Received: 14 June 2022 | Accepted: 30 January 2023

DOI: 10.1111/2041-210X.14076

Methods in Ecology and Evolution

RESEARCH ARTICLE

Identifying direct and indirect associations among traits by merging phylogenetic comparative methods and structural equation models

James T. Thorson[1] | Aurore A. Maureaud[2,3,4] | Romain Frelat[5] | Bastien Mérigot[6] | Jennifer S. Bigman[7] | Sarah T. Friedman[8,9] | Maria Lourdes D. Palomares[10] | Malin L. Pinsky[4] | Samantha A. Price[11] | Peter Wainwright[8]

[1]Habitat and Ecological Processes Research, Alaska Fisheries Science Center, Seattle, Washington, USA; [2]Department of Ecology & Evolutionary Biology, Yale University, New Haven, Connecticut, USA; [3]Center for Biodiversity & Global Change, Yale University, New Haven, Connecticut, USA; [4]Department of Ecology, Evolution, and Natural Resources, Rutgers University, New Brunswick, New Jersey, USA; [5]Aquaculture and Fisheries Group, Wageningen University & Research (WUR), Wageningen, The Netherlands; [6]MARBEC, Université de Montpellier, CNRS, IFREMER, IRD, Sète, France; [7]Recruitment Processes Program, Alaska Fisheries Science Center, NOAA Fisheries, Seattle, Washington, USA; [8]Department of Evolution and Ecology, University of California Davis, Davis, California, USA; [9]Current address: Groundfish Assessment Program, Alaska Fisheries Science Center, Seattle, Washington, USA; [10]Sea Around Us, Institute for the Oceans and Fisheries, University of British Columbia, Vancouver, British Columbia, Canada and [11]Department of Biological Sciences, Clemson University, Clemson, South Carolina, USA

Correspondence
James Thorson
Email: james.thorson@noaa.gov

Funding information
National Science Foundation, Grant/Award Number: 2109411 and DEB-1556953

Handling Editor: Arthur Porto

Abstract

1. Traits underlie organismal responses to their environment and are essential to predict community responses to environmental conditions under global change. Species differ in life-history traits, morphometrics, diet type, reproductive characteristics and habitat utilization.
2. Trait associations are widely analysed using phylogenetic comparative methods (PCM) to account for correlations among related species. Similarly, traits are measured for some but not all species, and missing continuous traits (e.g. growth rate) can be imputed using 'phylogenetic trait imputation' (PTI), based on evolutionary relatedness and trait covariance. However, PTI has not been available for categorical traits, and estimating covariance among traits without ecological constraints risks inferring implausible evolutionary mechanisms.
3. Here, we extend previous PCM and PTI methods by (1) specifying covariance among traits as a structural equation model (SEM), and (2) incorporating associations among both continuous and categorical traits. Fitting a SEM replaces the covariance among traits with a set of linear path coefficients specifying potential evolutionary mechanisms. Estimated parameters then represent regression slopes (i.e. the average change in trait Y given an exogenous change in trait X) that can be used to calculate both direct effects (X impacts Y) and indirect effects (X impacts Z and Z impacts Y).

4. We demonstrate phylogenetic structural-equation mixed-trait imputation using 33 variables representing life history, reproductive, morphological, and behavioural traits for all >32,000 described fishes worldwide. SEM coefficients suggest that one degree Celsius increase in habitat is associated with an average 3.5% increase in natural mortality (including a 1.4% indirect impact that acts via temperature effects on the growth coefficient), and an average 3.0% decrease in fecundity (via indirect impacts on maximum age and length). Cross-validation indicates that the model explains 54%–89% of variance for withheld measurements of continuous traits and has an area under the receiver-operator-characteristics curve of 0.86–0.99 for categorical traits.
5. We use imputed traits to classify all fishes into life-history types, and confirm a phylogenetic signal in three dominant life-history strategies in fishes. PTI using phylogenetic SEMs ensures that estimated parameters are interpretable as regression slopes, such that the inferred evolutionary relationships can be compared with long-term evolutionary and rearing experiments.

KEY-WORDS
evolutionary mechanisms, life history strategies, phylogenetic trait imputation, population and community ecology, structural equation model, trait-based approach, phylogenetic comparative methods

1 | INTRODUCTION

Trait-based approaches are essential for improving our understanding of ecological and evolutionary processes. For example, they are used to identify population and community responses to global change (Pacifici et al., 2017), community assembly rules (Gross et al., 2021; Legras et al., 2019), and predict how changes in community diversity affect ecosystem functioning (Díaz et al., 2013) and ecosystem services (Hevia et al., 2017). They can also be used to test theory regarding evolutionary mechanisms (Baker et al., 2020) and support biodiversity conservation (Cardillo et al., 2008). Traits of floristic and faunal species can be quantitative (discrete or continuous) and/or qualitative (binary, nominal, or ordinal variables). For instance, continuous traits include growth rates, body or leaf size, and age at maturity, while categorical traits encompass behaviours (e.g. solitary or gregarious species), diet (autotroph, heterotroph, mixotroph) or reproduction (dispersal modes, guarding vs. nonguarding young) (Hadj-Hammou et al., 2021; Violle et al., 2007).

Trait values are not available for every species of interest, both due to limited scientific resources and ongoing difficulties in collecting and/or sharing trait information across taxa and systems (although see Gallagher et al., 2020). Consequently, there are many potential methods available to impute these missing trait values (Azur et al., 2011; Goolsby et al., 2017; Schrodt et al., 2015). Comparisons of phylogenetic trait imputation (PTI) methods generally show that performance is improved by including phylogenetic information (Debastiani et al., 2021; Penone et al., 2014; Taugourdeau et al., 2014), or even using taxonomy as a proxy for phylogeny (Johnson et al., 2021) wherein related taxa are more likely to share similar traits than unrelated taxa.

PTI generally involves specifying a statistical process for how trait values change along a phylogenetic (Goolsby et al., 2017) or taxonomic tree (Schrodt et al., 2015; Thorson, 2020; Thorson et al., 2017). This involves estimating parameters to represent correlations $\mathbf{R}$ among n_g taxa for a given trait, as well as covariance $\boldsymbol{\Sigma}$ among n_j traits. For example, the function phylopars in R-package RPHYLOPARS is a common implementation for PTI but it cannot be implemented for categorical traits (Johnson et al., 2021; Penone et al., 2014), while such traits are generally easier to assess and collect than continuous ones. Additionally, estimating $\boldsymbol{\Sigma}$ without constraints (beyond the requirement that it is symmetric and positive definite) has three main limitations (Grace, 2006): (1) results cannot be compared easily with slopes estimated in conventional regression models, such that results are difficult to interpret or validate using experimental data; (2) existing methods cannot use evolutionary theory and experiments to specify the structure of covariance among traits; and (3) the number of parameters in $\boldsymbol{\Sigma}$ without other constraints is $n_j(n_j+1)/2$, which becomes computationally challenging to fit when interpolating a large number of traits.

As an alternative to estimating the covariance among traits directly, we propose to use structural equation modelling (SEM) to specify a parsimonious structure for this trait-covariance $\boldsymbol{\Sigma}$. Given a set of n_j traits $\left\{Y_1, Y_2, \ldots, Y_{n_j}\right\}$ with measurements $\left\{y_1, y_2, \ldots, y_{n_j}\right\}$, SEM allows the user to specify a set of dependencies linking these, where each dependency is represented by a path coefficient. These links can be interpreted as a graph wherein

