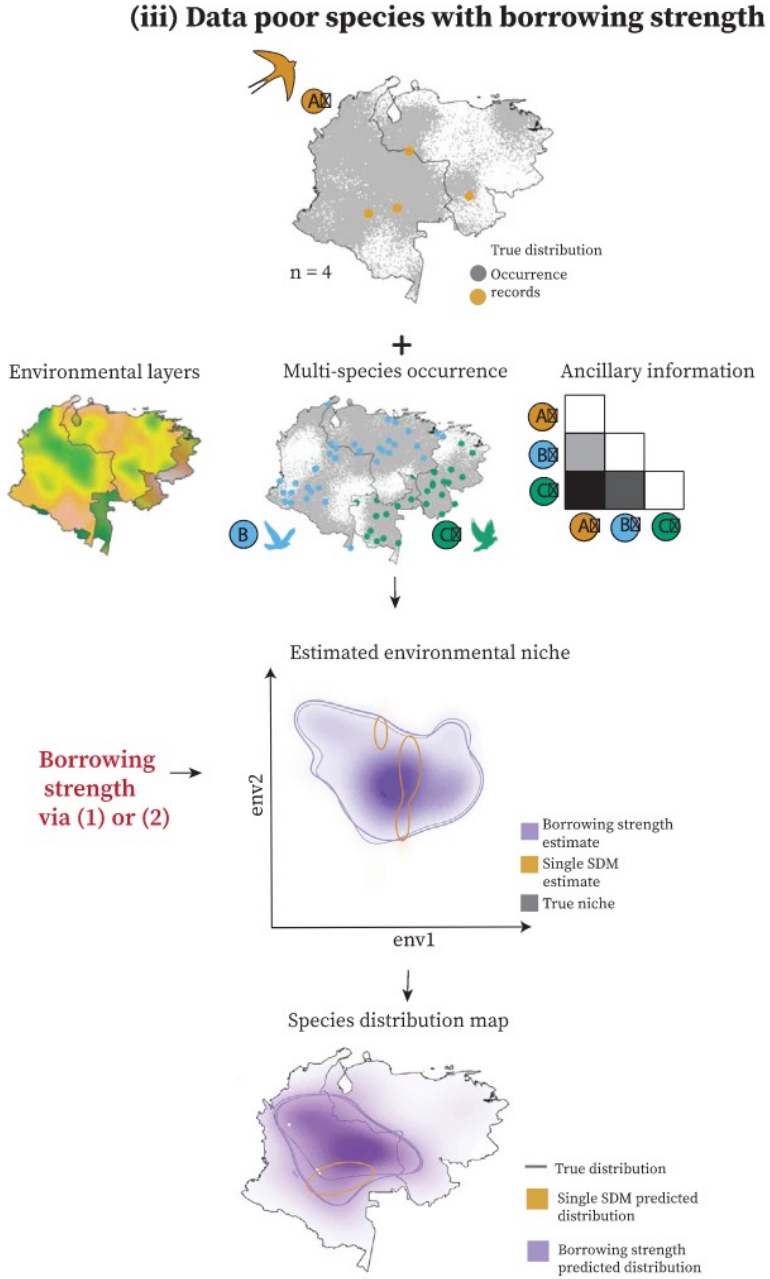

Shubhi Sharma, Kevin Winner, Laura J. Pollock, James T. Thorson, Jussi Mäkinen, Cory Merow, Eric J. Pedersen, Kalkidan F. Chefira, Julia M. Portmann, Fabiola Iannarilli, Sara Beery, Riccardo De Lutio, Walter Jetz, 2025. No species left behind: borrowing strength to map data-deficient species. Trends in Ecology & Evolution 40, 699–711. https://doi.org/10.1016/j.tree.2025.04.010

