uncertainty. For instance, a *possibly missing* link may actually be a *recurrent* one that was missed locally due to error in local detection. We thus expand the deterministic approach by attaching a confidence to each category assignment: instead of one label per link, the evidence supports several categories as plausible alternatives among which probability is divided.

We treat the true category $c \in C$ as unknown and ask how probable each category is given the evidence E. Given the error rate of each source of evidence (ϵ_d), Bayes' theorem inverts the observed evidence into posterior probability:

$$\underbrace{P(C \mid E)}_{\text{posterior}} = \frac{\overbrace{P(E \mid C)}^{\text{likelihood}}\ \overbrace{P(C)}^{\text{prior}}}{\underbrace{P(E)}_{\text{evidence}}}.$$

The error rates can be used in calculating the likelihood. We write $z_{c,d} \in \{0, 1\}$ for the value that component $d \in \{Y, l, r\}$ takes in category $c \in C$, that is, the evidence that category c would produce if no source erred. We denote the sources of error as $\epsilon_d \in \{\epsilon_Y, \epsilon_l, \epsilon_r\}$. Assuming the error sources are independent given the category, each contributes $1 - \epsilon_d$ when the observed value E_d matches $z_{c,d}$, and ϵ_d when it differs:

$$P(E \mid C = c) = \prod_{d \in \{Y, l, r\}} \epsilon_d^{\mathbf{1}[E_d \neq z_{c,d}]} (1 - \epsilon_d)^{\mathbf{1}[E_d = z_{c,d}]},$$

where $\mathbf{1}[\cdot]$ is the indicator function equal to 1 when its input is true and 0 when false.

This framing acknowledges the idea of probabilistic interactions [28], but flips the point of view. Instead of calculating link probability under given conditions, which are in many cases unknown, we place a posterior on link categorisation. The factors which affect link probability (e.g., co-occurrence, abundance, traits, environment, sampling) enter in two places: they govern realisation and detectability, which underlie the error rates, and they inform the prior on the categories. This view is actionable: the posterior identifies which category the uncertainty favours, and thus the most appropriate action to resolve it.

The simplest derivation uses binary evidence ($E_d \in \{0, 1\}$), fixed error rates, a uniform prior, and independent errors. We present here two extensions (Fig. 3), and in the Supplementary Information we state the full set of assumptions explicitly, discuss how they can be relaxed, and provide worked calculations. In addition, we provide a rich interactive visualisation of how the posterior responds to the error rates, following the same development, available at http://lpguide.ecomplab.com.

One straightforward extension of the binary framework is to consider cumulative O_r (see Supplementary Information for the mathematical development). As replicates accumulate consistent evidence (repeated observation or repeated absence) the posterior concentrates on one category and confidence grows (Fig. 3a). However, confidence does not grow without limit. Once the replicate evidence is resolved, the posterior cannot exceed a maximum contextual confidence $\kappa = (1 - \epsilon_Y)(1 - \epsilon_l)$, set by the model and the local sampling alone. Replicates

therefore increase the rate to reach that maximum rather than raising it. Raising κ requires a better model or a complementary method, which is the methodological dimension of the framework. The rate of accumulation is ecologically informative because, under repeated detection, a posterior that sharpens quickly marks a link consistent across contexts, whereas slow saturation signals context-dependence.

Confidence can also be aggregated across categories. We denote by L_l and L_r the truths that O_l and O_r observe with error: whether the link is realised where we sampled, and whether it is realisable in the replicates. The signature of a category, $z_{c,l}$ and $z_{c,r}$, is the value each of these truths takes in it (Supplementary Information). Then, the *feasibility confidence* $\phi = P(L_l = 1 \text{ or } L_r = 1 \mid E)$ is then the posterior probability that at least one of them holds. Ecologically, it is a lower bound on the probability that the link is feasible, because a feasible link may go unrealised everywhere we looked, and unlike confidence in any single category, it is not bounded by κ (Fig. 3a; Supplementary Information).

The prior is the second place where ecology enters. A uniform prior gives all eight categories equal weight before any evidence, which is rarely what we believe because interactions are sparse, and a pair of generalists is far more likely to interact than a pair of specialists [59]. Researchers can implement such knowledge by changing the priors (Fig. 3b). For example, suppose the evidence we have at hand is $E = (1, 0, 1)$, the signature of a *possibly missing* link. Without changing that evidence, the prior shifts the categorisation towards *phantom* for a rare, specialist pair (reading the replicate record as a false detection, $L_r = 0$ although $O_r = 1$), or towards *recurrent* for a pair of strong generalists, which we believe will consistently interact (reading the local non-detection as a miss, $L_l = 1$ although $O_l = 0$) (Fig. 3b).

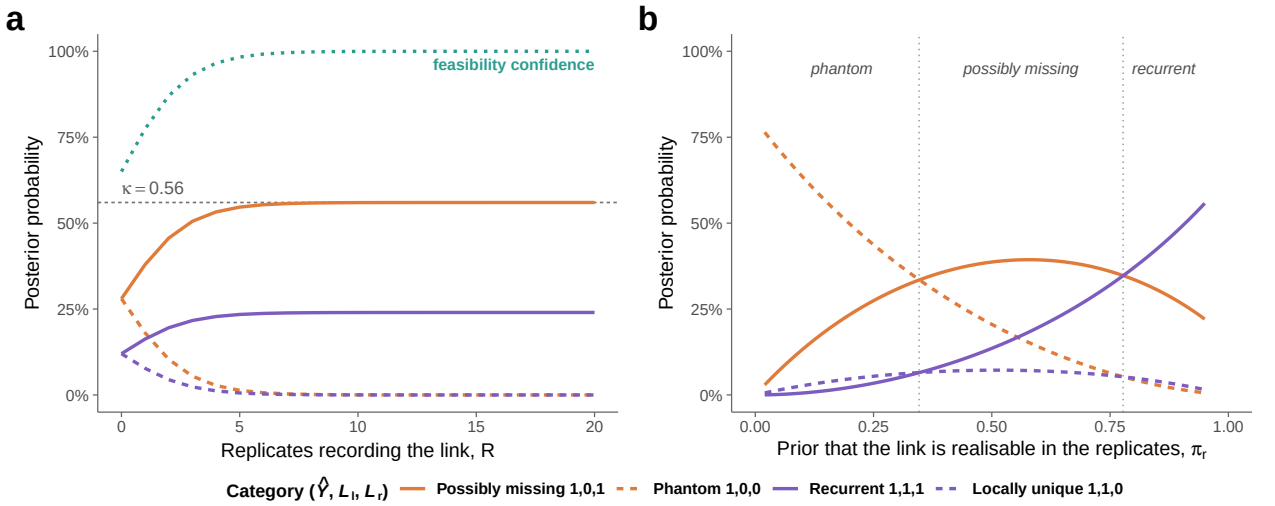

Fig. 3: Using the Bayesian framework to improve guided sampling. In this example the link is predicted ($Y = 1$), unobserved locally ($O_l = 0$), and searched for in every replicate. The legend gives each category's error-free signature. The lines differ in whether the link was realised where we sampled (distinguished by colour), and whether it is realisable in the replicates (solid where it is and dashed where it is not). **(a)** Posterior probability of four categories as within-system replicates accumulate. Confidence in *possibly missing* is limited to the maximum contextual confidence $\kappa = (1 - \epsilon_Y)(1 - \epsilon_l)$ (horizontal dotted line). The aggregate *feasibility confidence* ϕ is not bounded by κ. **(b)** The same four categories under an informed (non-uniform) prior, with the binary replicate evidence after five replicates ($O_r = 1$). Both priors (on local and replicate truths) rise with the pair's degree, so the axis runs from two interacting specialists on the left to two interacting generalists on the right. Dotted vertical lines mark where the leading category changes, named in italics. The same evidence and the same error rates therefore support three different readings, depending only on what is ecologically known about the pair. The illustrative rates are $\epsilon_Y = 0.2$ and $\epsilon_l = 0.3$, giving $\kappa = 0.56$, with per-replicate probabilities of recording the link of 0.105 where it is realisable in the replicates and 0.05 where it is not; they show the behaviour of the framework rather than describing any particular system. Together the panels turn the posterior into a sampling decision: (a) how many replicates is it worth sampling, and (b) how sampling would be affected by what we know about the species. We provide interactive versions of these plots in https://lpguide.ecomplab.com

Empirical demonstration of the evidence accumulation framework

To show how the framework guides gathering evidence in a real system, we applied it to plant-pollinator interactions on Cabrera island (Balearic archipelago, western Mediterranean Sea). Serra-Marin et al. [56] monitored pollination at six sites on the island by direct field observation and, critically, ran automated cameras in parallel at every site. This second, independent method lets us test the framework's predictions for the link categories that call for the use of additional methodology (Fig. 2).

We predicted all possible interactions among co-occurring species at each site using network embedding, which draws on latent structural patterns in each site's own pollination network, following the pipeline of Abramov et al. [37] under a leave-one-out approach (each potential interaction, observed or not, was withheld and predicted). Comparing predictions to field observations yields a within-site confusion matrix (Fig. 4A, first column). For every prediction, we then asked whether the same interaction appeared at any other site (second column), using this contextual evidence to assign each prediction to a taxonomy category (third column). At the final step, we tested targeted categories against the independent camera records (fourth column).

Across the six sites, field observation recorded 1624 potential interactions, of which 1232 (75.9%) were unobserved. Contextual evidence flagged 51 of these unobserved interactions as possibly missing while 103 observed links were weakly supported. The 51 possibly missing links were the higher priority of the two for validation, and targeting them alone reduced detection effort by roughly 24-fold from the original 1232 unobserved links.

Two other categories show why our guiding principles are useful: 941 links were possibly forbidden and 104 were phantom links with no support in the data at all, both signalling that increasing sampling effort of direct observations would be an inefficient course of action. Instead, the recommended course of action is to add an orthogonal method (Fig. 2). Cameras confirmed this reasoning directly: 8.5% of the possibly forbidden and 16.3% of the phantom links were captured on camera. Cameras also provided support for the categorisation system by recovering 51% of the links flagged as possibly missing, supporting the high priority that the framework assigns this category. Since direct observation is capable of detecting these links at other sites, the framework's prescribed action, more of the same effort, would have recovered them without the cost of a second method.

This guidance helps researchers make decisions at fine resolutions, such as whether to inspect specific sites or species of interest, and where to focus observation efforts (Fig. 4B). For instance, *Teucrium capitatum*, with multiple possibly missing links, rewards more intensive observation with the same method; *Helichrysum stoechas*, whose interactions largely escaped both observation and prediction, calls for the complementary sampling method instead; and *Daucus carota*, a generalist rich in locally unique links, shows how replication across sites turns single-network noise into ecological signal. Researchers can further direct their discovery efforts to high-priority link categories of threatened or keystone species.

Assumptions of evidence accumulation

Like any other inferential framework, contextual evidence rests on several assumptions. The deterministic categorisation treats the evidence at face value, which is not an assumption, but rather the starting point from which the probabilistic version (Section Bayesian Inference) builds confidence by attaching an error rate to each source of evidence.

The assumptions underlying the confidence concern the ecology, the sampling, and the predictive model. On the **ecology**, we assume that feasibility is regional (sensu Banville et al. [28]) while realisation varies across replicates, which is the signal the framework exploits. The simplest probabilistic version further assumes that all eight categories are equally probable before any evidence is seen, although this is rarely true. We already show one way to relax this assumption in Fig. 3b). On **sampling**, we assume that replicates share effort, method and taxonomic resolution (otherwise apparent evidence reflects sampling differences rather than feasibility), and that each source errs independently at a known rate. We further assume that detection error rates are known and common to all links. On the **predictive model**, we assume the model errs similarly for all links

and independently of the observations, although it inherits the biases of the data it was trained on. In the Supplementary Information we expand on these assumptions and show how to relax them.

Fig. 4: Distribution of evidence-based categories for plant–pollinator interactions in Cabrera island. (A) Flow of predicted links (for direct observations data set) through within-network evaluation classes, contextual evidence from other sites, which yields the taxonomy categories, and complementary-method (camera) evidence for targeted categories. Counts represent unique potential interactions summed across sites; percentages are out of the total number of potential links. An interactive version of the figure, with controls for sampling method and classification threshold, is available at https://lpguide.ecomplab.com/. **(B)** A detailed map of predictions between co-occurring species at the most species-rich site in the system, coloured by link category. Each cell represents an interaction between a plant species (row) and a pollinator species (column).

Future challenges and research directions

Establishing that variation between networks is evidence opens a set of questions about what counts as a replicate and what its evidence is worth. Replicates differ in how closely they resemble the focal network, so relating ecological dissimilarity to evidential weight is a natural next step [37]. Contextual evidence can be

drawn across space or time, yet these sources of variation are not ecologically equivalent. It remains open whether they should carry the same evidential weight, and if specific adaptations to temporal networks are necessary. Beyond space and time, other axes of variability may serve as well, and identifying them would widen the framework's reach. A special case are communities affected by perturbations, because range shifts and invasions generate assemblages for which no sufficiently similar reference network exists. Finding contextual evidence for predicted interactions in these emerging communities is therefore a critical research route.