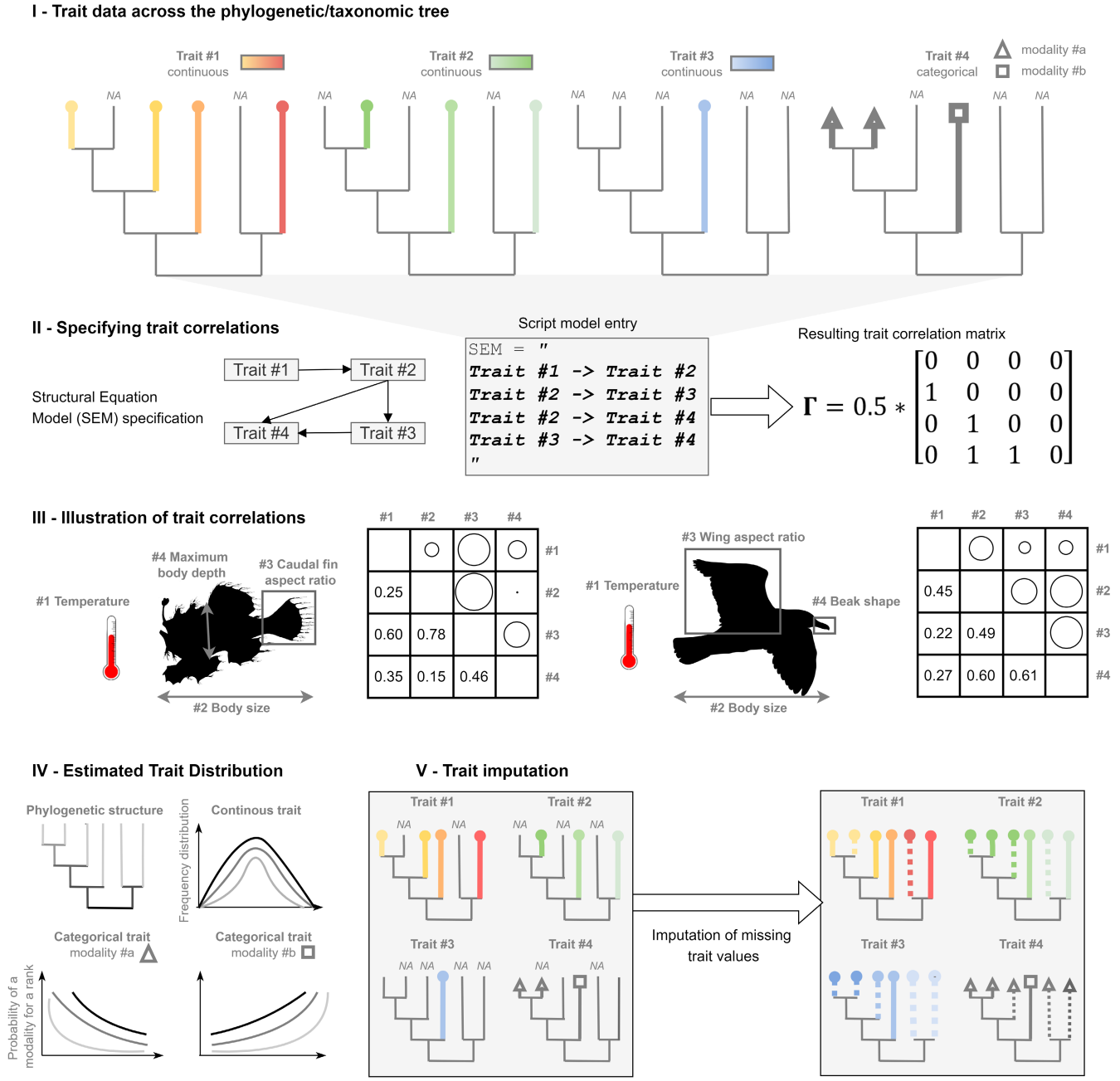

FIGURE 1 Conceptual diagram to illustrate trait correlations using two hypothetical examples involving fish or avian responses to temperature, assuming that temperature affects body size, which in turn affects one continuous and one categorical trait in each example. Analyses start by assembling trait measurements, where values are available for some but not all of six taxa. These conceptual models are then formalized by specifying a text file listing associations, and this in turn can generates the matrix $\boldsymbol{\Gamma}$ (for illustration we assume $\gamma = 0.5$ for all associations), and then are used to compute the covariance among traits $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^t$, where $\mathbf{L} = (\mathbf{I} - \boldsymbol{\Gamma})^{-1}\mathbf{V}^{0.5}$ and $\mathbf{V}$ represents exogenous covariance (evolutionary drift). For illustration we specify $\text{diag}(\mathbf{V}) = \mathbf{1}$ and convert the covariance to a correlation matrix, shown for each taxon. In practice, associations $\boldsymbol{\gamma}$ (used to form $\boldsymbol{\Gamma}$) and exogenous variances $\mathbf{V}$ are estimated from the fit to data (rather than specified as shown here). The covariance $\boldsymbol{\Sigma}$ is then used to generate a probabilistic prediction of missing trait values for each taxon.

each trait is a node and linkages are a directed edge, such that e.g., $Y_1 \rightarrow Y_2$ indicates that a change in trait Y_1 will cause a subsequent change in Y_2. The value of path coefficients can then be estimated as fixed effects by identifying their values that maximize the likelihood of data. This use of SEM then allows a user to replace the $n_j(n_j+1)/2$ parameters in a covariance matrix with any set of parameters (from 1, up to the maximum of $n_j(n_j+1)/2$ when not using Bayesian priors). For example, a trait-imputation model with $n_j = 30$ traits would require estimating 465 parameters for covariance $\boldsymbol{\Sigma}$ without other constraints, but could be restricted to fewer important parameters using SEM. Furthermore, SEM can be used to estimate the correlation between two traits that are connected by a directed edge ('direct pathways') or mediated by a third trait (called 'indirect pathways'). In this way, SEM decomposes the

correlation between two traits into the contribution from both direct and indirect trait effects.

In this study, we extend PTI to (1) incorporate both continuous and categorical traits, and (2) represent the trait covariance matrix using SEM, while using a Brownian motion model for simplicity of presentation. The approach can be implemented for any traits of floristic or faunistic species, and either using phylogeny for evolutionary distance or using taxonomy as a proxy for relatedness. To demonstrate the benefits of our extensions of PTI, we applied the approach to fishes which have evolutionary trade-offs that are highly structured by temperature and individual length, while also having extensive information about a variety of behavioural, reproductive, and life-history traits (Barnett et al., 2019). We specifically use data for 34 traits for >32,000 described fishes, obtained by combining existing in situ trait data (FishBase; Froese, 1990) and morphometric trait data from National Museum of Natural History fish specimens (Price et al., 2019, 2022). We interpret results by computing the direct and indirect impacts of temperature and maximum body length on other traits, and using traits to classify fishes into life-history strategies. Finally, we discuss how phylogenetic imputation of mixed traits using SEMs can help to unify experimental (micro-evolutionary) and comparative (macro-evolutionary) studies of life-history trade-offs.

2 | MATERIALS AND METHODS

We extend existing PTI methods in the following two ways:

1. *Structural equation modelling*: We model the covariance $\boldsymbol{\Sigma}$ among multiple traits using methods derived from SEM. This allows us to specify a small set of path coefficients, despite conducting multivariate trait imputation on many traits.
2. *Including categorical traits*: We fit our phylogenetic model to a mixture of continuous and categorical traits. Fitting to a categorical trait with M levels involves estimating $M-1$ latent variables, and we transform these to the probability of each level using a multivariate logistic transformation given the constraint that these probabilities sum to one. We then model the association between these $M-1$ latent variables and other continuous traits in a way that permits efficient statistical inference.

We provide further details below (see Supporting Information A for summary of all notation), and implement the approach in the package FishLife release 3.0.1 (Thorson, 2023) in the R statistical environment (R Core Team, 2021).

2.1 | Overview of phylogenetic structural equation modelling

We seek to estimate a vector of traits $\boldsymbol{\beta}_g$ for each taxon g in a rooted and additive tree (i.e., including ultrametric phylogenies), including trait-values for both tips (species) and ancestral nodes as well as the trait-vector $\boldsymbol{\beta}_0$ for the root of the tree. We assume that evolution follows a standard model (e.g., Brownian motion, Pagel's lambda, etc) that can be expressed using a multivariate normal distribution (Paradis, 2012). This model allows calculating a correlation matrix $\mathbf{R}$ with dimension $n_g \times n_g$, where n_g is the total number of taxa (tips and ancestral nodes), representing the correlation for a single trait along the phylogeny. We similarly construct the covariance $\boldsymbol{\Sigma}$ among n_j traits using methods drawn from structural equation modelling.

This then results in a separable covariance for $\mathbf{B}$ containing latent trait $\beta_{g,j}$ all taxa g and traits j:

$$\mathrm{vec}(\mathbf{B}) \sim \mathrm{MVN}\left(\mathrm{vec}\left(\mathbf{1} \otimes \boldsymbol{\beta}_0\right), \mathbf{R} \otimes \boldsymbol{\Sigma}\right), \quad (1)$$

where $\mathbf{R} \otimes \boldsymbol{\Sigma}$ is the Kroenecker ('outer') product of the correlation among taxa and covariance among traits, $\mathbf{1}$ is a vector of 1s with length n_i such that $\mathbf{1} \otimes \boldsymbol{\beta}_0$ forms the intercept for every taxon and trait, and MVN is a multivariate normal distribution with these moments. This separable covariance $\mathbf{R} \otimes \boldsymbol{\Sigma}$ can often be implemented more efficiently in some software as a conditional or simultaneous autoregressive model (Ver Hoef et al., 2018), although we present the separable covariance here to agree with standard notation in phylogenetic comparative methods (e.g., Paradis, 2012). In the following we only explore a Brownian motion (a.k.a. random-walk) process for $\mathbf{R}$, although future software developments could easily generalize this to Ornstein-Uhlenbeck, Pagel's delta, or other evolutionary models (see Supporting Information B).