Class focus area: Phylogenies to help with data deficiency

EEB603: Brian O'Meara

All quotes and images from the above paper unless otherwise noted

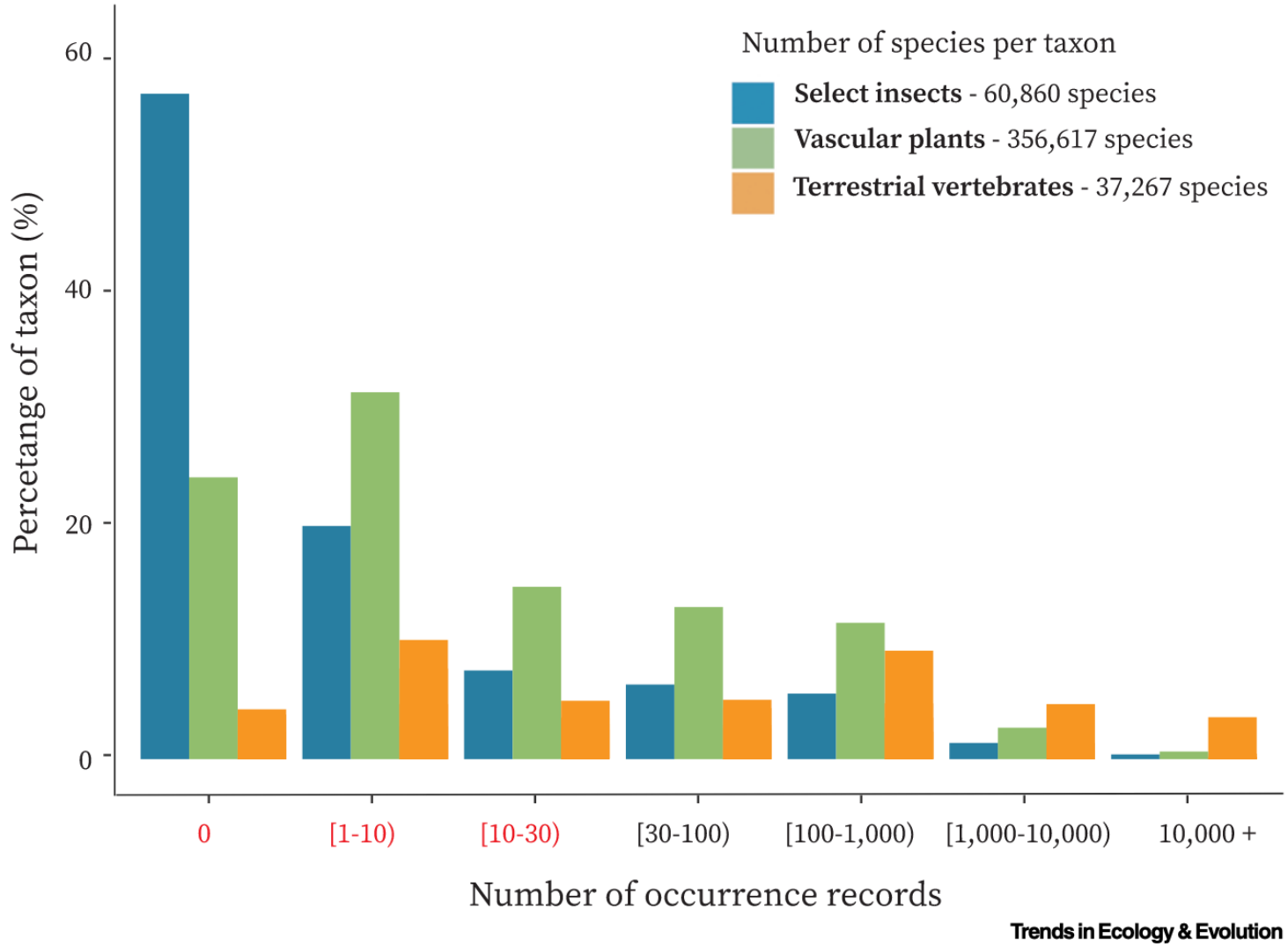

Figure 1. Data deficiency across taxonomic groups. Gaps in species occurrence data. Bars show the number of occurrence records per species for all extant members of select large taxa. Species groups covered include (i) terrestrial vertebrates: nonmarine mammals, birds, reptiles, and amphibians; (ii) select insect taxa: dragonflies, ants, butterflies, and bees; and (iii) vascular plants. Data were accessed from the Global Biodiversity Information Facility (GBIF) (accessed January 2024, https://doi.org/10.15468/dl.re4tqy), harmonized in Map of Life (mol.org) against a master list of 454 744 species in a total of which 13% (60 860) are insect species, 78% (356 617) are vascular plant species, and 8% (37 267) are terrestrial vertebrates (details are given in Table S1 in the supplemental information online). Species under 30 occurrence records are considered 'data-deficient' according to traditional statistical theory (highlighted in red), although this number can vary depending on the type of statistical model.