A second set of questions concerns estimating the error rates and priors that the probabilistic version takes as input. The sampling rates are within reach of established designs, through repeated visits, effort- and abundance-based estimators, or negative controls [27, 60], and the realisation rate can be estimated from the replicates themselves, though only as a lower bound [28]. The model error is the hardest, because cross-validation measures a model against the observations it was fitted to rather than against the truth, so recovering it calls for calibration against an orthogonal method or latent-class estimation from several imperfect ones. The prior raises a different question: traits, phylogeny and independently surveyed abundances are external to the interaction data, whereas degree and generalism are read off the network and are already part of the evidence. Establishing which sources are external, and how to weight their contribution, will increase our confidence in the guided sampling.

A third set concerns turning confidence into design. Because replication buys speed towards a maximum set by the model and the sampling method, but cannot raise it, the practical question is when to add replicates and when to raise that maximum instead, which is a decision about cost that the accumulation curves can inform. The maximum is raised either by a complementary sampling method or by a better model, and the latter invites treating model ensembles as evidence [40, 61]: much as a replicate joins the local observation, a second model can enter as a further evidence source with its own error rate. In the same spirit, matching sampling and prediction methods to the specific needs of each link category would maximise coverage while minimising redundancy with evidence already in hand. Finally, the feasibility confidence offers an entry point to tasks beyond validation, such as assembling metawebs or investigating the functional consequences of validated links.

Concluding remarks

Ecological link prediction is maturing rapidly, yet two problems persist: interaction data are incomplete and biased (e.g., by species abundance) [6], and validating predictions is usually prohibitively costly. We address the second, and in doing so offer a route to the first. The remedy is not more sophisticated models alone, but a shift in how predictions are interpreted to guide data collection. Reframing link prediction as evidence accumulation provides an ecologically grounded reading of what a predicted link means. Our taxonomy of eight link categories and the two dimensions of evidence accumulation, contextual and methodological, turn that shift into a practical roadmap for the field. The framework therefore helps researchers to focus fieldwork on the links whose resolution would change what we conclude about the community, with the researchers' knowledge and hypotheses deciding which those are. In practice, evidence accumulation offers concrete guidelines for validating link predictions and should be built into study design from the outset. Crucially, it is testable with data that many research groups already gather, since study designs often include replicated networks across sites or time and the capacity to employ more than one method. The challenges ahead are therefore not only about building better models, but about better ways of knowing what our models mean.

Code and data availability

The full code and technical description of how to run it are available on the GitHub repository https://github.com/Ecological-Complexity-Lab/LP_validation. The data used for the empirical demonstration are available in the repository set up in the original publication by Serra-Marin et al. [56] at https://doi.org/10.5281/zenodo.17130777.

Acknowledgements

We thank Mercedes Pascual, Stefano Allesina and Barry Biton for insightful discussions.

Funding

The study was funded by the Israel Science Foundation (grant 491/25) to SP and the Human Frontiers Science Program (grant RGY0064/2022) to SP. SP was also supported by the Dartmouth Kalaniyot in the Sciences Visiting Scholar Program. PJM was supported by the NSF (grant DEB-2308460). KA was supported by the Kreitman School of Advanced Graduate Studies, Ben-Gurion University of the Negev, and by the Planning and Budgeting Committee (VATAT) of the Council for Higher Education, through the Levtzion Scholarship for Outstanding Doctoral Students.

References

1. Jordano, P. Chasing ecological interactions. en. *PLoS Biology* **14,** e1002559. doi:10.1371/journal.pbio.1002559 (2016).
2. Chacoff, N. P., Vázquez, D. P., Lomáscolo, S. B., Stevani, E. L., Dorado, J. & Padrón, B. Evaluating sampling completeness in a desert plant–pollinator network. *Journal of Animal Ecology* **81,** 190–200. doi:10.1111/j.1365-2656.2011.01883.x (2012).
3. Souza, C. S., Oliveira, P. E., Rosa, B. B. & Maruyama, P. K. Integrating nocturnal and diurnal interactions in a Neotropical pollination network. en. *The Journal of ecology* **110,** 2145–2155. doi:10.1111/1365-2745.13937 (2022).
4. Jordano, P. Sampling networks of ecological interactions. en. *Functional ecology* **30,** 1883–1893. doi:10.1111/1365-2435.12763 (2016).
5. Llopis-Belenguer, C., Balbuena, J. A., Blasco-Costa, I., Karvonen, A., Sarabeev, V. & Jokela, J. Sensitivity of bipartite network analyses to incomplete sampling and taxonomic uncertainty. en. *Ecology* **104,** e3974. doi:10.1002/ecy.3974 (2023).
6. Peralta, G., CaraDonna, P. J., Rakosy, D., Fründ, J., Pascual Tudanca, M. P., Dormann, C. F., Burkle, L. A., Kaiser-Bunbury, C. N., Knight, T. M., Resasco, J., Winfree, R., Blüthgen, N., Castillo, W. J. & Vázquez, D. P. Predicting plant-pollinator interactions: concepts, methods, and challenges. en. *Trends in ecology & evolution* **39,** 494–505. doi:10.1016/j.tree.2023.12.005 (2024).
7. Derocles, S. A. P., Lunt, D. H., Berthe, S. C. F., Nichols, P. C., Moss, E. D. & Evans, D. M. Climate warming alters the structure of farmland tritrophic ecological networks and reduces crop yield. en. *Molecular ecology* **27,** 4931–4946. doi:10.1111/mec.14903 (2018).
8. Poisot, T., Ouellet, M.-A., Mollentze, N., Farrell, M. J., Becker, D. J., Brierley, L., Albery, G. F., Gibb, R. J., Seifert, S. N. & Carlson, C. J. Network embedding unveils the hidden interactions in the mammalian virome. en. *Patterns (New York, N.Y.)* **4,** 100738. doi:10.1016/j.patter.2023.100738 (2023).
9. Tylianakis, J. M., Didham, R. K., Bascompte, J. & Wardle, D. A. Global change and species interactions in terrestrial ecosystems. en. *Ecology letters* **11,** 1351–1363. doi:10.1111/j.1461-0248.2008.01250.x (2008).
10. Pearse, I. S. & Altermatt, F. Predicting novel trophic interactions in a non-native world. en. *Ecology letters* **16,** 1088–1094. doi:10.1111/ele.12143 (2013).
11. Dallas, T., Park, A. W. & Drake, J. M. Predicting cryptic links in host-parasite networks. en. *PLoS computational biology* **13,** e1005557. doi:10.1371/journal.pcbi.1005557 (2017).
12. Desjardins-Proulx, P., Laigle, I., Poisot, T. & Gravel, D. Ecological interactions and the Netflix problem. en. *PeerJ* **5,** e3644. doi:10.7717/peerj.3644 (2017).

13. Pichler, M., Boreux, V., Klein, A.-M., Schleuning, M. & Hartig, F. Machine learning algorithms to infer trait-matching and predict species interactions in ecological networks. en. *Methods in ecology and evolution / British Ecological Society* **11,** 281–293. doi:10.1111/2041-210x.13329 (2020).

14. Terry, J. C. D. & Lewis, O. T. Finding missing links in interaction networks. en. *Ecology* **101,** e03047. doi:10.1002/ecy.3047 (2020).

15. Strydom, T., Catchen, M. D., Banville, F., Caron, D., Dansereau, G., Desjardins-Proulx, P., Forero-Muñoz, N. R., Higino, G., Mercier, B., Gonzalez, A., Gravel, D., Pollock, L. & Poisot, T. A roadmap towards predicting species interaction networks (across space and time). en. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **376,** 20210063. doi:10.1098/rstb.2021.0063 (2021).

16. Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M. J., Fortin, M.-J., Hemming, V., Mercier, B., Pollock, L. J., Runghen, R., Dalla Riva, G. V. & Poisot, T. Food web reconstruction through phylogenetic transfer of low-rank network representation. en. *Methods in ecology and evolution* **13,** 2838–2849. doi:10.1111/2041-210x.13835 (2022).

17. Biton, B., Puzis, R. & Pilosof, S. Inductive link prediction facilitates the discovery of missing links and enables cross-community inference in ecological networks. en. *Nature ecology & evolution.* doi:10.1038/s41559-025-02715-6 (2025).

18. Guimerà, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. en. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 22073–22078. doi:10.1073/pnas.0908366106 (2009).

19. Stock, M., Poisot, T., Waegeman, W. & De Baets, B. Linear filtering reveals false negatives in species interaction data. en. *Scientific reports* **7,** 45908. doi:10.1038/srep45908 (2017).

20. Rossi, A., Firmani, D., Merialdo, P. & Teofili, T. *Explaining link prediction systems based on knowledge graph embeddings* in *Proceedings of the 2022 International Conference on Management of Data* (ACM, New York, NY, USA, 2022). doi:10.1145/3514221.3517887.

21. Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M., Fortin, M.-j., Mercier, B., Pollock, L. J., Runghen, R., Dalla Riva, G. V. & Poisot, T. Graph embedding and transfer learning can help predict potential species interaction networks despite data limitations. *Methods in ecology and evolution / British Ecological Society* **14,** 2917–2930. doi:10.1111/2041-210X.14228 (2023).

22. Dansereau, G., Barros, C. & Poisot, T. Spatially explicit predictions of food web structure from regional-level data. en. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **379,** 20230166. doi:10.1098/rstb.2023.0166 (2024).

23. Quintero, E., Isla, J. & Jordano, P. Methodological overview and data-merging approaches in the study of plant–frugivore interactions. en. *Oikos (Copenhagen, Denmark)* **2022.** doi:10.1111/oik.08379 (2022).

24. Cuff, J. P., Deivarajan Suresh, M., Dopson, M. E. G., Hawthorne, B. S. J., Howells, T., Kitson, J. J. N., Miller, K. A., Xin, T. & Evans, D. M. en. in *Advances in Ecological Research* 1–34 (Elsevier, 2023). doi:10.1016/bs.aecr.2023.09.002.

25. Olesen, J. M., Bascompte, J., Dupont, Y. L., Elberling, H., Rasmussen, C. & Jordano, P. Missing and forbidden links in mutualistic networks. en. *Proceedings. Biological sciences / The Royal Society* **278,** 725–732. doi:10.1098/rspb.2010.1371 (2011).

26. González-Varo, J. P. & Traveset, A. The Labile Limits of Forbidden Interactions. *Trends in Ecology & Evolution. Multiplicity in Unity: Plant Subindividual Variation and Interactions with AnimalsHandbook of the Birds of the World (Vol. 2)The Complete Birds of the Western Palearctic (CD-ROM)Genetics and Analysis of Quantitative Traits* **31,** 700–710. doi:10.1016/j.tree.2016.06.009 (2016).

27. Catchen, M. D., Poisot, T., Pollock, L. J. & Gonzalez, A. The missing link: Discerning true from false negatives when sampling species Interaction networks. *EcoEvoRxiv.* doi:10.32942/x2dw22 (2023).

28. Banville, F., Strydom, T., Blyth, P. S. A., Brimacombe, C., Catchen, M. D., Dansereau, G., Higino, G., Malpas, T., Mayall, H., Norman, K., Gravel, D. & Poisot, T. Deciphering probabilistic species interaction networks. en. *Ecology Letters* **28,** e70161. doi:10.1111/ele.70161 (2025).

29. Poelen, J. H., Simons, J. D. & Mungall, C. J. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecological informatics* **24,** 148–159. doi:10.1016/j.ecoinf.2014.08.005 (2014).

30. Fortuna, M. A., Ortega, R. & Bascompte, J. The web of life. *arXiv [q-bio.PE]* (2014).

31. Farrell, M. J., Elmasri, M., Stephens, D. A. & Davies, T. J. Predicting missing links in global host-parasite networks. en. *The Journal of animal ecology* **91,** 715–726. doi:10.1111/1365-2656.13666 (2022).

32. Stock, M., Piot, N., Vanbesien, S., Meys, J., Smagghe, G. & De Baets, B. Pairwise learning for predicting pollination interactions based on traits and phylogeny. *Ecological modelling* **451,** 109508. doi:10.1016/j.ecolmodel.2021.109508 (2021).

33. Fu, X., Seo, E., Clarke, J. & Hutchinson, R. A. Link prediction under imperfect detection: Collaborative filtering for ecological networks. *IEEE Transactions on Knowledge and Data Engineering* **33,** 3117–3128. doi:10.1109/tkde.2019.2962031 (2021).

34. CaraDonna, P. J., Petry, W. K., Brennan, R. M., Cunningham, J. L., Bronstein, J. L., Waser, N. M. & Sanders, N. J. Interaction rewiring and the rapid turnover of plant-pollinator networks. en. *Ecology letters* **20,** 385–394. doi:10.1111/ele.12740 (2017).

35. Trøjelsgaard, K., Jordano, P., Carstensen, D. W. & Olesen, J. M. Geographical variation in mutualistic networks: similarity, turnover and partner fidelity. en. *Proceedings. Biological sciences / The Royal Society* **282,** 20142925. doi:10.1098/rspb.2014.2925 (2015).

36. Poisot, T., Stouffer, D. B. & Gravel, D. Beyond species: why ecological interaction networks vary through space and time. en. *Oikos* **124,** 243–251. doi:10.1111/oik.01719 (2015).

37. Abramov, K., Galai, G., Biton, B., Puzis, R. & Pilosof, S. Predicting ecological interactions across space through pairwise integration of latent network patterns. en. *Methods in Ecology and Evolution.* doi:10.1111/2041-210x.70357 (2026).