We next introduce how to construct trait covariance $\boldsymbol{\Sigma}$ using methods drawn from structural equation modelling. We assume that the user specifies:

1. the structure of a path matrix $\boldsymbol{\Gamma}$ with dimension $n_j \times n_j$. The user specifies *a priori* which elements of this matrix are fixed at zero or are instead freely estimated as fixed effects (including cases when multiple path coefficients are constrained to the same estimated value). For example, specifying that $\gamma_{j,j^*} = 0$ involves assuming that trait j has no direct impact on trait j^*.
2. a Cholesky matrix $\mathbf{S}$ where $\mathbf{S}\mathbf{S}^t$ represents the covariance in exogenous variation with dimension $n_j \times n_j$. At a minimum, this covariance $\mathbf{S}\mathbf{S}^t$ involves estimating diagonal entries, $\mathrm{diag}(\mathbf{S}) = (\sigma_1, \sigma_2, \ldots, \sigma_{n_j})$ resulting in an independent exogenous variance σ_j^2 for each trait j (where these can again be constrained to the same estimated value). However, traits can also have exogenous covariance by estimating lower-triangle elements of $\mathbf{S}$, which then results in off-diagonal elements for exogenous covariance $\mathbf{V}$. Nonzero elements of $\mathbf{S}$ are then freely estimated as fixed effects.

This path matrix (and resulting path diagram) is central to structural equation modelling, which has been reviewed elsewhere for describing interaction networks and physiological performance (Frauendorf et al., 2021; Garrido et al., 2022). However, structural equation models have not to our knowledge been fitted

simultaneously with phylogenetic covariance. Previous studies have either adjusted data or estimated residual covariance based on phylogeny and then fitted a SEM to those residuals (Mason et al., 2016; Santos, 2012) or have fitted a series of phylogenetic linear models to represent dependencies in a path diagram (van der Bijl, 2018). Specifying a path diagram requires enumerating the set of variables (graph vertices) and dependencies (directed edges), where these dependencies can be interpreted as mechanisms for causal inference (Pearl, 2009). The reliability of causal inference requires correct specification of the path diagram (Grace & Irvine, 2020), and we recommend further simulation and case-study evaluation of causal inference within phylogenetic comparative methods.

These two matrices are then used to solve a simultaneous equation for $\mathbf{x} \sim \mathrm{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$, i.e., a hypothetical draw from covariance among traits $\boldsymbol{\Sigma}$ (Kaplan, 2001):

$$\begin{aligned} \mathbf{x} &= \boldsymbol{\Gamma}\mathbf{x} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim \mathrm{MVN}\left(0, \mathbf{S}\mathbf{S}^t\right). \end{aligned} \tag{2}$$

where $\boldsymbol{\Gamma}$ represents endogenous mechanisms linking variables and $\boldsymbol{\varepsilon}$ represents exogenous variation with variance $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{S}\mathbf{S}^t$. We then solve for the Cholesky of trait covariance as:

$$\mathbf{L} = (\mathbf{I} - \boldsymbol{\Gamma})^{-1}\mathbf{S}, \tag{3}$$

where trait covariance $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^t = \mathrm{Var}(\mathbf{x})$.

Constructing trait covariance $\boldsymbol{\Sigma} = (\mathbf{I} - \boldsymbol{\Gamma})^{-1}\mathbf{S}\mathbf{S}^t\left(\mathbf{I} - \boldsymbol{\Gamma}^t\right)^{-1}$ in this way generalizes several existing models:

1. *Brownian motion*: The analyst might specify $\boldsymbol{\Gamma} = \mathbf{0}$ and $\mathbf{S}$ as a diagonal matrix, and this then reduces to a standard Brownian motion model.
2. *Phylogenetic path analysis*: In some cases, variables can be reordered such that $\boldsymbol{\Gamma}$ is lower-triangular. In these cases, the model can be estimated using phylogenetic path analysis, for example fitted using piecewise SEM or d-separation methods (van der Bijl, 2018; von Hardenberg & Gonzalez-Voyer, 2013). However, $\boldsymbol{\Gamma}$ might also include loops, where for example, trait j_1 affects j_2, j_2 affects j_3, and j_3 affects j_1. This cannot be represented using standard phylogenetic path analysis but can be using SEM (Equations 2 and 3).
3. *Phylogenetic factor analysis*: In other cases, the analyst might specify $\boldsymbol{\Gamma} = \mathbf{0}$ and $\mathbf{S}$ having lower-diagonal entries that are nonzero for only a few columns. In this case, $\boldsymbol{\Sigma} = \mathbf{S}\mathbf{S}^t$ where the nonzero columns of $\mathbf{S}$ represent 'factors loadings' in a phylogenetic factor analysis (Hassler et al., 2022; Thorson et al., 2017).

In general, covariance $\boldsymbol{\Sigma}$ among n_j traits involve $n_j(n_j+1)/2$ moments, and the analyst can specify anywhere from one to $n_j(n_j+1)/2$ parameters within the two matrices $\boldsymbol{\Gamma}$ and $\mathbf{S}$. To simplify the user-interface, we require the user to specify linkages as a text file following the format of R-package SEM (Fox et al., 2020), and then parse this text file to construct $\boldsymbol{\Gamma}$ and $\mathbf{S}$ from a vector of estimated parameters.

2.2 | Fitting both continuous and categorical variables

We next outline how this model is fitted to a set of n_c continuous and n_d categorical traits, for a total of $n_t = n_c + n_d$ traits. This has been done previously using a 'threshold model' to combine categorical and continuous traits (e.g., Cybis et al., 2015; Felsenstein, 2012; Tolkoff et al., 2018), although we instead fit categorical traits using a Categorical distribution based on estimated probabilities for each categorical level (similar to Hadfield & Nakagawa, 2010). These traits are assembled in a matrix $\mathbf{Y}$ with dimension $n_i \times n_t$, where missing values are recorded as NAs and are excluded when computing the likelihood across available data. We also record the number of levels m_t for each trait t, where categorical traits have $m_t \geq 2$ by definition and we adopt the convention that $m_t = 1$ for continuous traits. Categorical traits are modelled via a probability vector that is constrained to sum to 1, so it requires $m_t - 1$ variables to describe a categorical trait with m_t levels. For trait-matrix $\mathbf{Y}$ with m_t levels for each trait t, we therefore must estimate latent trait matrix $\mathbf{B}$ with $n_j = n_c + \sum_{t=1}^{n_t}(m_t - 1)$ columns and n_g rows (where n_g is the total number of taxa in the tree). We also define a vector $\mathbf{h}$ with length n_j, where $h_j \in \{1, 2, \ldots, n_t\}$; this vector associates each column of $\mathbf{B}$ with a corresponding column of $\mathbf{Y}$. If trait t is continuous then only one value of $h_j = t$. Alternatively, if trait t is categorical then $h_j = t$ for $m_t - 1$ elements. Finally, we associate n_i rows of $\mathbf{Y}$ with n_g rows of $\mathbf{B}$ by defining a vector $\mathbf{g}$ with length n_i where g_i provides the taxon associated with sample i. The process of fitting latent traits $\mathbf{B}$ to trait measurements $\mathbf{Y}$ differs somewhat between continuous and categorical traits, as we explain next.

For a continuous trait t, we extract column $\mathbf{y}_t$ from $\mathbf{Y}$. We also extract the column from $\mathbf{B}$ for which $h_j = t$ and call this submatrix $\mathbf{B}^{(t)}$. We then specify a normal distribution for residual (measurement) variation:

$$y_{i,t} \sim \mathrm{Normal}\left(\beta_{g_i,1}^{(t)}, \sigma_j^2\right), \tag{4}$$

where σ_j^2 is the magnitude of measurement errors and is estimated as a fixed effect, although we fix $\sigma_j = 0.01$ (i.e., forcing $\beta_{g_i,j}$ to approach $y_{i,j}$) for any trait j that does not have replicated measurements and hence cannot estimate σ_j^2.