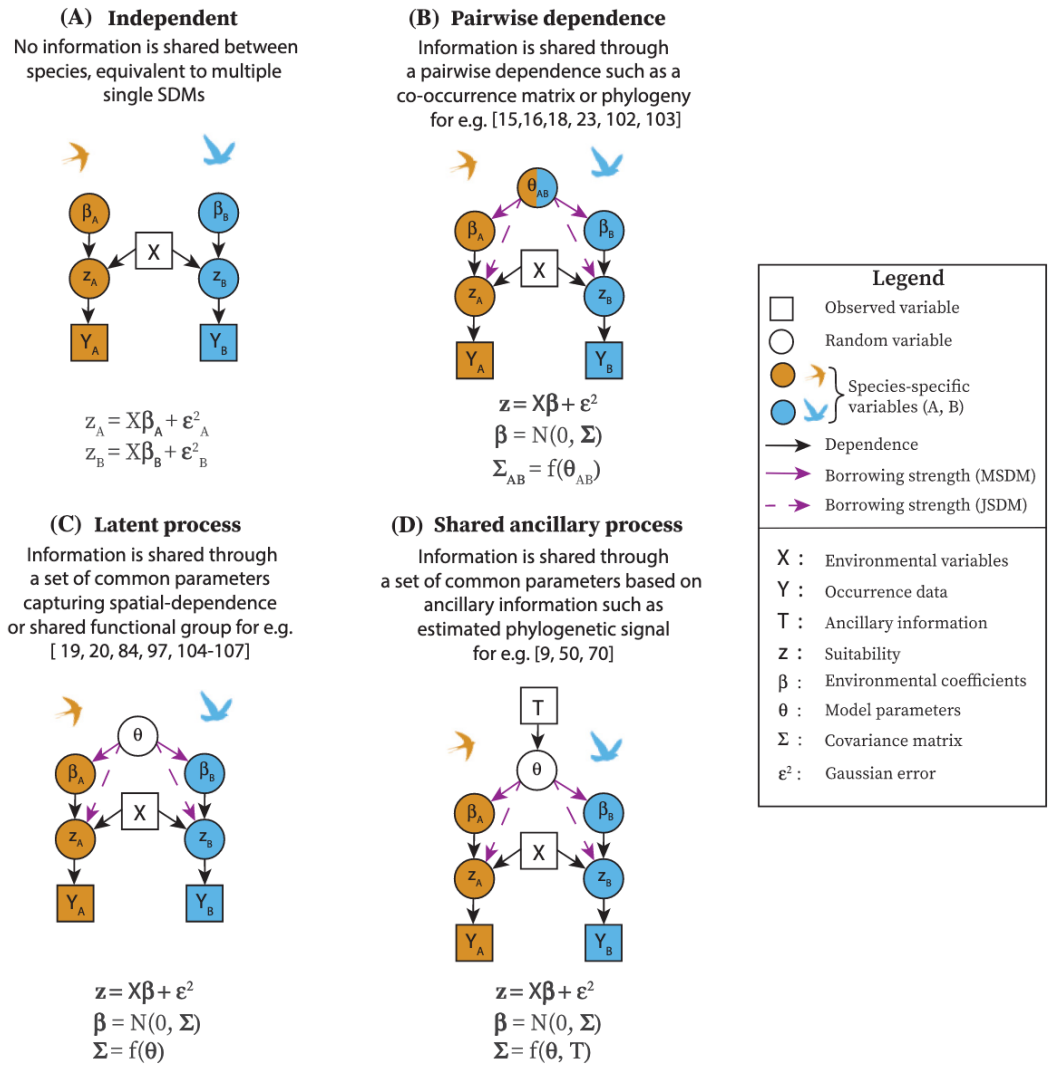

Figure 2. Classification of models that borrow strength. Principal avenues for information sharing in species distribution models (SDMs). The graphs illustrate information sharing between two species, based on their respective occurrence data (Y), environmental covariates (X), and the resulting SDM parameters (β) of the environmental niche. The Z variables then represent the species' suitability or else the true occupancy of sites, which is unknown. In (A) the independent (null) case, species A and B are conditionally independent of one another given the environmental covariates. In model (B), the species' suitability or occurrence are related directly to one another, perhaps through a model of co-occurrence [15,16,18,23,102,103]. In model (C), the dependence between species is estimated by a set of shared parameters that measure an unstructured latent process [19,20,84,97,104–107]. (D) This illustrates two mechanisms for bringing in ancillary data (T), such as species traits, expert knowledge, and/or phylogenetic information [16,50,70]. The key difference between (C) and (D) is the addition of the ancillary information.