38. He, X., Ghasemian, A., Lee, E., Schwarze, A. C., Clauset, A. & Mucha, P. J. Link prediction accuracy on real-world networks under non-uniform missing-edge patterns. en. *PloS one* **19,** e0306883. doi:10.1371/journal.pone.0306883 (2024).

39. Yates, L. A., Aandahl, Z., Richards, S. A. & Brook, B. W. Cross validation for model selection: A review with examples from ecology. en. *Ecological monographs* **93.** doi:10.1002/ecm.1557 (2023).

40. Poisot, T. Guidelines for the prediction of species interactions through binary classification. en. *Methods in ecology and evolution / British Ecological Society* **14,** 1333–1345. doi:10.1111/2041-210x.14071 (2023).

41. Dormann, C. F., Castillo, W. J., Tudanca, M. P. P., CaraDonna, P., Burkle, L. A., Rakosy, D., Winfree, R., Zoller, L., Benadi, G., Kaiser-Bunbury, C. N., *et al.* Predicting interaction frequency in plant-pollinator networks (2025).

42. Pellissier, L., Albouy, C., Bascompte, J., Farwig, N., Graham, C., Loreau, M., Maglianesi, M. A., Melián, C. J., Pitteloud, C., Roslin, T., Rohr, R., Saavedra, S., Thuiller, W., Woodward, G., Zimmermann, N. E. & Gravel, D. Comparing species interaction networks along environmental gradients: Networks along environmental gradients. en. *Biological Reviews of the Cambridge Philosophical Society* **93,** 785–800. doi:10.1111/brv.12366 (2018).

43. Poisot, T., Canard, E., Mouillot, D., Mouquet, N. & Gravel, D. The dissimilarity of species interaction networks. en. *Ecology letters* **15,** 1353–1361. doi:10.1111/ele.12002 (2012).

44. Hervías-Parejo, S., Colom, P., Beltran Mas, R., E. Serra, P., Pons, S., Mesquida, V. & Traveset, A. Spatio-temporal variation in plant–pollinator interactions: a multilayer network approach. en. *Oikos (Copenhagen, Denmark)* **2023.** doi:10.1111/oik.09818 (2023).

45. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F. & Dormann, C. F. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. en. *Ecography* **40,** 913–929. doi:10.1111/ecog.02881 (2017).

46. Fortuna, M. A., Nagavci, A., Barbour, M. A. & Bascompte, J. Partner fidelity and asymmetric specialization in ecological networks. en. *The American naturalist* **196,** 382–389. doi:10.1086/709961 (2020).

47. Dorado, J., Vázquez, D. P., Stevani, E. L. & Chacoff, N. P. Rareness and specialization in plant-pollinator networks. en. *Ecology* **92,** 19–25. doi:10.1890/10-0794.1 (2011).

48. Ward, C. A., Tunney, T. D., Hale, K. R. S., O'Connor, R. F. & McCann, K. S. The rewiring of ecological networks in a variable world. en. *Nature Reviews Biodiversity* **2,** 355–369. doi:10.1038/s44358-026-00159-9 (2026).

49. Peralta, G., Vázquez, D. P., Chacoff, N. P., Lomáscolo, S. B., Perry, G. L. W. & Tylianakis, J. M. Trait matching and phenological overlap increase the spatio-temporal stability and functionality of plant-pollinator interactions. en. *Ecology letters* **23,** 1107–1116. doi:10.1111/ele.13510 (2020).

50. García, Y., Giménez-Benavides, L., Iriondo, J. M., Lara-Romero, C., Méndez, M., Morente-López, J. & Santamaría, S. Addition of nocturnal pollinators modifies the structure of pollination networks. en. *Scientific reports* **14,** 1226. doi:10.1038/s41598-023-49944-y (2024).

51. Strydom, T., Dunhill, A. M., Dunne, J. A., Poisot, T. & Beckerman, A. P. Scaling from Metawebs to Realised Webs: A Hierarchical Approach to Network Ecology (2026).

52. Trøjelsgaard, K. & Olesen, J. M. Ecological networks in motion: micro- and macroscopic variability across scales. en. *Functional ecology* **30,** 1926–1935. doi:10.1111/1365-2435.12710 (2016).

53. Miller, D. A., Nichols, J. D., Mcclintock, B. T., Grant, E. H. C., Bailey, L. L. L. & Weir, L. A. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. en. *Ecology* **92,** 1422–1428. doi:10.1890/10-1396.1 (2011).

54. Chambert, T., Miller, D. A. W. & Nichols, J. D. Modeling false positive detections in species occurrence data under different study designs. en. *Ecology* **96,** 332–339. doi:10.1890/14-1507.1 (2015).

55. Brimacombe, C., Bodner, K., Michalska-Smith, M., Poisot, T. & Fortin, M.-J. Shortcomings of reusing species interaction networks created by different sets of researchers. en. *PLoS Biology* **21,** e3002068. doi:10.1371/journal.pbio.3002068 (2023).

56. Serra-Marin, P. E., Solé-Ribalta, A., Lana, A., Borge-Holthoefer, J., Hervías-Parejo, S. & Traveset, A. Comparative assessment of automated and manual monitoring in comprehensive plant–pollinator communities. en. *Methods in Ecology and Evolution* **16,** 2960–2978. doi:10.1111/2041-210x.70165 (2025).

57. Oreskes, N. Evaluation (not validation) of quantitative models. en. *Environmental health perspectives* **106 Suppl 6,** 1453–1460. doi:10.1289/ehp.98106s61453 (1998).

58. Benadi, G., Dormann, C. F., Fründ, J., Stephan, R. & Vázquez, D. P. Quantitative Prediction of Interactions in Bipartite Networks Based on Traits, Abundances, and Phylogeny. en. *The American naturalist* **199,** 841–854. doi:10.1086/714420 (2022).

59. Vázquez, D. P., Melián, C. J., Williams, N. M., Blüthgen, N., Krasnov, B. R. & Poulin, R. Species abundance and asymmetric interaction strength in ecological networks. en. *Oikos (Copenhagen, Denmark)* **116,** 1120–1127. doi:10.1111/j.0030-1299.2007.15828.x (2007).

60. Royle, J. A. N-mixture models for estimating population size from spatially replicated counts. en. *Biometrics* **60,** 108–115. doi:10.1111/j.0006-341X.2004.00142.x (2004).

61. Adhurya, S. & Park, Y.-S. A novel method for predicting ecological interactions with an unsupervised machine learning algorithm. en. *Methods in ecology and evolution* **15,** 1247–1260. doi:10.1111/2041-210x.14358 (2024).

Supplementary Information for:

Contextual evidence for classifying interactions and guiding their discovery

Kesem Abramov, Shir Miryam Nehoray, Rami Puzis, Peter J. Mucha, Shai Pilosof

Purpose and scope

This Supplement gives the formal version and definitions for the categorisation used in the main text in Section S1, and a worked example of its simplest version, the one presented in the main text, in Section S2. That version rests on six assumptions, which we state in Section S3. The sections that follow relax them in turn: Section S4 derives the error rates from sampling and realisation, Section S5 retains the full count of replicate detections rather than a single bit, Section S6 lets links differ from one another, Section S7 lets the local observation and the replicates err together, and Section S8 replaces the uniform prior. From Section S4 onwards the illustrations use directional error rates for the ecological sampling, so their values differ from those of the worked example. The illustrative error rates used throughout are chosen to show the behaviour of the framework, not to describe any particular system. Finally, Section S9 reviews what it would take to measure each of the four underlying quantities, a long-discussed problem in ecology.

S1 Formal setup

S1.1 Definitions

Table S1 provides all the notation used in this document. We follow the terminology of Banville et al. [1], who separate a regional metaweb of potential interactions M, which records whether two taxa are biologically capable of interacting, from its ecological restriction M^* and from the local network of realised interactions. Writing L_{ijk} for the realisation of a link between taxa i and j in local context k, a link is **feasible** when $M_{ij} = 1$ and **realised** when $L_{ijk} = 1$, with $P(L_{ijk} = 1 \mid M_{ij} = 0) = 0$: a link that is not feasible cannot be realised anywhere. Because every quantity below refers to one candidate link, we drop the indices i, j and k from here on and write the evidence, the categories and the error rates without them, with L_l for the realisation where we sampled and L_r for whether it is realisable within the space of the considered replicates (but not necessarily realised therein).

In the main text we place each candidate link into one of eight categories using three binary pieces (**bits**) of evidence: the model prediction Y, the local observation O_l, and the replicate observation O_r, each equal to 1 (predicted or observed) or 0. We collect them as $E = (Y, O_l, O_r)$. The eight categories correspond one to one with the eight possible error-free evidence patterns (Table S2). We write $z_c = (z_{c,Y}, z_{c,l}, z_{c,r})$ for the pattern that category c produces when no evidence source errs, and $z_{c,d}$ for its value on source $d \in \{Y, l, r\}$. These components are the underlying truths of Fig. S1, with $z_{c,Y} = \hat{Y}$, $z_{c,l} = L_l$ and $z_{c,r} = L_r$, so that $C = (\hat{Y}, L_l, L_r)$.

Table S1: Notation and definitions. Every quantity refers to a single candidate link, whose taxon and context indices are suppressed throughout (Section S1).

Symbol	Meaning
Truths	
M, M^*	Metaweb of potential interactions and its ecological restriction [1].
L_l, L_r	Realisation of the link where we sampled (L_l), and whether it is realisable in the replicate systems (L_r). A given replicate realises a realisable link at rate ρ (Section S4).
ρ	Realisation rate: probability that a feasible link is realised in a given context. It is ecological, not methodological. Importantly, within our notation, $(\rho > 0) \Leftrightarrow (L_r = 1)$.
$\hat{Y}$	Prediction (1 or 0) of the model that would be obtained by fitting to the truth rather than to the observations. It is never available.
C	The true (latent) category of the link, $C = (\hat{Y}, L_l, L_r)$, one of eight.
$z_{c,d}$	Value (bit) of component $d \in \{Y, l, r\}$ for category c; $(z_{c,Y}, z_{c,l}, z_{c,r})$ is the category's signature.
$P(C = c)$	Prior probability of category c; uniform, $P(c) = 1/8$, unless stated.
Evidence	
Y	Prediction (1 or 0) of the model at hand, which is fitted to the observations.
O_l	Local observation, 1 if the link was recorded where we sampled.
O_r	Replicate observation, 1 if the link was recorded in at least one replicate.
R, n	Number of replicates, and the number of them in which the link was recorded.
E	The evidence, $E = (Y, O_l, O_r)$, or $E = (Y, O_l, n, R)$ from Section S5.
q	Calibrated model probability for the link, of which Y is the thresholded version (Section S8).
Error rates	
ϵ_d	Error rate of source $d \in \{Y, l, r\}$, one per source, relative to its respective underlying truth. Sections S1 and S2 use this single rate; from Section S4 each source has two rates, one per direction of error.
ϵ_Y	Rate at which Y differs from $\hat{Y}$, describing error due to fitting model to observations (cf. truth).
δ	Error of Y against the observations, $\delta = 1 -$ accuracy, e.g. from cross-validation.
$\hat{\delta}$	Error of $\hat{Y}$ against the underlying truth. It is never available.
ϵ_l	Miss rate: probability that a link realised where we sampled goes unrecorded.
f	False-detection rate: probability that a link that is not there is erroneously recorded.
ϵ_r	Probability that a link realised in at least one replicate is recorded in none of them. It is not a primitive but follows from ϵ_l, f and ρ (Section S4).
ϵ_d^+, ϵ_d^-	Rate at which source d errs on a link that is truly present (+, a false negative) or truly absent (−, a false positive).
$\epsilon_d(z_{c,d})$	The same pair written as one state-dependent rate: $\epsilon_d(1) = \epsilon_d^+$, $\epsilon_d(0) = \epsilon_d^-$.
Derived quantities	
p_1, p_0	Per-replicate probability of recording a link that is realisable in the system, $p_1 = \rho(1 - \epsilon_l)$ (i.e., realised and detected), or unrealisable within the replicates, $p_0 = f$ (i.e., falsely detected).
κ	Maximum contextual confidence, the ceiling that the model and local errors set on the posterior. Under symmetric rates, $(1 - \epsilon_Y)(1 - \epsilon_l)$. Under asymmetric rates ($\epsilon_l \neq f$): $(1 - \epsilon_Y)(1 - f)/((1 - f) + \epsilon_l)$ when $O_l = 0$, and $(1 - \epsilon_Y)(1 - \epsilon_l)/((1 - \epsilon_l) + f)$ when $O_l = 1$.
ϕ	Feasibility confidence, a lower bound on the posterior probability that the link is feasible.

S1.2 Errors in evidence

Because every source is error-prone, the observed E is a noisy reading of the true category's pattern. We draw this mapping for the model and local observation errors in Fig. S1, where each arrow runs from a quantity to the quantity it recovers and carries the error rate of that recovery. The replicates add a further source of error, since a realisable link need not be realised in all, or even any, of the R selected replicates. The model prediction's error against the underlying truth is the diagonal of the square. It is never measured directly, but it is useful to recognize it as the model error δ composed with the observation errors ϵ_l and f (Table S3).