For a categorical trait t, we again extract column $\mathbf{y}_t$ from $\mathbf{Y}$. However, we then expand $\mathbf{y}_t$ to an indicator matrix $\mathbf{Z}^{(t)}$ with dimension $n_i \times m_t$, such that a trait with m_t possible levels is converted to a matrix with m_t columns where each row i contains a 1 in the column corresponding to level $y_{i,t}$ and zeros otherwise. We also extract the $m_t - 1$ columns from $\mathbf{B}$ for which $h_j = t$ and again call this submatrix $\mathbf{B}^{(t)}$. We calculate the probability $\boldsymbol{\pi}_g^{(t)}$ for each level $k \in \{1, 2, \ldots, m_t\}$ of categorical trait t via a multivariate logistic transformation of each row $\boldsymbol{\beta}_g^{(t)}$ of $\mathbf{B}^{(t)}$:

$$\pi_{g,k}^{(t)} = \begin{cases} \dfrac{e^{\beta_{g,k}^{(t)}}}{1 + \sum_{k'=1}^{m_t-1} e^{\beta_{g,k'}^{(t)}}} & \text{if } k \leq m_t - 1 \\ \dfrac{1}{1 + \sum_{k'=1}^{m_t-1} e^{\beta_{g,k'}^{(t)}}} & \text{if } k = m_t \end{cases}. \tag{5}$$

This multivariate logistic transformation converts $m_t - 1$ unbounded values in $\boldsymbol{\beta}_g^{(t)}$ to m_t probabilities $0 < \boldsymbol{\pi}_g < 1$ where $\sum_{k=1}^{m_t} \pi_{g,k} = 1$ by construction. We then fit a categorical distribution:

$$\mathbf{z}_i^{(t)} \sim \text{Categorical}\left(\boldsymbol{\pi}_{g(i)}^{(t)}\right). \quad (6)$$

This differs from previous specifications of a 'threshold model' to predict categorical traits, which have typically predicted a response of $z_{i,k} = 1$ whenever a 'liability' variable $\beta_{g(i),j}$ exceeds an estimated threshold and zero otherwise (e.g., Felsenstein, 2012). Such a threshold model must integrate across values of $\beta_{g(i),j}$ that fall on the right side of a given threshold for a measurement $y_{i,j}$, typically accomplished using Bayesian hierarchical models and MCMC sampling. By contrast, we specify that latent traits $\boldsymbol{\beta}_g^{(t)}$ for each taxon g are transformed to the probability $\boldsymbol{\pi}_g^{(t)}$ for each level of a categorical variable.

Parameters of this model remain identifiable given missing data (i.e. entries of $y_{i,t} = \text{NA}$). In these cases, the model continues to integrate across latent variables $\mathbf{B}$, and simply does not include these missing values of $\mathbf{Y}$ in the likelihood. We note that we assume trait measurements $\mathbf{Y}$ are missing at random. If the probability of having an available trait measurement (termed 'sampling intensity') is correlated with latent traits $\boldsymbol{\beta}_j$, then this assumption will result in 'preferential sampling' bias (Diggle et al., 2010). We recommend further research regarding model-based mitigation of this bias (e.g., Conn et al., 2017), but do not explore the topic further here.

2.3 | Parameter estimation and interpretation

We identify maximum likelihood estimates for all model parameters (see Supporting Information B for estimation details). This requires calculating an objective function as the product of the likelihood (Equation 4/6) and the probability of random effects (Equation 1). We obtain the marginal likelihood by integrating the objective function across random effects $\mathbf{B}$, composed of random effects $\beta_{g,j}$ for all taxa g (including tips and ancestors) and traits j. This multivariate integral is approximated using the Laplace approximation and implemented using R-package TMB, and this is computationally efficient because the inverse-covariance $(\mathbf{R} \otimes \boldsymbol{\Sigma})^{-1}$ has a value of 0 for any two taxa that are not adjacent in the specified tree (Kristensen et al., 2016). We then maximize the marginal likelihood with respect to remaining fixed effects ($\boldsymbol{\Gamma}$, $\mathbf{S}$, $\boldsymbol{\beta}_0$, and $\boldsymbol{\sigma}^2$), export the estimate of SEM-coefficients $\boldsymbol{\Gamma}$ and $\mathbf{S}$, extract 'empirical Bayes' predictions for latent traits $\mathbf{B}$ (which includes imputed values for missing trait values), and use R-package SEM to visualize the estimated path diagram.

Path coefficients $\boldsymbol{\Gamma}$ can be interpreted as a regression slope, but the precise interpretation depends upon the transformation that was chosen by the analyst for connected variables $Y_1 \rightarrow Y_2$. For example, if Y_1 is untransformed (e.g. temperature in Celsius) and Y_2 is log-transformed (e.g., log-maximum body length), then e.g., $\gamma_{1,2} = 0.1$ indicates that a 1 Celsius increase in Y_1 is associated on average with a 10% increase in Y_2. By contrast, if Y_1 is log-transformed (e.g. log-maximum body length), and Y_2 and Y_3 are two levels of a categorical variable, then $\gamma_{1,2} = 0.1$ and $Y_{1,3} = -0.1$ indicates that a 10% increase in maximum body length is associated on average with a $e^{0.1(0.1)} / e^{0.1(-0.1)} = 2\,\%$ increase in the odds of level Y_2 relative to level Y_3. We also note that the covariance among traits $\boldsymbol{\Sigma}$ is estimated as being constant across the entire phylogenetic tree (i.e., that $\text{Var}(\mathbf{B}) = \mathbf{R} \otimes \boldsymbol{\Sigma}$). In reality, slope and variance parameters may be nonstationary, representing different evolutionary trade-offs and rates resulting from environmental context and ecological traits that are not being modelled. We recommend further research extending the approach to include nonstationarity, and interpret parameters in this study as representing a sample-weighted average across the tree being analysed.

2.4 | Case study: Estimating life-history traits of fishes

To test and apply these methodological advances, we seek to estimate life-history traits for all described fishes (Chondrichthyes and Osteichthyes) included in FishBase in November 2019, where previous research has validated that these data are likely unbiased (Thorson et al., 2014). There is no phylogeny available for all fishes, despite ultrametric phylogenies existing separately for a subset of bony (Rabosky et al., 2018) and cartilaginous fishes (Stein et al., 2018). We therefore follow past research (Johnson et al., 2021; Thorson et al., 2017) in approximating phylogeny via taxonomy, that is, where all taxonomic classes are assumed to have a single common ancestor, and then including ancestral levels for order, family and genus. Package FISHLIFE then converts taxonomy to a tree using R-package APE (Paradis & Schliep, 2019), and when using taxonomy we specify phylogenetic distance $d_g = 1$ for each level of the taxonomic tree (i.e., for family to genus, genus to species, etc). We later provide a sensitivity analysis with a novel merged phylogeny.

We analyse 17 continuous-valued traits and four categorical traits, where the latter include 16 levels in total. These trait data include at least one measurement for 26,622 fish species. However, life-history data in particular are missing for many species (Figure B1), where 2%–27% of species have at least one measurement of a given trait related to growth, mortality, or body size. These 'inclusion rates' are higher for genera (7%–24%), and family levels (26%–76%), suggesting that phylogenetic information is necessary to infer trait-values for many species based on their genus or family.

We classify these 33 variables into six trait categories, expanding upon the list from Hadj-Hammou et al. (2021) where traits are broadly classified into five categories: (1) behaviour, (2) life history, (3) morphology, (4) diet and (5) physiology. The list includes at least one variable in each category (see Table 1 for details). The morphometric traits are composed of continuous measures of body shape traits that describe overall body shape for 5940 extant species of actinopterygian fishes spanning 392 families, taken on specimens at the Smithsonian Museum of Natural History and averaged by species (Price et al., 2019, 2022). These data include eight linear measurements in three dimensions: standard body and jaw length; head,

body, and caudal peduncle depth; and body, jaw, and caudal peduncle width. We standardized specimen morphometrics to account for variation in individual development for museum specimens, by dividing each measurement by the geometric mean of specimen length, width, and height.

We use several design principles to assemble the SEM for fishes, and this in turn defines the structure of SEM coefficients $\boldsymbol{\Gamma}$ and exogenous covariance $\mathbf{V}$. Specifically we specify that:

1. temperature (in Celsius) is the exogenous 'root' of the path diagram. This recognizes that life-history studies typically use temperature as a covariate to predict size and mortality (Gislason et al., 2010; Palomares et al., 2022; Pauly, 1980), and hence our estimates are comparable to widely reported slopes.
2. von Bertalanffy length (L_∞) in units mm has the greatest number of impacts on other traits, in recognition of the central role of body size in size-structured evolutionary theory (Andersen, 2019). Von Bertalanffy length is the asymptotic body size of a fish. We include linkages to other measurements of size (in mm or g), growth (in year^{-1}), and mortality parameters (in units year^{-1}), as well as to categorical traits representing reproductive behaviour, feeding mode, and habitat (Denéchère et al., 2022; Palomares et al., 2022).

TABLE 1 Life-history traits included in the analysis, listing the variable name, trait category (using five defined by Hadj-Hammou et al. (2021) while also adding 'Reproductive' as a sixth category), whether the trait is continuous or categorical, the transformation applied to continuous variables achieve a close-to-normally distributed process for evolutionary changes, and the levels for factor-valued traits.