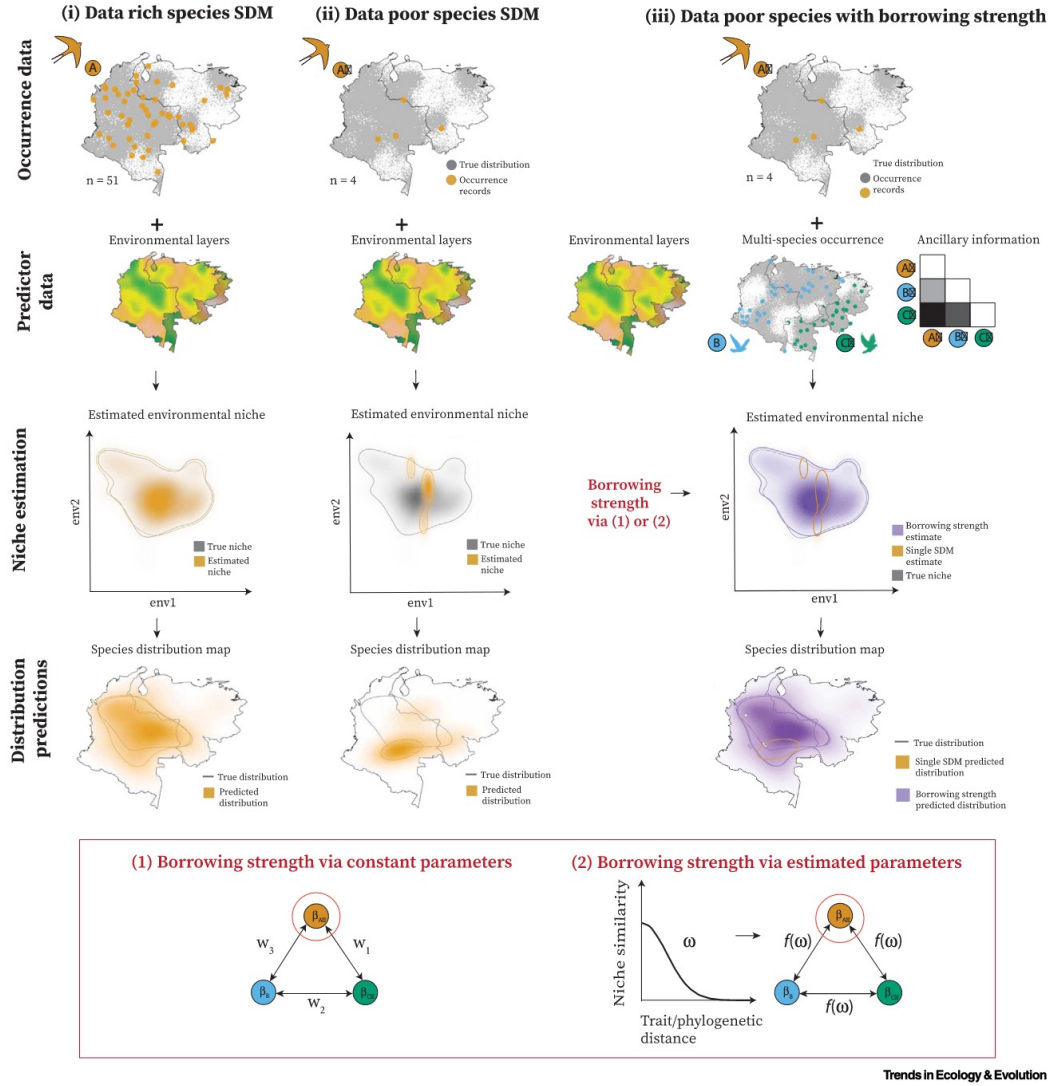

Figure 3. Illustrated example of borrowing strength. Borrowing strength from data-rich to data-deficient species through models of the shared ancillary process. Consider a data-rich (i) and data-deficient (ii) scenario for a simulated species. The data-rich scenario follows a typical species distribution modeling workflow where occurrence data are combined with environmental data to produce estimates of the species' environment niche, which in turn is used to predict suitable habitat across a landscape. When the species is data-deficient, however, the estimate of the environmental niche is often biased or incomplete, leading to poor spatial predictions. In scenario (iii), the distribution model for species A is additionally informed by ancillary information such as from traits or phylogeny. Ancillary information is incorporated as a distance matrix based on, for example, the similarity of traits or phylogenetic relatedness. This information is combined with the niche estimates for related species in one of two ways. (i) Constant parameters that are not estimated in the model. Here there is an *a priori* assumed similarity between species that are used as weights to share information across species, for example, an assumption that similarity is proportional to phylogenetic distance. (ii) Model estimated parameters – ω parameterizes the relationship between niche similarity among the three species as a function of a trait or phylogenetic distance. Information is shared between species-level parameters according to this structure resulting in a borrowing strength estimate (purple) that captures the true niche and, therefore, the predicted geographic distribution more accurately. We compare the niche estimates from a single-species distribution model (SDM) (using occurrence data only) with the borrowing strength estimate (using ancillary data, occurrence, and environmental data for species A, B, and C) to the true distribution (centroid drawn in gray). The proposed model captures the shared ancillary process by estimating the relationship between niche correlation (or similarity) and phylogenetic distance by the parameters in ω. When borrowing strength between parameters in the model, the amount of information shared is weighted by ω – for example, estimates for species A are pulled 'more' towards species B than species C estimates.