Given the errors, we can ask what is the probability of a link's true category, given the evidence. To do that, we treat the true category C as unknown and obtain its posterior by Bayes' theorem,

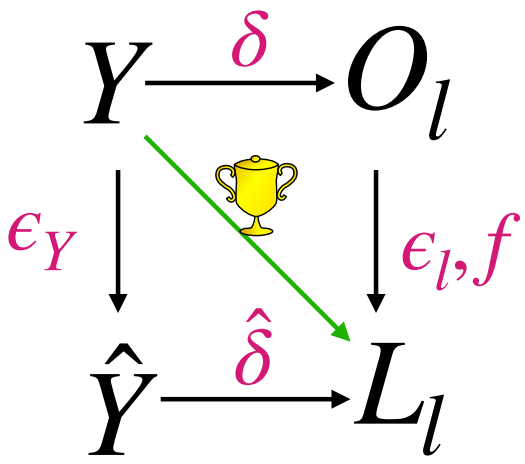

Fig. S1: How the evidence maps onto the truth through the error rates. Each arrow runs from a quantity to the quantity it is trying to recover, and its label is the error rate of that recovery. From Y to O_l: the model we have was fitted to the local observation and departs from it at rate $\delta = 1 -$ accuracy, which is what standard within-model evaluation returns. From O_l to L_l: the observation recovers a realised link with miss rate ϵ_l and false-detection rate f. From Y to $\hat{Y}$: the prediction we have differs at rate ϵ_Y from the one that fitting to the underlying truth would have given. From $\hat{Y}$ to L_l: $\hat{\delta}$ is the counterpart of δ for a model that was fitted to the underlying truth. The green diagonal from Y to L_l is the quantity of interest (marked by the trophy), the error of the prediction against the underlying truth. It is never measured directly, because L_l is never seen, but composing $Y \to O_l$ with $O_l \to L_l$ gives it, which is the lower block of Table S3. Note that O_r does not appear here because, as discussed in this document, the errors of each separate element (e.g., location) of O_r are those of O_l.

$$P(C \mid E) = \frac{P(E \mid C)\, P(C)}{P(E)}, \qquad P(E) = \sum_{c} P(E \mid C = c)\, P(c). \tag{S1}$$

Assuming the sources err independently given the category, the likelihood factorises over the three sources, each contributing $1 - \epsilon_d$ when the observed value matches the category's pattern and its error rate ϵ_d when it differs, which we can compactly write as

$$P(E \mid C = c) = \prod_{d \in \{Y,l,r\}} \epsilon_d^{\mathbf{1}[E_d \neq z_{c,d}]} \, (1 - \epsilon_d)^{\mathbf{1}[E_d = z_{c,d}]}. \tag{S2}$$

The error rates describe the uncertainty from the model (ϵ_Y) and the observations (ϵ_l, f and ϵ_r). Unless stated otherwise we use a uniform prior $P(C) = 1/8$ for all C. We discuss these assumptions from Section S3 onwards.

Table S2: The eight categories and their evidence signatures.

Category	Y	O_l	O_r
Recurrent	1	1	1
Locally unique	1	1	0
Possibly missing	1	0	1
Phantom	1	0	0
Model-elusive	0	1	1
Weakly-supported	0	1	0
Locally absent	0	0	1
Possibly forbidden	0	0	0

Table S3: Confusion matrices for the errors of the local observation and for the model prediction, both against the underlying truth of the local link. Columns give the true state, rows the reported one, and each column sums to one separately for each block. The upper block is the detection channel, with miss rate ϵ_l and false-detection rate f; the lower block adds the model error $\delta = 1 -$ accuracy, the rate at which the prediction departs from the observation it was fitted to. A model that is perfect for the observations, $\delta = 0$, recovers the upper block and not the underlying truth, inheriting the detection error in full, while perfect detection, $\epsilon_l = f = 0$, leaves only model error.

		Truth	
		$+$	$-$
O_l	$+$	$1 - \epsilon_l$	f
	$-$	ϵ_l	$1 - f$
Y	$+$	$(1-\epsilon_l)(1-\delta) + \epsilon_l\delta$	$f(1-\delta) + (1-f)\delta$
	$-$	$(1-\epsilon_l)\delta + \epsilon_l(1-\delta)$	$(1-f)(1-\delta) + f\delta$

S2 Worked example: binary evidence and a uniform prior

We work through the simplest case to develop an intuitive understanding of the idea. Consider a link that is predicted, unobserved locally, but observed in the replicates, so the evidence is $E = (1, 0, 1)$: under error-free conditions, this is the signature of a *possibly missing* link. We take illustrative rates $\epsilon_Y = 0.2$, $\epsilon_l = f = 0.3$, and $\epsilon_r = 0.1$.

Each category's likelihood is a product of three factors (Eq. S2), one per source, contributing $1 - \epsilon_d$ where the category's signature matches the evidence and ϵ_d where it does not. For the *possibly missing* category itself, all three bits match, giving

$$P(E = (1,0,1) \mid C = \text{possibly missing}) = (1-\epsilon_Y)(1-\epsilon_l)(1-\epsilon_r) = 0.8 \times 0.7 \times 0.9 = 0.504. \tag{S3}$$

Any other alternative category disagrees with the evidence on one or more sources, and, because of the symmetric error rates in this example (relaxed below), its likelihood is the same product with $1-\epsilon_d$ replaced by ϵ_d on each disagreeing source. The likelihood therefore falls as the number of disagreeing sources grows, provided every rate is below one half, and falls more steeply when the disagreeing source is the more reliable one. For a possibly missing link, the three categories that differ on a single source are the strongest competitors: *recurrent* (differs on O_l), *locally absent* (differs on Y), and *phantom* (differs on O_r); for instance $P(E = (1,0,1) \mid \text{recurrent}) = (1-\epsilon_Y)\epsilon_l(1-\epsilon_r) = 0.8 \times 0.3 \times 0.9 = 0.216$. Categories that differ on two or three sources yield even lower likelihoods.

Under the assumption of a uniform prior, $P(C) = 1/8$ (relaxed below), the posterior and the likelihood become proportional to one another. Moreover, by the assumed form of the likelihood in Eq. S2, the eight likelihoods for a specified evidence E conditional on different C sum to one. It thus follows that the denominator of Eq. S1 is $P(E) = 1/8$ and the posterior equals the likelihood. The resulting posterior is given in Table S4. The category whose signature matches the evidence is the most probable assignment, the three single-error neighbours account for most of the remainder, and the four multi-error categories together sum to under 10%. **To develop intuition, we recommend interactively playing with this example in the online version in `http://lpguide.ecomplab.com`, where it is also possible to shift the error rates to change this illustrative example.**

Table S4: Posterior over the eight categories for binary evidence $E = (1, 0, 1)$ under a uniform prior, with $\epsilon_Y = 0.2$, $\epsilon_l = 0.3$, $\epsilon_r = 0.1$. Mismatched sources are relative to the observed evidence. Because of the uniform prior and the assumed form of the likelihood, the posterior equals the likelihood.

Category	Signature $(\hat{Y}, L_l, L_r)$	Mismatched sources	Posterior
Possibly missing	$(1, 0, 1)$	none	50.4%
Recurrent	$(1, 1, 1)$	local	21.6%
Locally absent	$(0, 0, 1)$	model	12.6%
Phantom	$(1, 0, 0)$	replicates	5.6%
Model-elusive	$(0, 1, 1)$	model, local	5.4%
Locally unique	$(1, 1, 0)$	local, replicates	2.4%
Possibly forbidden	$(0, 0, 0)$	model, replicates	1.4%
Weakly-supported	$(0, 1, 0)$	model, local, replicates	0.6%

Three features of this example are worth noticing to generalise. First, under a uniform prior and $\epsilon_l = f$, every category attains the same probability $(1 - \epsilon_Y)(1 - \epsilon_l)(1 - \epsilon_r)$ when the evidence matches its signature. Hence, the confidence in the matching category, Eq. S3, depends only on the error rates and not on which evidence pattern was observed.

Second, the local and model errors set a limit on the confidence even under error-free replicate evidence: as $\epsilon_r \to 0$ the confidence rises to the *maximum contextual confidence*, which we denote as κ. For the selected parameters, $\kappa = (1 - \epsilon_Y)(1 - \epsilon_l) = 0.56$ is the most that within-system replication alone can deliver, bounded by the model and local-sampling errors. Increasing κ is not a matter of more replicates but of methodological evidence: a better model to lower ϵ_Y, or a complementary method (or increased sampling effort) to lower ϵ_l. Fig. S2 gives κ over a range of both rates, for the symmetric case used here and for the directional rates of Section S4.

Third, the framework also aggregates categories into ecologically meaningful summaries. The *feasibility confidence* $\phi = 1 - P(\text{phantom} \mid E) - P(\text{possibly forbidden} \mid E)$ is the posterior probability that the link is either realised locally or (at least) realisable in the replicate systems, since *phantom* and *possibly forbidden* are the two categories in which $L_l = L_r = 0$. Equivalently, $\phi = P(\text{TP}) + P(\text{FN}) + P(\textit{possibly missing} \mid E) + P(\textit{locally absent} \mid E)$: every category in which the link is realised where we sampled, whether the model predicted it (TP) or not (FN), together with the two in which it is not realised locally but is realisable in the replicates. Ecologically, it is a lower bound on the posterior probability that the link is feasible, since a feasible link may go unrealised everywhere we looked. In this example, $\phi = 93.0\%$.

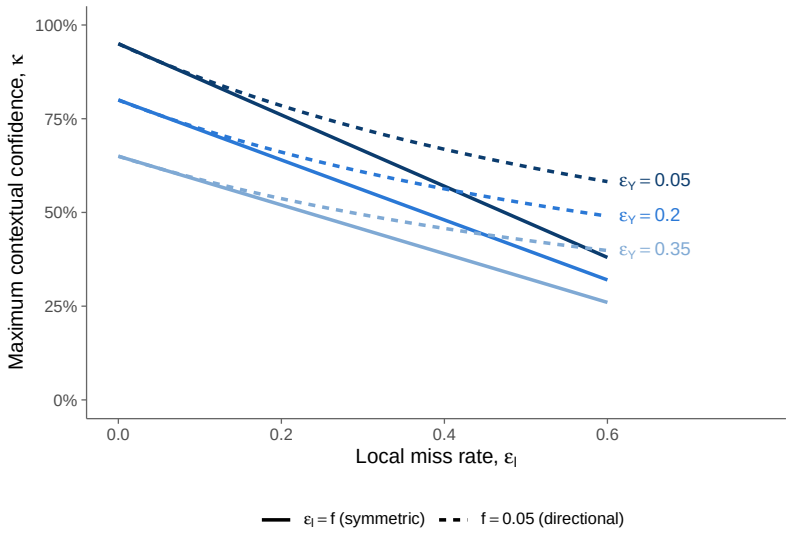

Fig. S2: Only a better model or a better local method raises the maximum contextual confidence. κ against the local miss rate ϵ_l, for three model error rates ϵ_Y. Solid lines are the symmetric case ($\epsilon_l = f$), $\kappa = (1-\epsilon_Y)(1-\epsilon_l)$, used in this section; dashed lines are the directional rates with $O_l = 0$ and $f = 0.05$, $\kappa = (1-\epsilon_Y)(1-f)/((1-f)+\epsilon_l)$, used from Section S4 onwards. The two forms agree at $\epsilon_l = f$ and diverge above it, because a low f keeps a local non-detection weak evidence of absence.

S3 Assumptions

The version in Section S1 (and in the main text) is deliberately minimal, resting on the following assumptions.

1. **Symmetric observation errors.** Failing to detect a link and accidentally marking a link that is not there are two different errors (ϵ_l vs. f): they have different origins and behave differently as effort increases. We deal with this in Section S4.

2. **Binary evidence.** Each source is reduced to a single bit (Section S1). In particular O_r records only whether the link was seen in at least one replicate, discarding how many. The discarded count is, however, informative. We deal with this in Sections S4 and S5. Similarly, we can use model outcome probabilities instead of setting a threshold, which we detail in Section S8.

3. **Interchangeable replicates.** A replicate is treated as a repeat of the same observation within the replicate set, so a link is either there to be seen in all of them or in none. A regionally feasible link is, however, realisable in some contexts and not in others, which is what separates, for example, *locally absent* (feasible but unrealised in a patch) from *possibly forbidden* (infeasible anywhere) links. We introduce the realisation rate ρ for this in Section S4.

4. **Homogeneous links.** One set of rates applies to all links. These rates are not arbitrary, since the false-negative rate of interaction sampling can be estimated from sampling effort and species relative abundances [2], as we discuss in Section S9. Links differ, however, in how often they are realised and in their detectability. We deal with this in Sections S6 and S7.

5. **Independent errors.** The sources err independently given the category, which is what lets the likelihood factorise (Eq. S2). However, evidence sources can share biases, and the local observation and the replicates are usually the same method applied in different places, so they share its blind spots. A feasible link the method cannot detect then stays absent locally and in every replicate, masquerading as *possibly forbidden*. We discuss this in Section S7. Furthermore, the predictive model is assumed to err independently of the observations, even though its error rate normally comes from cross-validation against those same observations (Box 1 in the main text). Blocking the withholding by site or clade rather than at random [3],

and keeping the replicate networks out of the training data, are design choices that make this assumption more defensible, as we discuss in Section S9.