Name	Trait category	Continuous (C) or categorical (F)	Transformation (if continuous)	Levels (if factor-valued)
age_max	Life-history	C	Natural log	—
trophic_level	Diet	C	Identity	—
aspect_ratio	Morphology	C	Natural log	—
fecundity	Reproduction	C	Natural log	—
growth_coefficient	Physiology	C	Natural log	—
temperature	Physiology	C	Identity	—
length_max	Physiology	C	Natural log	—
length_infinity	Physiology	C	Natural log	—
length_maturity	Physiology	C	Natural log	—
age_maturity	Physiology	C	Natural log	—
natural_mortality	Physiology	C	Natural log	—
weight_infinity	Physiology	C	Natural log	—
max_body_depth	Morphology	C	Natural log	—
max_body_width	Morphology	C	Natural log	—
lower_jaw_length	Morphology	C	Natural log	—
min_caudal_peduncule_depth	Morphology	C	Natural log	—
offspring_size	Reproduction	C	Natural log	—
spawning_type	Reproduction	F	—	nonguarders guarders bearers
habitat	Behaviour	F	—	demersal benthopelagic reef-associated bathymetric pelagic
feeding_mode	Diet	F	—	macrofauna planktivorous_or_other generalist
body_shape	Morphology	F	—	elongated fusiform_normal short_and_or_deep eel-like other

3. both growth and mortality rates affect age and length at maturity, in recognition that their ratio affects the optimal maturation timing (Holt, 1958).
4. for each categorical trait t (i.e., all columns j of $\mathbf{B}$ where $h_j = t$), the exogenous covariance $\mathbf{V}$ is symmetric and positive definite but otherwise unconstrained (i.e., the body-shape trait has five measured levels and involves estimating $\frac{4\times5}{2} = 10$ covariance parameters in $\mathbf{S}$), while continuous traits have independent exogenous variance (i.e., $\mathbf{S}$ is diagonal for these rows and columns).

Future research could compare fit with alternative assumptions about life-history trade-offs (e.g., Mason et al., 2016).

2.4.1 | Sensitivity, validation and performance

We assess the performance of the model, validate results, and explore sensitivity to alternative assumptions using several auxiliary analyses.

First, we compare phylogenetic structural equation modelling with the R-package PHYLOLM (Tung Ho & Ané, 2014) as widely used example of standard phylogenetic comparative methods (Supporting Information D). We specifically compare model structure, and also using a short simulation experiment with 500 replicates to confirm that the phylogenetic SEM can generate identical estimates of regression coefficients to an existing phylogenetic linear model package. For each replicate, we simulate an additive tree with 100 'tips' and randomized branch lengths and structure. We then simulate two variables under a Brownian motion model from this tree, exploring scenarios either with complete data for each taxon, or 60% of taxa missing measurements for each trait. We record the estimated slope parameter for these two models.

We also assess sensitivity of results to using taxonomic information as a proxy for evolutionary relatedness. To do so, we first merge publicly available chondrichthyan (Stein et al., 2018) and actinopterygian (Rabosky et al., 2018) ultrametric trees, using branch lengths to infer the location of their common ancestor. We then subset our data to the 11,070 species that can be matched between trait data and the merged phylogeny, and repeat the analysis on this subset. Subsetting to these matched species reduces the number of available trait measurements from 246,736 to 152,596, so we present these estimates using phylogenetic information as a sensitivity analysis.

Next, we validate the predictive performance of the model by conducting a 4-fold cross-validation experiment. To do so, we randomly partition each row of original data matrix $\mathbf{Y}$ into one of four bins (labelled $\{A, B, C, D\}$). For the first experiment, we then fit the model to all data in bins $\{B, C, D\}$ and use the estimated parameters to predict $\boldsymbol{\beta}_t$ for continuous traits and level probabilities $\boldsymbol{\pi}_g^{(t)}$ for categorical traits corresponding to data in bin A. We record these and then repeat this process for the other three bins, comparing predictions with the withheld data. This experiment evaluates performance when predicting new data that are collected via the same process as the original data set (Roberts et al., 2017), and we recommend future research use a blocked cross-validation design to explore performance when predicting traits for taxa that are systematically under-represented in available data.

We then evaluate performance separately for continuous and categorical traits:

- *Continuous traits*: for continuous trait t, we plot unfitted observations $\mathbf{y}_t$ against the out-of-bag predictions $\boldsymbol{\beta}_j^{(t)}$ (where $\boldsymbol{\beta}_j^{(t)}$ is the column of $\mathbf{B}$ for which $h_j = t$), and also calculate the percent variance explained relative to a null model that predicts $\mathbf{y}_t$ based on its mean value $\bar{y}_t$:

$$PVE_t = 1 - \frac{\sum_{i=1}^{n_i}\left(y_{i,t} - \beta_{g(i),j}^{(t)}\right)^2}{\sum_{i=1}^{n_i}\left(y_{i,t} - \bar{y}_t\right)^2}. \quad (7)$$

  PVE_t predicts the proportion of variance that would be explained for a hypothetical 'new' sample, where a value of 0 indicates no out-of-bag explanatory power (i.e. no improvement relative to predicting new samples as the mean of all data) and a value of 1 implies perfect explanatory power.
- *Categorical traits*: for latent trait $\boldsymbol{\beta}_j^{(t)}$ representing a level of a categorical trait, remember that we expand original data $\mathbf{y}_t$ to an indicator matrix $\mathbf{Z}^{(t)}$ where $\mathbf{z}_k^{(t)}$ is the column corresponding to level k of latent trait t. This indicator column has value 0 when a taxon does not have that level and 1 when it does, while the model estimates the probability $\boldsymbol{\pi}_k^{(t)}$ for that level of the categorical trait, and these probabilities sum to one across levels $k \in \{1, 2, \ldots, m_t\}$. To evaluate model performance, we plot the receiver operator characteristics (ROC) curve for each level, which involves calculating the rate of false-positives and false-negatives when converting the predicted probability to a predicted indicator using different potential threshold values. We then calculate the area under the ROC (AUC) using R package pROC (Robin et al., 2011), where an AUC of 0.5 indicates no out-of-bag ability to discriminate between 0 and 1 values for an indicator, and an AUC of 1 implies perfect ability to discriminate between these.

2.4.2 | Identifying life-history strategies

We illustrate results by identifying a small number of life-history strategies for fishes, defined as an extreme combination of trait values that frequently occur together, such that all fishes can be characterized as some mixture of strategies (i.e., following the usage in Winemiller & Rose, 1992). Previous studies have applied clustering methods to a smaller subset of species than we have available, e.g., for North American fishes (Winemiller & Rose, 1992), selected North Pacific marine fishes (King & McFarlane, 2003), freshwater fishes (Mims et al., 2010), or European marine fishes (Pecuchet et al., 2017). However, our study is the first to predict the life-history strategies for all described fishes worldwide, representing more than 34,000 species.

We specifically follow Winemiller and Rose (1992) in estimating 'archetypes' that represent an extreme combination of life-history characteristics. All fish taxa are then described as a finite mixture of these archetypes, and we refer to these archetypes as 'life-history strategies'. This contrasts with other studies that have clustered taxa continuously within the space of life-history traits (King & McFarlane, 2003). To do so, we extract predictions $\boldsymbol{\beta}^{(t)}$ for continuous traits and level probabilities $\boldsymbol{\pi}_g^{(t)}$ for categorical traits for all taxa that have at least one observation (i.e., are not purely drawn from the predictive distribution based on its taxa). We then apply 'archetypal analysis' (Cutler & Breiman, 1994) following methods from Pecuchet et al. (2017), using R package archetypes (Eugster & Leisch, 2009). Archetypal analysis involves estimating n_b 'archetypes' $\boldsymbol{\alpha}_b$ composed of values $\alpha_{b,j}$, representing the value of variable j in archetype b. Each taxon $\boldsymbol{\beta}_g$ is then predicted as a finite mixture of these archetypes, with mixture coefficients $p_{g,b}$ defined such that $\sum_{k=1}^{n_k} p_{g,b} = 1$ and $p_{g,b} > 0$. Archetypal analysis then estimates the value of $\alpha_{g,b}$ and $p_{b,j}$ to minimize the sum of squared distance (SSD) between predicted and inputted $\mathbf{B}$. We use a scree-plot to visualize how the SSD decreases when using 1–6 archetypes and we select the number by visually identifying when further increases generate little improvement in SSD. We then explore the results in two ways:

- *Archetype trait values*: We extract trait values for estimated archetypes, $\boldsymbol{\alpha}_b$, to interpret which traits are associated with each. We specifically convert $\alpha_{b,j}$ to a percent score $\alpha^*_{b,j}$ by calculating the proportion of fishes having a predicted trait $\boldsymbol{\beta}_j < \alpha_{b,j}$.
- *Simplex by taxonomy*: Similarly, we extract mixture coefficients $\mathbf{p}_g$ for each taxon. We then use package archetypes to apply a skew-orthogonal transformation to visualize $p_{g,b}$ in a two dimensional simplex (Seth & Eugster, 2014). We specifically compare $p_{g,b}$ for major taxa and compare resulting assignments with previous studies (Winemiller & Rose, 1992).