“The existing methods utilize the phylogenetic distance matrix to model the covariance between species-level environment coefficients, implicitly assuming that similarity in species–environment relationships is proportional to phylogenetic distance (Figure 3). This is a pattern expected only under the neutral expectation of niche evolution, described by a Brownian motion (BM) model, whereas there is mounting evidence that other patterns, such as niche conservatism (more similar than expected under BM) are common [59,68,69]”

“The existing methods utilize the phylogenetic distance matrix to model the covariance between species-level environment coefficients, implicitly assuming that similarity in species–environment relationships is proportional to phylogenetic distance (Figure 3). This is a pattern expected only under the neutral expectation of niche evolution, described by a Brownian motion (BM) model, whereas there is mounting evidence that other patterns, such as niche conservatism (more similar than expected under BM) are common [59,68,69]”

False

True

"The existing methods utilize the phylogenetic distance matrix to model the covariance between species-level environment coefficients, implicitly assuming that similarity in species–environment relationships is proportional to phylogenetic distance (Figure 3). This is a pattern expected only under the neutral expectation of niche evolution, described by a Brownian motion (BM) model, whereas there is mounting evidence that other patterns, such as niche conservatism (more similar than expected under BM) are common [59,68,69]"

False

True

TABLE 1. A summary of some phenotypic covariance structures expected under microevolutionary models discussed in the text. Variables are defined in the main text, and some of the formulas are simplified for clarity.

	Model	$\mathrm{Cov}[\mathbf{X}_i, \mathbf{X}_j] \approx$	Correlation between species phenotypes
	Random genetic drift:		
Brownian →	alone	$(\mathbf{G}/N_e)t_z$	Decreases linearly with increasing phylogenetic distance
Brownian →	drift-mutation balance	$2\mathbf{G}_m t_z$	(but may take other forms)
	Directional selection:		
Brownian →	with random genetic drift	$(\mathbf{G}/N_e)t_z$	Decreases linearly with increasing phylogenetic distance.
Brownian →	fluctuating environment	$\mathbf{G}\mathbf{V}_s\mathbf{G}'t_z$	Trend cannot be detected from comparative data alone.
	Stabilizing selection and drift:		
	univariate	$\hat{\mathbf{V}}\mathrm{Exp}[-w\mathbf{V}_A t_{ij}]$	Decreases exponentially with increasing phylogenetic distance.
	multivariate	$\mathbf{Q}(t_{iz})\hat{\mathbf{V}}\mathbf{Q}'(t_{jz})$	Decreases with increasing phylogenetic distance as a sum of exponentials.
	Stabilizing selection + environmental change:		
Brownian →	Brownian motion of optimum	$\mathbf{E}_\theta t_z$(+ exponentials)	Decreases linearly with increasing phylogenetic distance.
	Lynch-Lande model	$\mathbf{Q}(t_{iz})\mathbf{E}_\theta\mathbf{Q}'(t_{jz})$	Decreases with increasing phylogenetic distance as a sum of exponentials.
Brownian →	Punctuated phenotypic change	$\left(\mathbf{H} + \frac{\sigma^2}{\mu}\mathbf{h}\mathbf{h}'\right)\frac{t_z}{\mu}$	Decreases linearly with increasing phylogenetic distance.
	Change correlated with speciation events	$\mathbf{H}\mathbf{N}_z$	Decreases with ratio of shared to total speciation events.

Hansen and Martins 1996 *Evolution* https://doi.org/10.1111/j.1558-5646.1996.tb03914.x

1. What are SDMs for?
2. Independent vs Multispecies vs Joint-species models?
3. Should we decide on the correlation (constant parameters) or let the data drive it (estimated parameters)?
4. What is machine learning? Is it like ChatGPT?
5. In borrowing strength, are we also borrowing weakness?
6. What about issues from allopatric speciation? Competition?
7. What is the point where you say, “there aren’t enough data to answer this question”?
8. “Wallacean shortfall” [also see “Linnaean shortfall”]
9. What ancillary data should we use?
10. Should we do this beyond SDMs – what other traits are missing data that we care about?