6. **Uniform prior.** All eight categories are equally probable a priori. This is rarely true: interactions are sparse, so the categories in which the link is absent are more probable before any evidence is seen, and a model score is itself prior information about the link. Co-occurrence, abundance, generalism, trait matching and phylogeny all bear on how probable a link is before it is sampled, and can inform the prior instead. A non-uniform prior also removes the arithmetic convenience of the worked example in Section S2, since the posteriors are then no longer simply proportional to the likelihoods. We deal with this in Section S8. Another issue regards the use of threshold to decide if model output predict or not interactions. We discuss using the model score instead of a thresholded bit to carry per-link model confidence in Section S8.

S4 Grounding the error rates in sampling and realisation

The three rates ϵ_d of Section S1 are a convenient description. However, underlying them are four quantities. Three are methodological: the sampling miss rate ϵ_l, the probability that a link realised where we sampled goes unrecorded; the false-detection rate f, the probability that a link that is not there is erroneously recorded through misidentification or contamination [4, 5]; and the model error ϵ_Y, the rate at which the prediction obtained through fitting to observations differs from what fitting to the truth would have given. The fourth quantity is ecological rather than methodological: the realisation rate ρ, the probability that a link is realised in a given place, a form of aleatoric variability [1]. This section derives the three ϵ_d from these four; estimating them is discussed in Section S9.

The local observation is one method with two directions of failure: a link realised where we sampled is missed with probability ϵ_l, while a link that is locally absent is recorded as present at the false-detection rate f. The realisation rate does not enter, because local realisation is not marginalised over but is the latent state itself.

In a replicate the same method applies, but realisation is now marginalised over, so a replicate observation is the product of two ecologically distinct steps: whether a feasible link is *realised* in that replicate, and whether, once realised, it is *detected*. The probability of being both realised in a given replicate and then detected is

$$p_1 = \rho\,(1 - \epsilon_l). \tag{S4}$$

For a link that is not realisable in the replicates ($z_{c,r} = 0$), a detection can still arise in a given replicate at the false-detection rate $p_0 = f$ (Table S5). The two directions of error are accordingly not symmetric in origin. Not detecting a realisable link in a replicate draws on both methodological and ecological properties, since $1 - p_1 = (1 - \rho) + \rho\,\epsilon_l$, while a spurious detection of a link that is not realised draws only on f. The ecological source dominates the lack of observation of a realisable link: under the illustrative values $\rho = 0.15$ and $\epsilon_l = 0.3$, non-realisation in a single given replicate contributes 0.85 of the $1 - p_1 = 0.895$ probability of not detecting the link in the replicate, while failed detection contributes only 0.045.

With R replicates, a realisable link ($L_r = 1$) goes unrecorded ($O_r = 0$) only if it is unobserved in every replicate, with probability $(1 - p_1)^R$ (Table S5). A truly absent and unrealisable link, meanwhile, is recorded as present if it is spuriously detected in at least one replicate[1]:

$$\epsilon_r^+(R) = (1 - p_1)^R, \qquad \epsilon_r^-(R) = 1 - (1 - p_0)^R. \tag{S5}$$

The single number ϵ_r presented in Section S1 is thus a simplification: the false-negative rate on the replicate

[1] For simplicity, in this calculation of ϵ_r^+ and in subsequent error rate calculations below, we do not account for the second-order "two wrongs make a right" possibilities, e.g. that a realisable link is spuriously detected in a replicate in which it is not realised, expecting these additional corrections to be relatively small in practice.

axis shrinks as replicates accumulate while the false-positive rate grows (Fig. S3). Since p_1 and p_0 are fixed, this trade-off comes from compressing accumulating evidence into a single bit rather than from the evidence itself.

Table S5: Confusion matrix for the replicate observation O_r against the truth L_r, whether the link is realisable within the space of the considered replicates. Columns give the true state, rows the observation, and each column sums to one. A realisable link is recorded in a given replicate with probability $p_1 = \rho(1 - \epsilon_l)$, which requires both realisation and detection, and an unrealisable one with probability $p_0 = f$, which requires a false detection. The two off-diagonal entries are the rates of Eq. S5. They are not symmetric in origin: the false negative $(1 - p_1)^R$ carries the ecological failure to be realised together with the methodological failure to detect, whereas the false positive $1 - (1 - p_0)^R$ is methodological alone. This is the replicate counterpart of the upper block of Table S3, from which it differs by marginalising over realisation, and to which it reduces when $R = 1$ and $\rho = 1$.

		L_r	
		$+$	$-$
O_r	$+$	$1 - (1 - p_1)^R$	$1 - (1 - p_0)^R$
	$-$	$(1 - p_1)^R$	$(1 - p_0)^R$

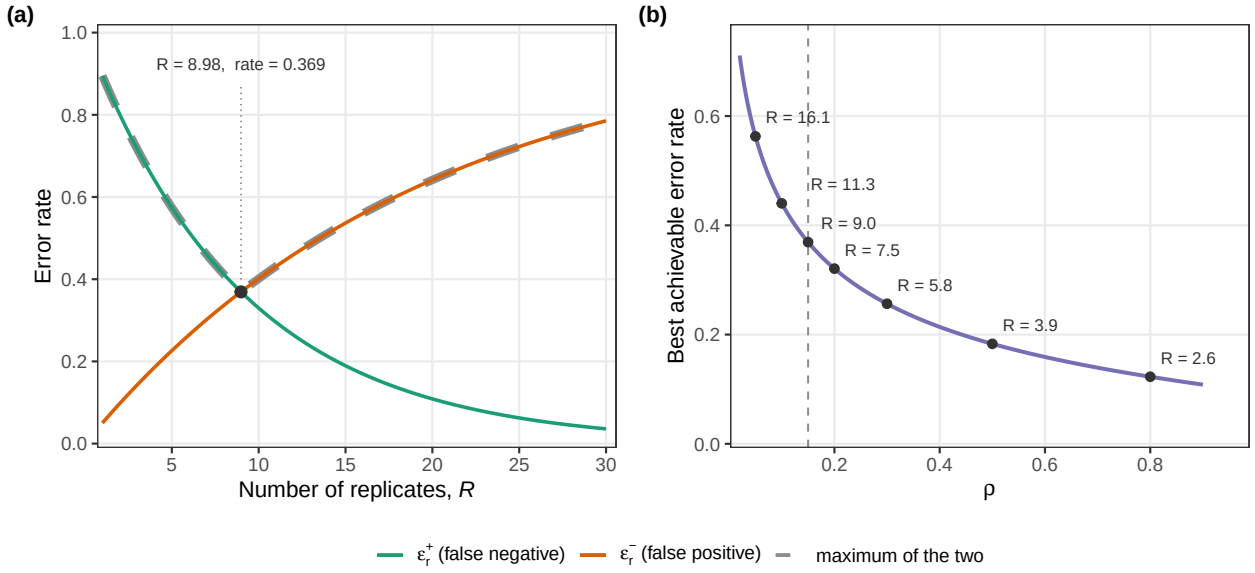

Fig. S3: The two directional error rates of the replicate axis pull against each other. **(a)** The false-negative rate $\epsilon_r^+(R) = (1 - p_1)^R$ falls as replicates accumulate while the false-positive rate $\epsilon_r^-(R) = 1 - (1 - p_0)^R$ rises. Example is shown for the illustrative values $\rho = 0.15$, $\epsilon_l = 0.3$ and $f = 0.05$, so that $p_1 = 0.105$ and $p_0 = 0.05$. The grey dashed line is the larger of the two rates at each R; its lowest point is where the two cross, at $R \approx 9$, where both equal 0.37. No number of replicates makes both errors small, and the best a single symmetric ϵ_r can be is 0.37 at these parameters. **(b)** The lowest error rate attainable in both directions, that is, the minimum over R of the larger of the two rates, which is the value of the grey dashed curve in (a) at its lowest point where $\epsilon_r^+ = \epsilon_r^-$, recomputed as the realisation rate ρ varies with $\epsilon_l = 0.3$ and $f = 0.05$ held fixed. Labelled points give the number of replicates at which it is reached and the dashed vertical line marks $\rho = 0.15$. Notably, no choice of R can put both errors below the value given by the curve, so what limits the replicate axis is how often the link is realised, not how many replicates are taken.

Writing $\epsilon_d^+ = P(E_d = 0 \mid z_{c,d} = 1)$ for a false negative and $\epsilon_d^- = P(E_d = 1 \mid z_{c,d} = 0)$ for a false positive, every rate in the framework now follows from the four underlying quantities ϵ_Y, ϵ_l, f and ρ (Table S6). Seen this way, the version of Section S1 is the special case in which each source's two rates are equated: it sets $f = \epsilon_l$ on the local axis and lets one ϵ_r stand for both replicate directions. One could of course also generalize the model error ϵ_Y to be asymmetric, though in the absence of detailed information about how the model changes due to fitting to the truth (cf. observations), it is difficult to estimate the direction and scale of asymmetric model error.

Table S6: The rates at which each evidence source errs, in terms of the four underlying quantities ϵ_Y, ϵ_l, ρ and f. R is the number of replicates.

Source	False negative ϵ_d^+	False positive ϵ_d^-
Model Y	ϵ_Y	ϵ_Y
Local O_l	ϵ_l	f
Replicates O_r	$(1-\rho(1-\epsilon_l))^R$	$1-(1-f)^R$

Letting $\epsilon_d(z_{c,d})$ denote the rate at which source d errs about a category whose signature bit is $z_{c,d}$, that is $\epsilon_d(1)=\epsilon_d^+$ and $\epsilon_d(0)=\epsilon_d^-$, the likelihood of Eq. S2 becomes

$$P(E \mid C=c) = \prod_{d\in\{Y,l,r\}} \epsilon_d(z_{c,d})^{\mathbf{1}[E_d\neq z_{c,d}]}\left(1-\epsilon_d(z_{c,d})\right)^{\mathbf{1}[E_d=z_{c,d}]}, \tag{S6}$$

which is Eq. S2 with the single rate ϵ_d replaced by one that depends on which way the error runs, and reduces to it when the two rates of each source are equal ($\epsilon_d^+=\epsilon_d^-=\epsilon_d$).

Introducing asymmetry in the direction of the errors no longer admits the arithmetic shortcut of Section S2: the eight likelihoods no longer sum to one, so $P(E)\neq 1/8$ and the posterior must be normalised explicitly (Eq. S1).

For the evidence $E=(1,0,1)$ of the worked example, with $\epsilon_Y=0.2$, $\epsilon_l=0.3$, $f=0.05$, $\rho=0.15$ and $R=5$ replicates, the likelihoods sum to 0.815 and $P(E)=0.102$. The consequences are substantial: the posterior probability for the *possibly missing* category falls from 50.4% to 39.7% and *phantom* rises from 5.6% to 21.1%, and the feasibility confidence falls from 93.0% to 73.6%.

Grounding the rates this way separates sampling error from the realisation rate and makes the role of R visible. One consequence is that the framework now requires a value for ρ, the only one of the four quantities that is not a property of the method. In exchange, ρ and f make it possible to retain the full count of detections rather than collapsing it to a single bit, so that confidence can accumulate as treated in the next section. That treatment inherits the two assumptions behind Eq. S5, that replicates are independent and share a common p_1. Section S6 lets $p1$ differ between links, and Section S7 lets the local observation and the replicates err together.

S5 Cumulative replicate evidence

Of the R replicates, the simple derivation from Section S1 retains only whether the count of detections was zero or not, yet the count n is available whenever the replicates are, and can be used with p_1 and p_0. With $E=(Y,O_l,n,R)$ the replicate factor of Eq. S6 becomes binomial:

$$P(E \mid C=c) = \left[\prod_{d\in\{Y,l\}} \epsilon_d(z_{c,d})^{\mathbf{1}[E_d\neq z_{c,d}]}\left(1-\epsilon_d(z_{c,d})\right)^{\mathbf{1}[E_d=z_{c,d}]}\right]\binom{R}{n} p_c^n (1-p_c)^{R-n}, \tag{S7}$$

where $p_c=p_1$ for the categories in which the link is realisable in the replicates, and $p_c=p_0$ for those in which it is not realisable. The binomial coefficient is the same for every category and cancels, so which replicates carried the observation is irrelevant and the count n is a sufficient summary of the replicate evidence. Where the link is not realisable, every record is a false detection at rate f, so no adjustment is needed. The binary version is the special case in which n is first passed through the rule $O_r=\mathbf{1}[n\geq 1]$.

The illustrations below show an example under the evidence $Y=1$ and $O_l=0$, using the directional rates of Section S4: $\epsilon_Y=0.2$, $\epsilon_l=0.3$, $f=0.05$ and $\rho=0.15$, so that $p_1=0.105$ and $p_0=0.05$. Evidence is consistent when every replicate points the same way, either $n=R$ or $n=0$, and these two scenarios bound what any mixed count can do (Fig. S4).