3 | RESULTS

The phylogenetic structural equation model quantified the direct impact of temperature on size and growth. Specifically, a one degree Celsius increase was associated with a 4% increase in growth coefficient (with standard error SE = 0.3%), 2% increase (SE = 0.2%) in mortality rate, and 2% decrease (SE = 0.3%) in asymptotic body length (Figure 2; Table E1), where these represent average associations across the wide range of fishes being analysed. In turn, a 10% increase in asymptotic body length was associated with an 8.2% (SE = 0.3%) decrease in natural mortality and a 6.6% decrease (SE = 0.2%) in growth coefficient. When both direct and indirect effects are included (Table E2), temperature had a slightly larger impact on the growth coefficient (0.051) than on the mortality rate (0.035). Temperature was estimated to have a minimal effect on reproductive behaviour, feeding mode, or spawning type, while asymptotic length had a larger effect on these traits (Table E2). For example, a 10% increase in asymptotic length was estimated to decrease the odds of guarding behaviour relative to non-guarding behaviour by 34% (Table E2). Finally, the model also captured previously documented life-history trade-offs, including the association between earlier maturation and higher relative mortality (Figure E1).

The simulation experiment confirmed that *FishLife* and the widely used R-package PHYLOLM give essentially identical estimates when fitting continuous traits and data are available for all species (Figure D1, left panel), and that *FishLife* shows a small improvement in estimation performance when data are missing at random (Figure D1, right panel). Four-fold cross-validation confirmed that the model fitted to real-world data had good performance when predicting records that were randomly dropped from the model fitting (Figure 3). Continuous-valued traits had a percent-variance explained (PVE) ranging from 51% to 89%. Among these variables, performance was particularly high (>80% PVE) for traits measuring length, weight, and fecundity, but lower for traits measuring age, growth, maturity, and trophic level. Similarly, levels of categorical traits had an area under the receiver-operator-characteristics curve (ROC) ranging from 0.86 to 0.99, with lower (but still high) power to discriminate levels for the feeding-mode trait. Comparing the model fitted using taxonomy with one using a subset of data and phylogeny to represent evolutionary distance (Figure E2) shows similar estimates of linkages for life-history parameters (i.e., for mortality, growth, size, and maturity parameters) between analyses. However, the estimated impact of body size on body shape was substantially larger when using phylogeny.

Our approach is further demonstrated by the archetype analysis, which identified three life-history strategies (Figure E2), in agreement with Winemiller and Rose (1992). The first archetype (purple in Figure 4 and top panel in Figure 5) was associated with higher maximum age, trophic level, slow growth, and low temperatures. This suite of traits corresponded to the 'equilibrium' strategy from Winemiller and Rose (1992). The third archetype (yellow in Figure 4 and bottom panel in Figure 5) corresponded to the opportunistic strategy from Winemiller and Rose (1992). It had the lowest maximum age and fecundity, while having high natural mortality and probability of guarding their young. Finally, the second archetype was somewhat intermediate in terms of growth and size, while typically having highest fecundity, being mainly pelagic and having the highest probability of a non-guarding reproductive strategy. As expected, there was strong phylogenetic signal in these life-history strategies, with Elasmobranchii (sharks and rays) representing the equilibrium strategy, Clupeidae (herrings and sardines) largely representing the periodic strategy, and Gobiidae (gobies) largely representing the opportunistic strategy (Figure 6).

4 | DISCUSSION

We extended phylogenetic trait-imputation methods to include two additional features: (1) representing the covariance among traits via a structural equation model, and (2) incorporating both continuous and categorical traits. We fit categorical traits using latent variables

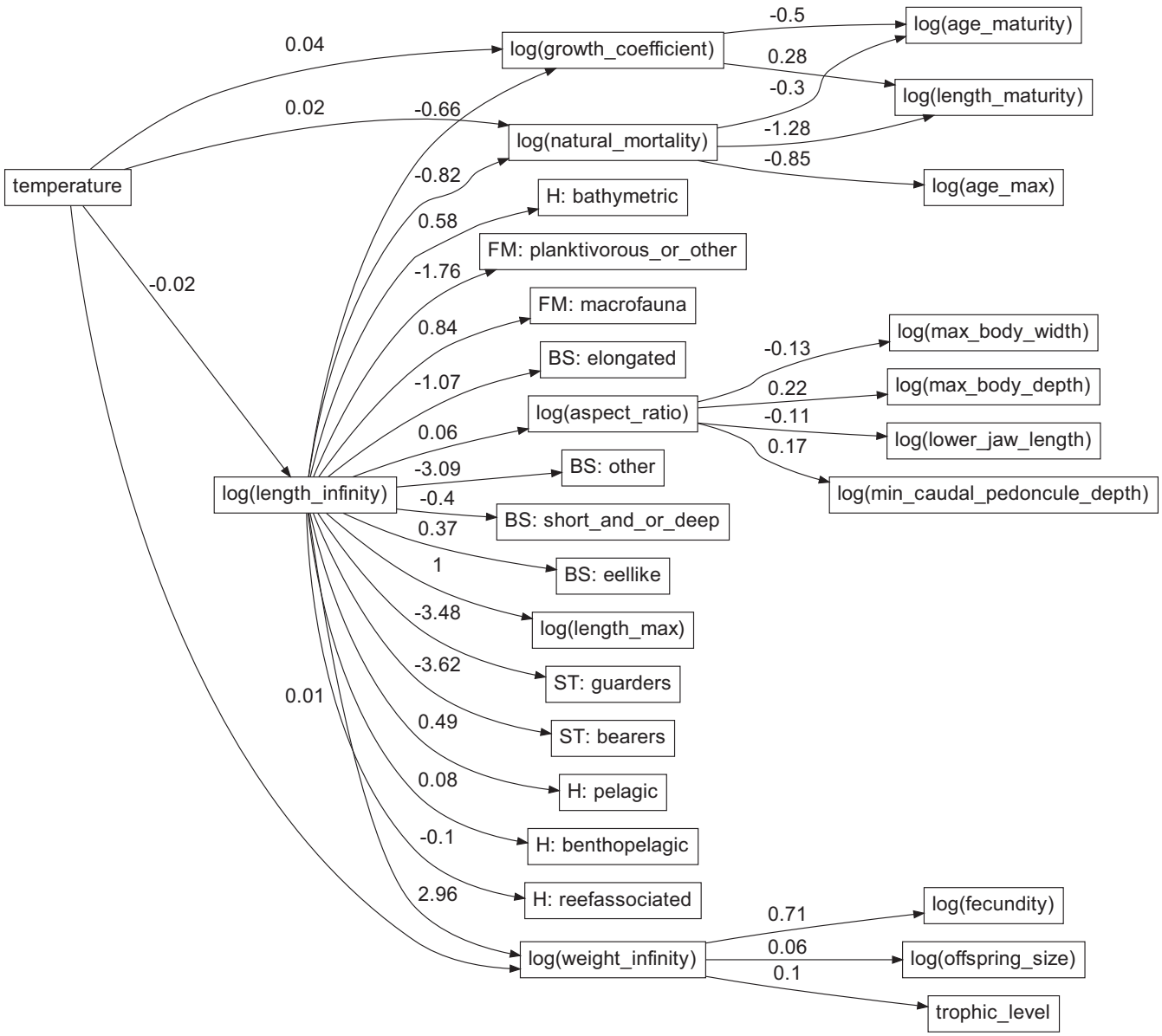

FIGURE 2 Path diagram representing specified causal linkages and estimated $\boldsymbol{\Gamma}$ coefficients (see Figure 1 for description) linking fish traits when using taxonomy to represent evolutionary distance, using package sem to generate the plot (Fox et al., 2020), where levels of the categorical variables are abbreviated (H: habitat; FM: feeding mode; BS: body shape; ST: spawning type) and coefficients for categorical variables represent the log-odds relative to a specified base level (H: demersal; FM: generalist; BS: fusiform/normal; ST: nonguarders). Note that evolutionary variance and covariance parameters $\boldsymbol{\Sigma}$ are not shown here for clarity of presentation.

that are then transformed to calculate the probability for each categorical level. Unlike past analyses (e.g. Felsenstein, 2012), however, we use a computational method (the Laplace approximation) that allows rapid inference on large trees. These two developments have wide relevance for applications across life-history databases for any taxonomic group within various ecosystems worldwide such as plants, mammals, fishes, birds, insects, as well as comparing across these taxa (Capdevila et al., 2020). For example, Kattge et al. (2011) documented 52 traits for 69,000 plant species in the TRY global plant database, of which 15 are categorical including Mycorrhiza type, nitrogen fixation capacity, and pollination mode. In addition, GRooT (Guerrero-Ramírez et al., 2021) includes 38 root traits, from 38,276 species-by-site mean values based on 114,222 trait records, for more than 1000 species, such as root mass fraction, root carbon and nitrogen concentration, lateral spread, root mycorrhizal colonization intensity, mean root diameter, root tissue density, specific root length and maximum rooting depth. Similarly, the bird trait database AvoNET (Tobias et al., 2022) includes continuous morphological traits but also categorical traits like trophic level (three levels), foraging niche (nine levels) and foraging locomotory behaviour (five levels) for 11,009 species. Likewise, the foraging database EltonTraits (Wilman et al., 2014) includes foraging time as a categorical trait for 9993 bird and 5400 mammal species. Clearly there is potential for both phylogenetic signal within these

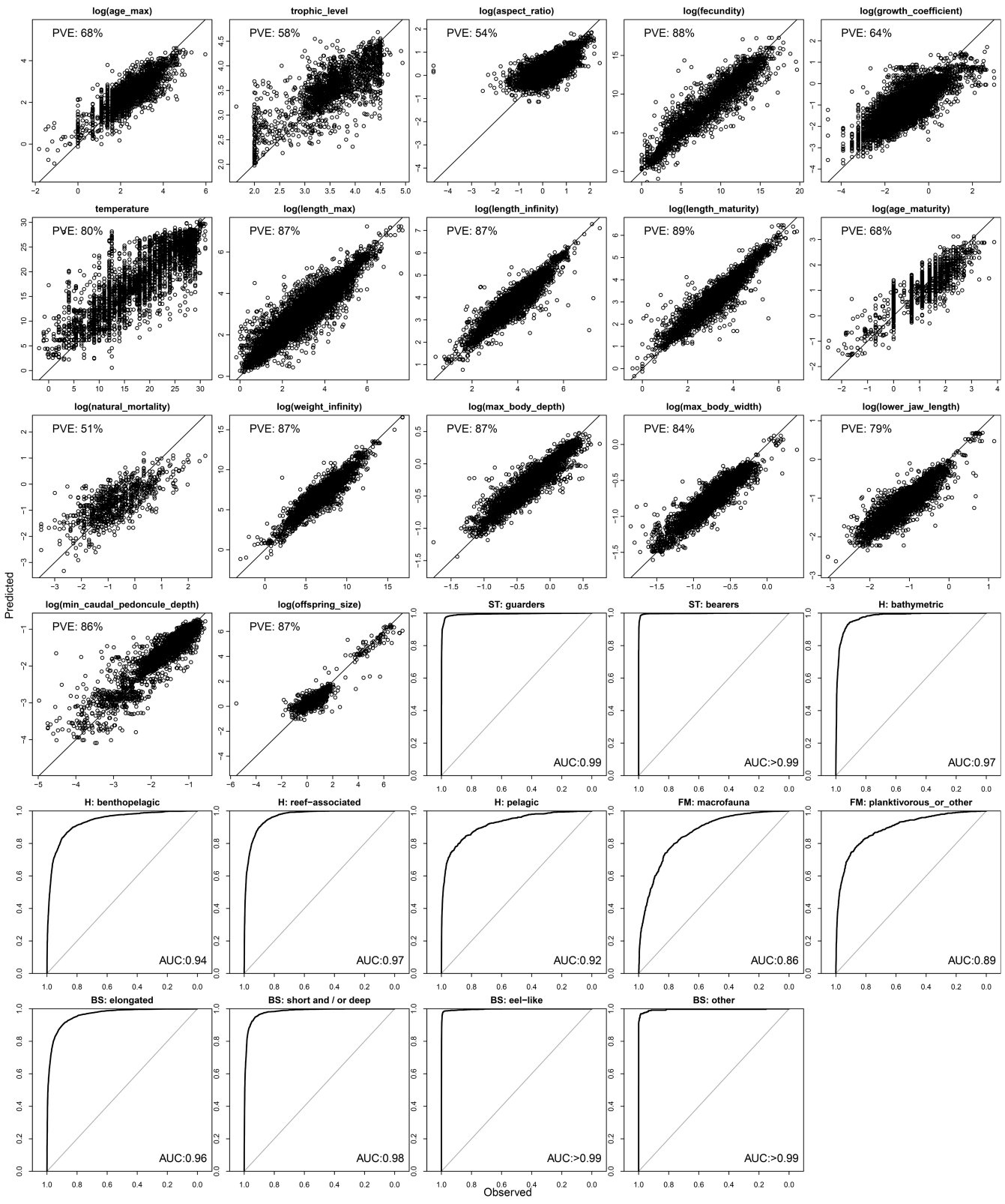

FIGURE 3 Evaluating predictive performance for all variables based on a four-fold cross-validation experiment. For continuous-valued traits, plots show the held-out value (*x*-axis) against the predicted value (*y*-axis), along with the one-to-one line (black line) and list the percent-variance-explained (PVE). A well-performing model will have predictions near the one-to-one line and a PVE approaching 100%. For discrete-valued traits, we used the held-out factor-level indicator (0 or 1) and the predicted class probability to calculate the receiver-operator characteristics curve (ROC). A well-performing model will have ROC in the upper-left corner and an AUC approaching 1.0.

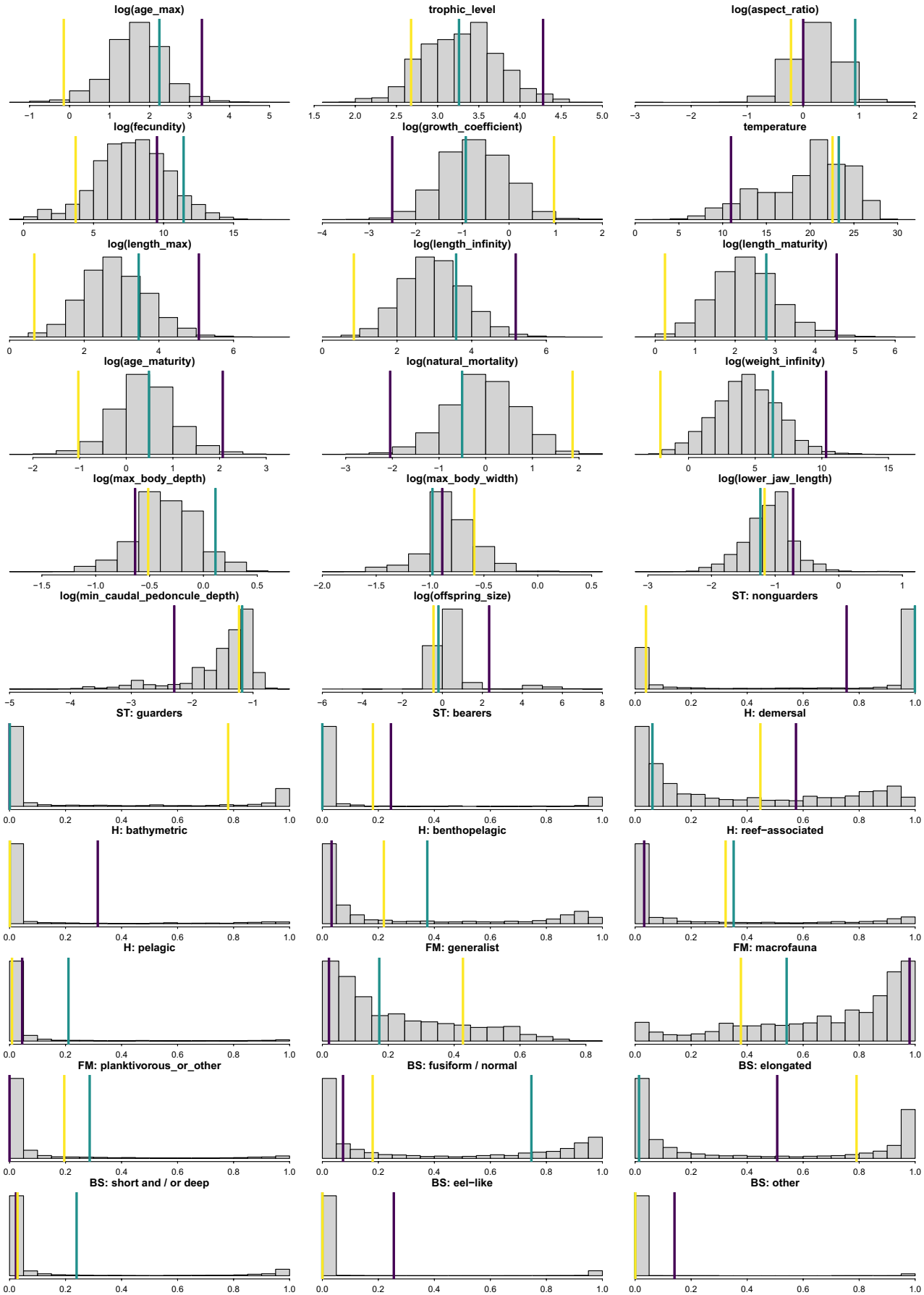

FIGURE 4 Frequency distribution (*y*-axis) for estimated values (*x*-axis) for each life-history trait (panels) with the trait-value for each of three life-history strategies identified using the 'archetype' analysis (vertical lines; purple: Equilibrium; green: Periodic; yellow: Opportunistic).