Under $n=R$ the posterior of *possibly missing* rises from 41.2% at $R=1$ to 54.9% at $R=3$ and 59.3% at $R=5$, approaching $\kappa=60.8\%$ by $R=10$ (Fig. S4a). At that point, the replicate axis is resolved completely and the

four surviving categories divide by their model and local factors alone. Under the binary version ($O_r = \mathbf{1}[n \geq 1]$) the same replicates push confidence the other way (Fig. S4a), from 41.2% at $R = 1$ to 39.7% at $R = 5$, the value reported in Section S4, and 35.4% at $R = 20$, converging on half of κ. This is because n is discarded but not R, which still enters through Eq. S5, so what declines is the contribution of each record rather than the amount of evidence: the more replicates were taken, the less surprising a single detection is under absence. The two summaries agree at $R = 1$.

The mirror scenario $n = 0$ (Fig. S4b) raises *phantom*, the replicate counterpart of *possibly missing*, to the same κ but far more slowly, reaching only 39.2% at $R = 10$. This is because a category and its counterpart differ only in the replicate term, so their log-odds grow linearly in R (for large R), but at very different rates: a detection is rare under either hypothesis and so is surprising, whereas an empty replicate is what both mostly predict.

Read as a design question, the same curves give R^*, the replicates needed to reach a target confidence: for a target of 50%, $R^* = 3$ under validation by detections (for a *possibly missing* link) and 26 under validation by absences (for a *phantom* link) (Fig. S4). The target must lie strictly below κ, since replicates buy only speed towards the maximum that ϵ_Y, ϵ_l and f impose; raising that maximum requires a better model, increased sampling effort or a complementary local method.

Equation S7 still treats replicates as independent draws sharing one p_1. Section S6 lets p_1 differ between links, and Section S7 treats the dependence between the local observation and the replicates.

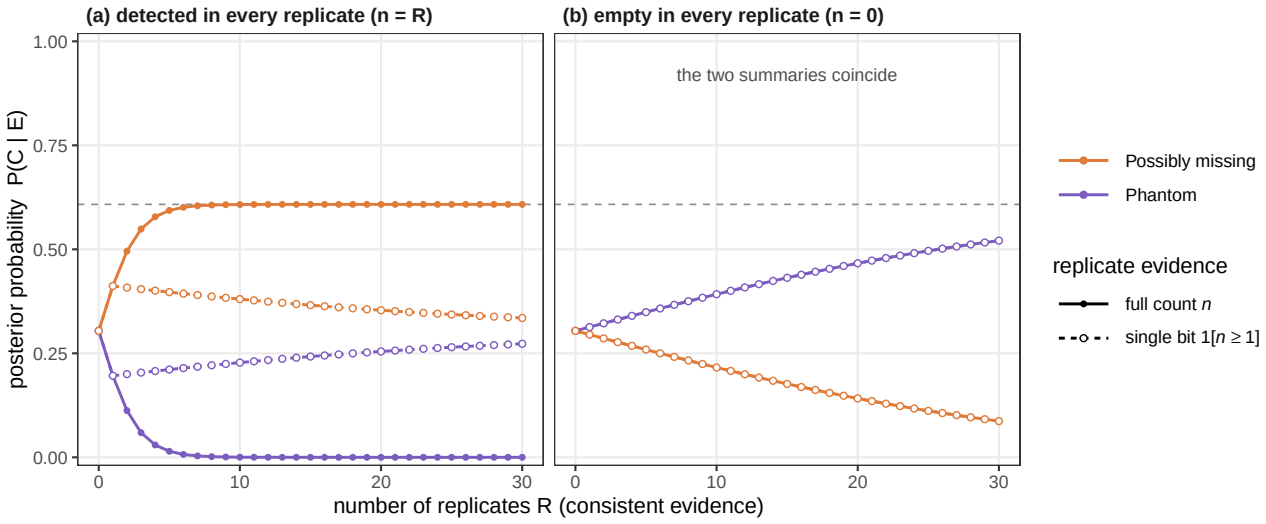

Fig. S4: Posterior probability as a function of the number of replicates R, for the evidence $Y = 1$, $O_l = 0$. At $R = 0$ there is no replicate term, so the categories pair up by $z_{c,r}$, each pair splits evenly, and both curves start at $\kappa/2 = 30.4\%$. Solid lines retain the full count n (Eq. S7); dashed lines pass it through the rule $O_r = \mathbf{1}[n \geq 1]$, which indicates if the link was observed at least once. The dashed curves still move with R (even though they are for the binary case) because their error rates do (Eq. S5): the more replicates stand behind a single record, the less that record contributes. **(a)** Detected in every replicate. The count carries *possibly missing* to the maximum contextual confidence $\kappa = 60.8\%$ (horizontal dashed line), which no number of replicates exceeds. In contrast, the single bit sends both categories back towards the $\kappa/2$ they began at, so that a bit drawn from many replicates ends up saying no more than no replicate at all. **(b)** Empty in every replicate. Here the two summaries are the same number, since a count of zero and an unset bit are the same statement, so the dashed curves lie on the solid ones. *Phantom* climbs towards the same κ, an order of magnitude more slowly. Rates are $\epsilon_Y = 0.2$, $\epsilon_l = 0.3$, $f = 0.05$ and $\rho = 0.15$, so that $p_1 = 0.105$ and $p_0 = 0.05$.

S6 Heterogeneous interactions

Equation S7 gives every link the same p_1. However, links are not homogeneous: both ϵ_l and ρ vary between them. In the replicate factor the two enter only through their product $p_1 = \rho(1 - \epsilon_l)$, so it is enough to let p_1 vary rather than to model them separately. Banville et al. [1] treat the number of networks in which an

interaction occurs as binomial with a constant interaction probability, and note that once that probability is itself uncertain, a beta distribution over it makes the count beta-binomial. The same holds here with detections in place of occurrences, so we take p_1 to be drawn once per link from $\mathrm{Beta}(\alpha, \beta)$, where α counts replicates in which the link was recorded and β those in which it was not, and hold it fixed across that link's replicates. The distribution is over links rather than over replicates, and for any one link it is a prior that its own detections update. We write the mean $\bar{p}_1 = \alpha/(\alpha + \beta)$ and the concentration $\nu = \alpha + \beta$, which sets how alike links are in their p_1: the larger ν, the narrower the distribution until, in the limit of large ν, all links share one rate and the common rate of Eq. S7 is recovered (Fig. S5a).

Because p_1 now differs between links, the replicate factor becomes beta-binomial. Only the categories in which the link is realisable in the replicates are affected, since $p_0 = f$ is methodological, and the count n remains sufficient because the binomial coefficient still cancels.

Heterogeneity in p_1 acts only on the replicate axis, so it tilts the probability between the four categories in which the link is realisable in the replicates (*recurrent*, *possibly missing*, *model-elusive* and *locally absent*) and the four in which it is not (*locally unique*, *phantom*, *weakly-supported* and *possibly forbidden*). Under empty replicates it tilts in favour of the first four, because a drawn rate always keeps weight near $p_1 = 0$ and a link with such a rate explains a run of empty replicates without difficulty, while under detections it changes almost nothing.

For example, consider a bee–plant interaction that is predicted by the model ($Y = 1$), is not observed at the focal meadow ($O_l = 0$), and is then looked for in ten further meadows and never recorded once ($O_r = 0$), the scenario of Fig. S5b at $R = 10$. That evidence is the signature of *phantom*, a predicted link neither realised locally nor realisable elsewhere. Its competitor is *possibly missing*, which holds that the interaction is realisable within the ten meadows but was missed every time. Under a common p_1 an empty run of ten becomes increasingly hard to reconcile with a link that is there, so *possibly missing* falls to 21.6%, the grey curve at $R = 10$, while *phantom* rises to 39.2%. Once links differ, a p_1 near zero is always available, and an interaction with such a rate explains the empty run without difficulty: at $\nu = 10$, the purple curve, *possibly missing* holds at 26.3% and *phantom* reaches only 34.5%. Empty replicates therefore stop distinguishing the two once links differ enough.

In this example, the local observation and the replicates are still treated as erring independently, which is the subject of Section S7.

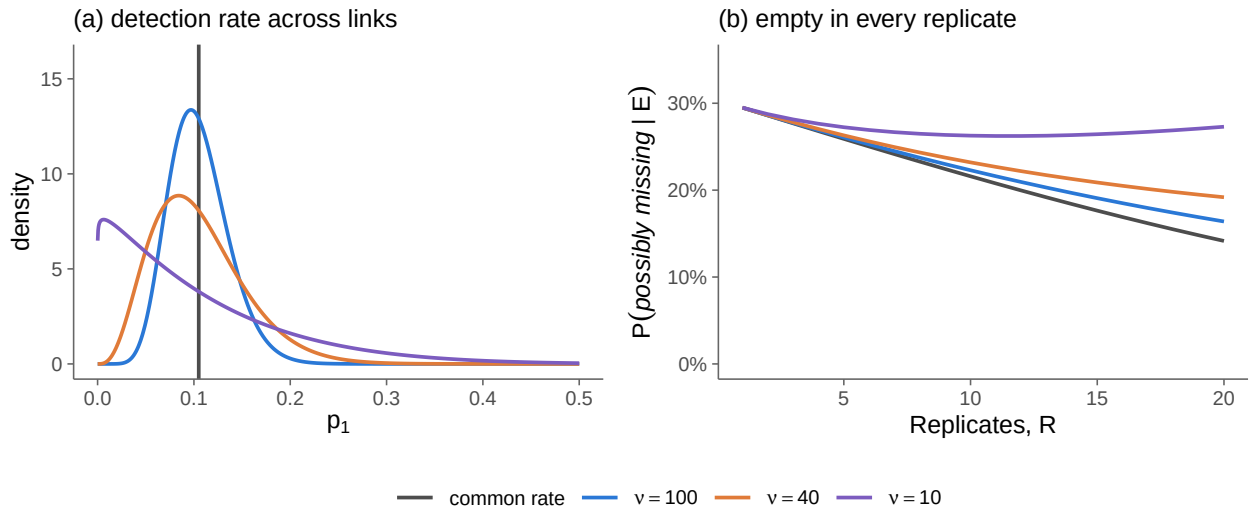

Fig. S5: Empty replicates tell a realisable link from an unrealisable one only when links share a detection rate. **(a)** The per-replicate detection rate across links, $p_1 \sim \text{Beta}(\alpha, \beta)$ with mean $\bar{p}_1 = \rho(1-\epsilon_l) = 0.105$ and concentration $\nu = \alpha + \beta$. The vertical line is the common rate of Eq. S7, which large ν recovers. The four cases span a range of ecological settings: under a common p_1 (grey) every link is recorded in the same 10.5% of replicates, about one visit in ten; at $\nu = 100$ (blue) links barely differ, the middle half of them between 0.083 and 0.124 with almost none below 0.05 or above 0.20; at $\nu = 40$ (orange) they differ moderately, 11% of links being recorded in fewer than 5% of replicates and 4% in more than 20%; and at $\nu = 10$ (purple) they differ strongly, 35% falling below 0.05 and 15% above 0.20, so a quarter of links would need some 29 visits for a single expected record while another quarter would need only seven. **(b)** Posterior probability of a *possibly missing* link when every replicate is empty, for the evidence $Y = 1$, $O_l = 0$ with $\epsilon_Y = 0.2$, $\epsilon_l = 0.3$, $f = 0.05$ and $\rho = 0.15$. A common p_1 (grey line) makes a run of empty replicates increasingly hard to reconcile with a link that is there, driving the posterior from 29.5% at $R = 1$ to 14.2% by $R = 20$, whereas a drawn p_1 can also keep weight near zero, so at $\nu = 10$ it holds near 27% throughout. Detection in every replicate is not shown: every curve reaches $\kappa = 60.8\%$, a replicate or two sooner as ν falls.

S7 Correlated local and replicate errors

Section S6 let p_1 vary but kept the local factor fixed, so the local observation and the replicates still erred independently. To consider the correlation between the local observation and the replicates, we keep ρ common, and draw the detectability $u = 1 - \epsilon_l$ once per link from the beta distribution $\text{Beta}(\alpha_u, \beta_u)$, where α_u and β_u are the numbers of replicates in which a realised link was recorded or missed, respectively (u stands for the detectability, to distinguish them from the α and β of Section S6).

We note that while Section S6 draws p_1, which mixes realisation and detectability but is confined to the replicate factor, here we draw u, which is methodological alone and touches both axes (the local observation and the replicates). When the link is detected in every replicate ($n = R$), Eq. S7 has the following two modifications,

$$\epsilon_l \longrightarrow \frac{\beta_u}{\alpha_u + \beta_u + n\, z_{c,r}}, \qquad p_1^n \longrightarrow \prod_{i=0}^{n-1} \frac{\rho\,(\alpha_u + i)}{\alpha_u + \beta_u + i}, \tag{S8}$$

where i indexes the replicates within the run of detections, from $i = 0$ for the first to $i = n - 1$ for the last. It is a running count of how many detections have already been absorbed. We keep $p_0 = f$ unchanged. Both modifications follow from the same update, in which each detection is added to the count α_u. The replicates sharpen the local rate only where the link is realisable ($z_{c,r} = 1$); where it is not the detections are attributed to f and do not inform detectability. One consequence is that κ is no longer a maximum because detections are evidence that the method sees this link well (Fig. S6a).

Sharing f runs the other way because a long run of detections could mean a link often recorded in error rather

than a true link seen in every replicate (Fig. S6b). Correlation therefore raises confidence when the shared rate is the failure to see a link that is there, and lowers it when it is the tendency to record one that is not.

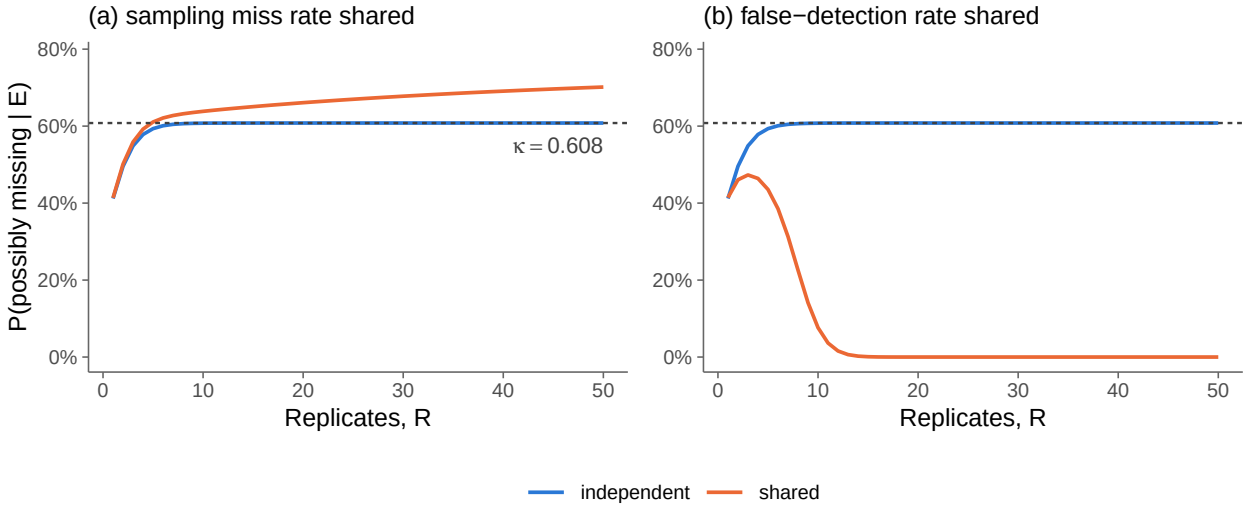

Fig. S6: Sharing the detectability raises confidence above κ, while sharing the false-detection rate reduces it. This example is for a posterior probability of *possibly missing* for $Y = 1$, $O_l = 0$ and detection in every replicate ($n = R$), with $\epsilon_Y = 0.2$, $\epsilon_l = 0.3$, $\rho = 0.15$, $f = 0.05$ and $\nu_u = 40$. Blue holds both $u = 1 - \epsilon_l$ and f fixed, so the local observation and the replicates err independently, and is the solid curve of Fig. S4a; orange draws one of them once per link, u in (a) and f in (b). The blue curve is the same in both panels, so the two differ only in which quantity is drawn. **(a)** The detectability $u = 1-\epsilon_l$, drawn with mean 0.7. The independent curve stops at $\kappa = 0.608$ (dashed), while the shared draw carries the same replicates past κ and towards $1 - \epsilon_Y$. **(b)** The false-detection rate f, drawn with mean 0.05. Blue again reaches κ, while orange rises briefly and then collapses. The probability then moves to the categories in which the link is not realisable in the replicates, primarily to *phantom*.

S8 Informed priors

Equation S1 has used $P(C) = 1/8$ throughout, placing half the prior on the link being present locally, since four of the eight categories have $z_{c,l} = 1$. Under an independence assumption, the prior takes the same product form as the likelihood, one probability for each of the three truths,

$$P(C = c) = \prod_{d \in \{Y,l,r\}} \pi_d^{z_{c,d}} \, (1 - \pi_d)^{\, 1 - z_{c,d}}, \tag{S9}$$

with π_d the prior probability that truth d is 1, and $\pi_Y = \pi_l = \pi_r = 1/2$ recovering the uniform prior.

Independence among the three truths is a strong assumption, since a link realised where we sampled is more likely realisable elsewhere. We keep it because it lets each truth carry its own prior, which is what makes departures from the uniform case tractable.

A uniform prior does not coincide with what we know about the properties of ecological networks, which have skewed degree distributions [6] and low connectance, indicating that most interactions are unlikely and that a few species have higher potential to interact [7]. Therefore, we can update our prior beliefs about a particular interacting pair using ecological knowledge (Table S7).

Predictions from species attributes have used abundances, morphological traits, phenologies, spatial and temporal distributions, and phylogenies [8]. Two conditions must be met for an interaction: individuals must meet, and once they meet their traits must match. Co-occurrence in space and time is therefore a prerequisite, and where it has not been recorded it can be borrowed from ecologically similar sites rather than taken as absent [9]. Abundance is also crucial, since the probability that two species encounter each other rises with how common each of them is, and for a rare pair it is so low that the link is either extremely hard to sample or simply does

not occur [7]. Abundance also promotes interactions between species whose traits make them unlikely partners [10]. Generalism follows from the same mechanism, because locally abundant species acquire more partners while rare ones appear specialised [11], so a species with many partners is more likely to have any particular one.

Trait matching is the second condition. A mismatch in morphology, phenology or physiology forbids the link outright, although trait values vary enough within species that database means overstate how many links are forbidden [10]. Phylogeny carries trait information indirectly, since related species tend to interact with related partners [12, 13]. It can therefore be used to inform priors as well.

The degrees of the two species can anchor π_l and π_r, at the local and the regional level respectively. Degree typically increases with abundance, so both priors do too. Trait compatibility and co-occurrence are particularly informative, because a mismatch in traits lowers both π_l and π_r, while partners that do not overlap in one place but do elsewhere lower π_l alone. Phylogeny enters π_r rather than π_l, because relatedness reflects the diversification of heritable traits that make an interaction feasible rather than local community assembly [12].

For example, suppose that our evidence is $E = (Y = 1, O_l = 0, O_r = 1)$, which under error-free conditions is the signature of *possibly missing* link. If we know that the species at hand are two specialists, we have low expectation that these species interact locally or anywhere in the region (low π_l and π_r), which will increase the confidence that the true category is *phantom* ($\hat{Y} = 1, L_l = 0, L_r = 0$). At the other extreme, if those are two generalists, our belief that they interact locally and regionally increases (high π_l and π_r), shifting the posterior probability towards *recurrent* ($\hat{Y} = 1, L_l = 1, L_r = 1$) (Fig. S7a).

We can also have prior knowledge on the site, rather than on species. If we know that the site has very different characteristics than the metaweb, then we expect the link to be realisable in the region but not realised in the focal site (π_l well below π_r), and we can be quite confident that the true category is *possibly missing*. On the other hand, if we know that the site resembles the overall characteristics of the metaweb (low between-site variation), then we expect that the interaction is also realised at the local site (π_l close to π_r), shifting our confidence towards *recurrent*, that is, reading the local zero as a miss ($L_l = 1$ although $O_l = 0$) rather than as an absence (Fig. S7b).

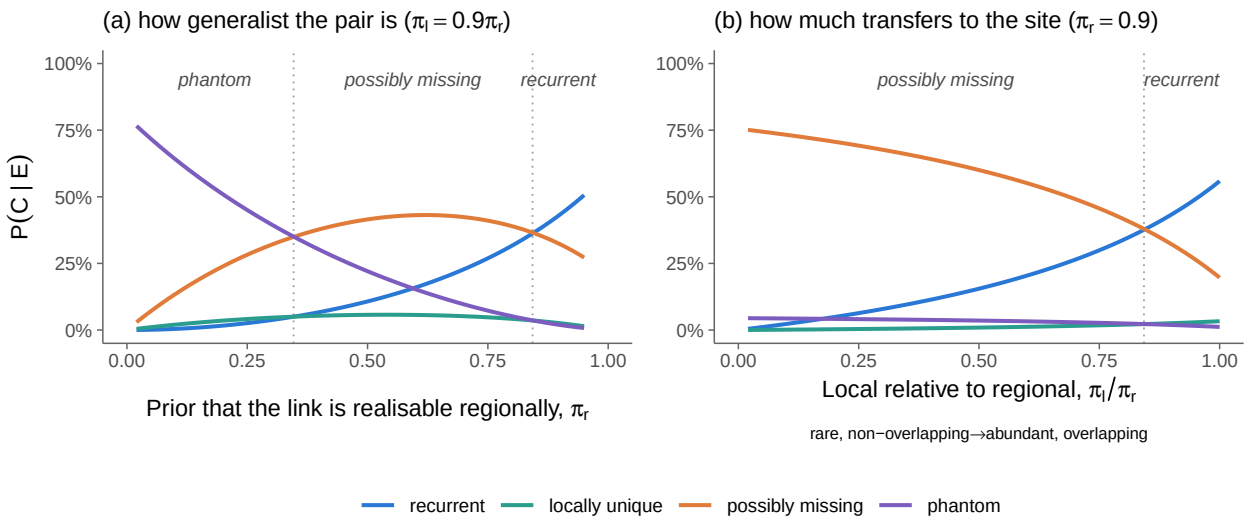

Fig. S7: An informative prior changes which source explains the same evidence. Posterior probability for the evidence used throughout, a link that is predicted ($Y = 1$), unrecorded at the focal site ($O_l = 0$) and recorded in every one of $R = 5$ replicates ($O_r = 1$). Dotted lines mark where the leading category changes, named in italics. The two panels vary different beliefs. **(a)** A belief about the species pair: how likely the link is to be realisable in the region at all, from two interacting specialists with little in common on the left to two interacting generalists on the right, with the local prior held just below the regional one. The same evidence is attributed first to a false detection elsewhere (*phantom*), then to a link that is genuinely elsewhere and was missed here (*possibly missing*), and above $\pi_r = 0.84$ to a link that is everywhere including here (*recurrent*). That third reading requires the local prior to stay close to the regional one; below about $\pi_l = 0.8\,\pi_r$ it leaves the range. **(b)** A belief about the site: the pair is fixed as strongly interacting generalists and the axis is how much of that regional expectation transfers locally. Low values are a site where the two species are rare or their phenologies do not overlap, so failing to record the link is expected and the reading stays *possibly missing* at 74%; high values are a site where both are abundant and overlapping, so failing to record it becomes surprising and *recurrent* takes over. The mapping from abundance and degree to π_l and π_r is a common assumption in ecology. Parameter values are $\epsilon_Y = 0.2$, $\epsilon_l = 0.3$, $f = 0.05$, $\rho = 0.15$ and $\pi_Y = 1/2$. Only the four categories in which the model is right are drawn; they hold exactly $1 - \epsilon_Y = 80\%$ of the posterior throughout and the other four are their mirrors.

Several of these factors also set the error rates, so the prior and the likelihood are not independent sources (Table S7). The prior increases with abundance, co-occurrence and generalism, and so does the ease of recording the link, through a lower ϵ_l or a higher ρ. For a common, widely co-occurring and generalist pair the same ecology is therefore counted twice, and the posterior is more confident than it should be. Relatedness allows phylogeny to inform the prior, since closely related species tend to share partners [12]. However, the same closeness raises f, because taxa that are hard to tell apart can be recorded as the same one [8]. Within a clade whose members are easily confused, the prior for a link therefore rises while the record that would confirm it becomes less reliable.

The model prior π_Y is where a model's output score, instead of a threshold, can be informative. If the model returns a calibrated probability q rather than a bit, which is a standard calibration step in machine learning (e.g., Platt scaling [14]), then q is prior information about the link rather than a further observation of it, and we set $\pi_Y = q$. The two treatments are exclusive because Y is q thresholded, so they are one output of one model. Either the model enters as evidence, through Y and ϵ_Y with $\pi_Y = 1/2$, or as prior, through $\pi_Y = q$ and no model factor in the likelihood. Using both would count the model twice.

In the latter case the four categories whose model bit is 1 carry q and the four whose bit is 0 carry $1 - q$, so the balance between them tips at $q = 0.5$ instead of at a chosen threshold. Setting $\pi_Y = q$ is an approximation, since q is calibrated against the observations while π_Y is a belief about $\hat{Y}$, and the distance between the two is what ϵ_Y measures. A thresholded Y discards the score at the first step, whereas $\pi_Y = q$ carries it through to the posterior. Thresholding is in fact equivalent to setting $q = 1 - \epsilon_Y$ for every positive, so a link the model

rates at 0.55 and one it rates at 0.95 both enter as though the model were 80% confident, whereas carrying q through separates them in proportion to their scores. Carrying the score through therefore replaces a step at the threshold with a ramp in q, and the two agree only where the model's confidence happens to equal $1 - \epsilon_Y$ (Fig. S8).

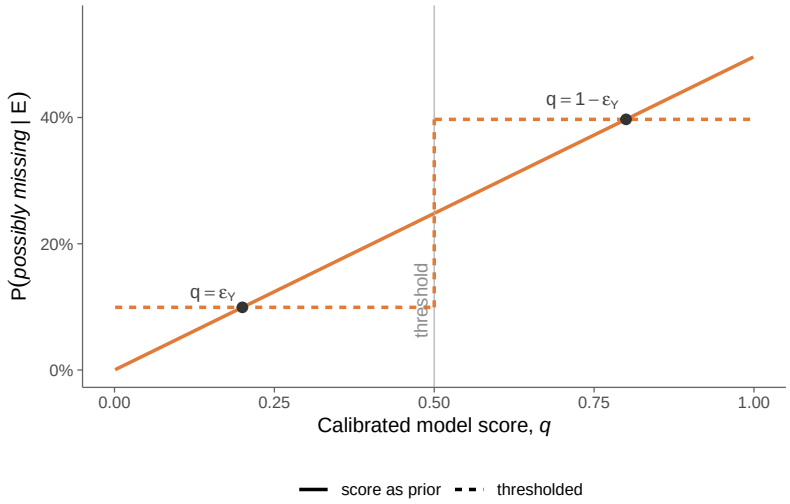

Fig. S8: Using model output probabilities provides a smoother posterior than using a threshold. Posterior probability of *possibly missing* for the evidence $O_l = 0$ and detection in every one of $R = 5$ replicates, as the calibrated model score q varies, with $\epsilon_Y = 0.2$, $\epsilon_l = 0.3$, $f = 0.05$ and $\rho = 0.15$. The solid line sets $\pi_Y = q$ and drops the model factor from the likelihood; the posterior is then exactly proportional to q, because the four categories in which the model is right and the four in which it is wrong differ only in that bit, so the normalising constant does not depend on q. The dashed line thresholds the same score at 0.5 and enters Y as evidence with rate ϵ_Y, which collapses every score on one side of the threshold onto a single value: 9.9% below and 39.7% above, so a score of 0.51 and a score of 0.99 are read identically. The two treatments agree only at $q = \epsilon_Y$ and $q = 1 - \epsilon_Y$ (points), and between them the threshold is over-confident about weak positives and under-confident about strong ones. *Locally absent*, the same category with the model bit flipped, is the mirror image of both curves.

Table S7: Ecological and methodological factors affecting the likelihood and prior. Arrows give the direction in which the quantity moves as the factor increases. The factors in the middle block inform both the likelihood and the prior, so Eq. S1, which multiplies prior by likelihood as though they were independent sources, counts the same ecology twice: for abundance and co-occurrence the two pull the same way and the posterior is overconfident, while for relatedness the prior rises and the record that would confirm it becomes less reliable. Relatedness here is between a species and others in its own guild, not between the two interacting species.

Factor	Likelihood	Prior
Evidence only		
Sampling effort	$\epsilon_l \downarrow$	–
Number of replicates	replicate factor	–
Taxonomic resolution	$\epsilon_l \downarrow$, $f \downarrow$	–
Model quality	$\epsilon_Y \downarrow$	–
Both		
Abundance	$\epsilon_l \downarrow$, $\rho \uparrow$	$\pi_l \uparrow$, $\pi_r \uparrow$
Co-occurrence, phenology	$\rho \uparrow$	$\pi_l \uparrow$, $\pi_r \uparrow$
Phylogenetic relatedness	$f \uparrow$	$\pi_r \uparrow$
Prior only		
Generalism (degree)	–	$\pi_l \uparrow$, $\pi_r \uparrow$
Trait matching	–	$\pi_l \uparrow$, $\pi_r \uparrow$
Model score q	–	$\pi_Y = q$

S9 Measuring the errors

The framework takes the four quantities of Section S4 ($\epsilon_Y, \epsilon_l, f, \rho$) as given. The model side is the only one with an established procedure for measurement, but it measures δ rather than ϵ_Y. Ecological network research has not yet settled how to obtain any of these quantities on a common footing. **However, our framework does not introduce this difficulty, but rather explicitly exposes it**: treating a recorded network as the interaction network is itself a choice of the four: $\epsilon_l = f = 0$ and $\rho = 1$, with δ standing in for ϵ_Y. Exposing the error rates make them addressable: each error rate can (and should) be bounded, varied or measured. We therefore provide this short discussion on measuring error rates.

The miss rate ϵ_l has been mainly approached indirectly, through sampling completeness and through null models that derive it from effort [2, 15]. One direct estimation was done through hierarchical observation models that estimate species-level detection probability from repeated sampling occasions [16]. The false-detection rate f is usually assumed rather than measured [2]; and the realisation rate ρ is rarely separated from detection at all. What follows is therefore an account of what each quantity would take to measure, and of what can be done when it is unavailable, rather than a settled protocol.

The model error, ϵ_Y. What cross-validation returns is δ, the error of Y against the observations it was fitted to [17], and its two directions can be read separately from the confusion matrix (Box 1 in the main text). That is not ϵ_Y. Had the observations been error-free, the fitted model would be the truth-fitted model and ϵ_Y would be zero, so what makes ϵ_Y non-zero is the same observation error the framework carries in ϵ_l and f, which no held-out set can reveal because the withheld labels are observations too. In practice, δ is the stand-in, and how it is obtained matters: a random hold-out gives the error on pairs of the same kind, whereas blocking by site or by clade, and keeping the withheld data disjoint from the evidence used to categorise, gives the error expected when the model is carried to a new context [3], which is closer to what the framework needs. Where several prediction methods are applied to the same candidate link, the margin of their vote is not a rate but a score, so it belongs in π_Y (Section S8) rather than here.

The miss rate ϵ_l. This is the easiest to estimate of the four because it is possible to use common detection estimates. These are mainly done via repeated sampling, over a period short enough that the local network can be treated as unchanged, holding realisation fixed so that what varies between repeats is detection alone, which returns $1 - \epsilon_l$ directly [16, 18]. The estimate is only as good as that constancy, and is badly biased when realisation or detection do vary between repeats [19]. Where such repeats do not exist, the rate can be derived from effort alone. In a neutral null model, where species are observed in proportion to their abundance, the probability that a realised link goes unrecorded follows from the number of observations, and the resulting rate falls with effort and rises with species richness [2]. Estimators of sampling completeness give a coarser version of the same quantity [15], though they attribute every unrecorded pair to sampling, including pairs that are forbidden and would not be recorded at any effort [7]. A third route works from the recorded network itself, since methods that infer which absences are likely to be missing links return a score or a probability for each unrecorded pair [20–22]. These could give an expected number of unrecorded links, whose ratio to the number recorded estimates ϵ_l. This would be an upper bound, since none of these methods separates a link missed by sampling from one that does not occur. These methods return a network-level average, while ϵ_l is in fact highest for the rarest pairs. We deal with this in Section S6.

The false-detection rate f. This rate is set a priori as a misidentification probability rather than estimated [2], but can be estimated via two routes. The first is re-examination via another method and calculating the proportion of detection overturns (e.g., obtaining molecular confirmation for a sample of recorded interactions obtained visually). The second uses impossible pairs as negative controls: species whose ranges or phenologies do not overlap cannot interact, so a recorded interaction between them is a false detection. However, this route is only as good as the researcher's knowledge of which pairs are impossible. That knowledge, when it rests on trait thresholds, can be labile such that pairs deemed forbidden from mean traits are recorded interacting [10].

Any wrongly forbidden pair inflates the estimate of f, so the criterion is better restricted to pairs that never co-occur, whose impossibility does not depend on a trait threshold at all.

The realisation rate ρ. This is the hardest to estimate, because it is not a property of a method. The most direct estimate comes from the replicates themselves: where the same pair is sampled across many sites or times, the proportion of networks in which the interaction is recorded estimates how often it is realised [1]. That proportion is, however, $p_1 = \rho(1-\epsilon_l)$ rather than ρ, so it is a lower bound. For example, a link realised half the time and detected always is indistinguishable from one realised always and detected half the time. Separating them requires the repeated-visit design above, which holds realisation fixed so that what varies between repeats is detection alone. Camera observation of hummingbird-plant visitation analysed this way returned per-day detection probabilities of 16% to 44% across hummingbird species [16]. Uncertainty in ρ is also not of the same type as in the other three: improving a method lowers ϵ_Y, ϵ_l and f, whereas ρ is a property of the system, so more replicates sharpen its estimate without reducing the variability it describes [1].

References

1. Banville, F., Strydom, T., Blyth, P. S. A., Brimacombe, C., Catchen, M. D., Dansereau, G., Higino, G., Malpas, T., Mayall, H., Norman, K., Gravel, D. & Poisot, T. Deciphering probabilistic species interaction networks. en. *Ecology Letters* **28,** e70161. doi:10.1111/ele.70161 (2025).
2. Catchen, M. D., Poisot, T., Pollock, L. J. & Gonzalez, A. The missing link: Discerning true from false negatives when sampling species Interaction networks. *EcoEvoRxiv*. doi:10.32942/x2dw22 (2023).
3. Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F. & Dormann, C. F. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. en. *Ecography* **40,** 913–929. doi:10.1111/ecog.02881 (2017).
4. Chambert, T., Miller, D. A. W. & Nichols, J. D. Modeling false positive detections in species occurrence data under different study designs. en. *Ecology* **96,** 332–339. doi:10.1890/14-1507.1 (2015).
5. Miller, D. A., Nichols, J. D., Mcclintock, B. T., Grant, E. H. C., Bailey, L. L. L. & Weir, L. A. Improving occupancy estimation when two types of observational error occur: non-detection and species misidentification. en. *Ecology* **92,** 1422–1428. doi:10.1890/10-1396.1 (2011).
6. Jordano, P., Bascompte, J. & Olesen, J. M. Invariant properties in coevolutionary networks of plant–animal interactions: Invariant properties in coevolutionary networks. en. *Ecology Letters* **6,** 69–81. doi:10.1046/j.1461-0248.2003.00403.x (2003).
7. Jordano, P. Sampling networks of ecological interactions. en. *Functional ecology* **30,** 1883–1893. doi:10.1111/1365-2435.12763 (2016).
8. Peralta, G., CaraDonna, P. J., Rakosy, D., Fründ, J., Pascual Tudanca, M. P., Dormann, C. F., Burkle, L. A., Kaiser-Bunbury, C. N., Knight, T. M., Resasco, J., Winfree, R., Blüthgen, N., Castillo, W. J. & Vázquez, D. P. Predicting plant-pollinator interactions: concepts, methods, and challenges. en. *Trends in ecology & evolution* **39,** 494–505. doi:10.1016/j.tree.2023.12.005 (2024).
9. Kampe, J. N., DeSisto, C. M. M. & Dunson, D. B. Link prediction in ecological meta-networks under extreme taxonomic bias. en. *Methods in Ecology and Evolution.* doi:10.1111/2041-210x.70368 (2026).
10. González-Varo, J. P. & Traveset, A. The Labile Limits of Forbidden Interactions. *Trends in Ecology & Evolution. Multiplicity in Unity: Plant Subindividual Variation and Interactions with AnimalsHandbook of the Birds of the World (Vol. 2)The Complete Birds of the Western Paleartic (CD-ROM)Genetics and Analysis of Quantitative Traits* **31,** 700–710. doi:10.1016/j.tree.2016.06.009 (2016).
11. Poisot, T., Stouffer, D. B. & Gravel, D. Beyond species: why ecological interaction networks vary through space and time. en. *Oikos* **124,** 243–251. doi:10.1111/oik.01719 (2015).

12. Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M. J., Fortin, M.-J., Hemming, V., Mercier, B., Pollock, L. J., Runghen, R., Dalla Riva, G. V. & Poisot, T. Food web reconstruction through phylogenetic transfer of low-rank network representation. en. *Methods in ecology and evolution* **13,** 2838–2849. doi:10.1111/2041-210x.13835 (2022).

13. Nunes Martinez, A. & Mistretta Pires, M. Estimated missing interactions change the structure and alter species roles in one of the world's largest seed-dispersal networks. en. *Oikos (Copenhagen, Denmark).* doi:10.1111/oik.10521 (2024).

14. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* **10,** 61–74 (1999).

15. Chacoff, N. P., Vázquez, D. P., Lomáscolo, S. B., Stevani, E. L., Dorado, J. & Padrón, B. Evaluating sampling completeness in a desert plant–pollinator network. *Journal of Animal Ecology* **81,** 190–200. doi:10.1111/j.1365-2656.2011.01883.x (2012).

16. Weinstein, B. G. & Graham, C. H. Persistent bill and corolla matching despite shifting temporal resources in tropical hummingbird-plant interactions. en. *Ecology Letters* **20,** 326–335. doi:10.1111/ele.12730 (2017).

17. Yates, L. A., Aandahl, Z., Richards, S. A. & Brook, B. W. Cross validation for model selection: A review with examples from ecology. en. *Ecological monographs* **93.** doi:10.1002/ecm.1557 (2023).

18. Royle, J. A. N-mixture models for estimating population size from spatially replicated counts. en. *Biometrics* **60,** 108–115. doi:10.1111/j.0006-341X.2004.00142.x (2004).

19. Link, W. A., Schofield, M. R., Barker, R. J. & Sauer, J. R. On the robustness of N-mixture models. en. *Ecology* **99,** 1547–1551. doi:10.1002/ecy.2362 (2018).

20. Stock, M., Poisot, T., Waegeman, W. & De Baets, B. Linear filtering reveals false negatives in species interaction data. en. *Scientific reports* **7,** 45908. doi:10.1038/srep45908 (2017).

21. Guimerà, R. & Sales-Pardo, M. Missing and spurious interactions and the reconstruction of complex networks. en. *Proceedings of the National Academy of Sciences of the United States of America* **106,** 22073–22078. doi:10.1073/pnas.0908366106 (2009).

22. Terry, J. C. D. & Lewis, O. T. Finding missing links in interaction networks. en. *Ecology* **101,** e03047. doi:10.1002/ecy.3047 (2020).