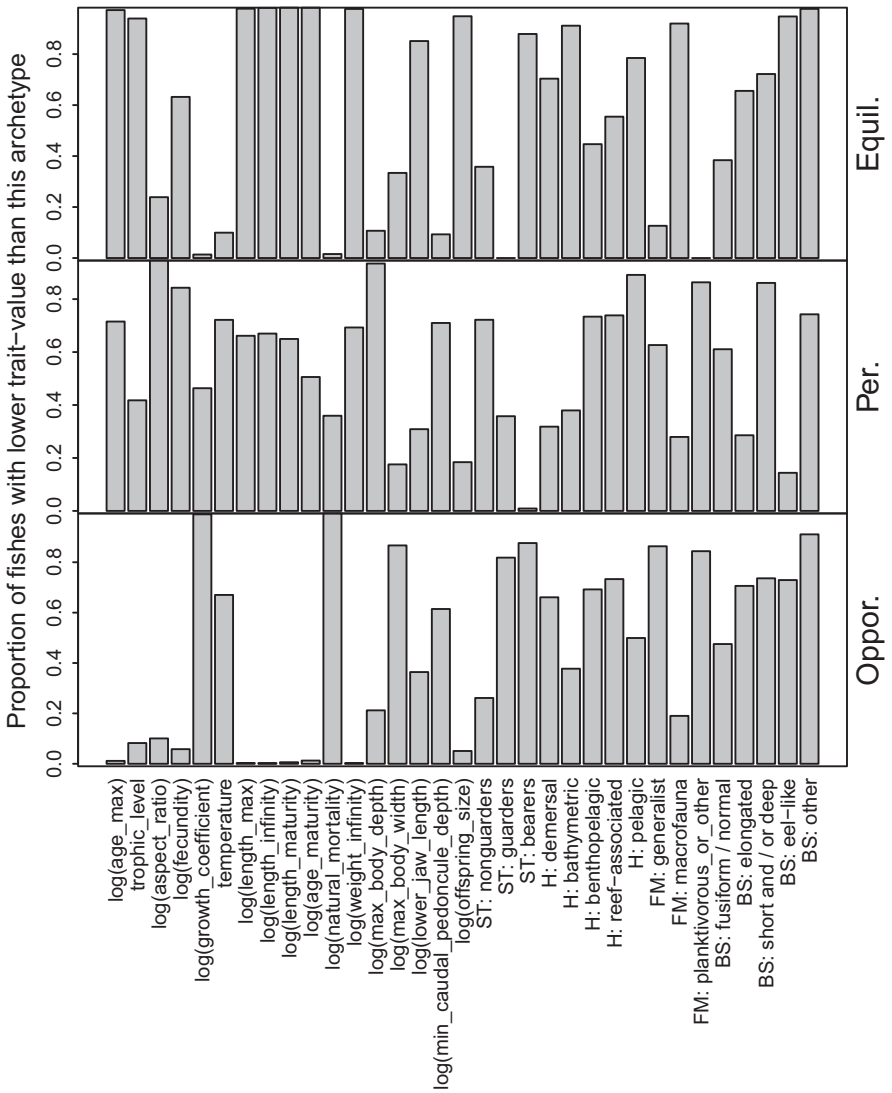

FIGURE 5 Illustration of traits associated with each estimated life-history strategy (equilibrium, periodic and opportunistic), specifically showing the proportion of species with a trait-value lower than that of a given archetype (*y*-axis) for each trait (*x*-axis) and archetype (panel).

categorical traits, and ecologically meaningful relationships between qualitative and quantitative traits. This is clearly illustrated in our case study, which estimates that fishes with smaller adult body sizes (a continuous trait) are more likely to guard their young (a categorical trait) in agreement with recent theoretical prediction (Denéchère et al., 2022). Similarly, categorical traits (e.g. reproductive behaviour in fishes, or nitrogen fixation in plants) can be highly relevant when measuring functional diversity or predicting responses to new competitors or climates. Overall, this underlines the importance of including categorical traits when imputing traits for both regional and macroecological studies.

We also argue that SEM will be increasingly attractive for phylogenetic trait imputation as the number of traits increases. This utility arises because phylogenetic trait imputation with n_j traits typically requires on the order of n_j^2 parameters for the covariance among traits (Bruggeman et al., 2009), or $n_j n_f$ parameters when specifying n_f factors that represent major axes of covariance among traits (Hassler et al., 2022; Thorson et al., 2017). These approaches scale rapidly with an increase in the number of traits, which becomes prohibitive when there are many traits to consider, such as in the TRY database version-5 containing 2100 traits. By contrast, SEM allows customized specification of the number of parameters, ranging from 1 (i.e., identical evolution rate for each trait) to $n_j(n_j+1)/2$. Furthermore, path parameters in $\mathbf{\Gamma}$ are interpretable as regression slopes, such that individual parameters can be compared with pre-existing theory about trait linkages, whether from field observations or laboratory experiments. In our study for example, we estimate a nearly isometric (2.96) scaling of asymptotic body length and body mass and a linear scaling of asymptotic and maximum length (0.99), and these parameters are easily corroborated when evaluating model plausibility. Indeed, future SEMs could consider fixing these and other parameters a priori to improve parsimony and the resulting precision for difficult-to-estimate trait linkages. Alternatively, we estimate the total (direct and indirect) impact of log(length) on log(natural mortality) of −0.82, and this differs somewhat from the inverse relationship claimed by Lorenzen et al. (2022), such that in some cases it is helpful to test for differences relative to existing theory. Finally, SEM starts by specifying a graph (where nodes represent variables, and edges represent dependencies), which can

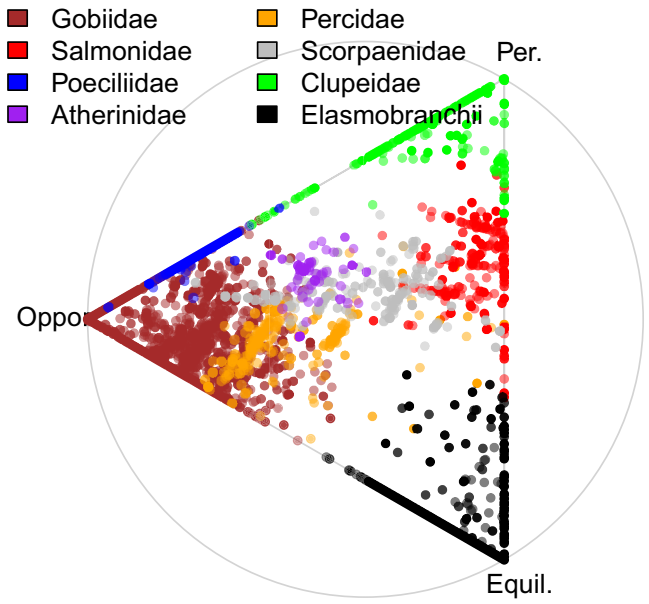

FIGURE 6 Illustration of where species (circles) within each family or class (colours, using legend in top-left) fall among three life-history strategies (Per = Periodic; Oppor = Opportunistic; Equil = Equilibrium), noting (1) that *Sebastes* within *Scorpaenidae* is expected to be between Periodic and Equilibrium, (2) *Etheostorna* within *Percidae* is expected between Equilibrium and Opportunistic, (3) *Salmonidae* are expected between Equilibrium and Periodic, and (4) *Gambusia* within *Poecilidae* are expected to be opportunistic (Winemiller & Rose, 1992). Similarly, Equilibrium fishes are expected to exploit a stable environment while Periodic exploit large-scale and/or seasonal patches (Winemiller & Rose, 1992), which we illustrate by showing *Elasmobranchii* and *Clupeidae*, respectively. Finally, Pecuchet et al. (2017) discussed *Gobiidae* as an example of the Opportunistic strategy.

be readily derived from existing conceptual or theoretical models for a given taxonomic group. For example, length-structured models for fish evolution have already derived boldness as a function of exogenous changes in mortality rate (Andersen et al., 2018) or temperature (Neubauer & Andersen, 2019), and future research could adapt these graphical models within multivariate trait imputation.

We suggest that in the future SEM and causal inference may help unite research at disparate scales. Causal inference is often described using graphical models, and if the path diagram accurately describes the real-world processes then the resulting estimates can be interpreted as causal mechanisms (Laubach et al., 2021). Phylogenetic trait imputation using SEM provides new avenues to combine laboratory and natural experiments (micro-evolution) with comparative studies of life-history trade-offs (macro-evolution). This may be particularly important for studies of global change biology. For example, natural experiments in insects suggest that increased temperatures can lead to longer flight season (e.g., a wider range of days where adults are present) (Merckx et al., 2021). This observed relationship could then be supplied when using SEM to conduct phylogenetic trait imputation of an insect database, such as the Odonate Phenotypic Database (Waller et al., 2019). Similarly, rearing experiments involving artificial harvest of fishes suggests that changes in mortality will negatively impact age-at-maturity (van Wijk et al., 2013). However, these experimental results have not previously been used in comparative analyses of life-history parameters. Ultimately, we hope that the use of SEM in phylogenetic trait imputation will contribute to the ongoing discussion between experimental and observational studies of trade-offs in floristic or faunal traits. Allowing the explicit recognition of the theoretical assumptions implied in phylogenetic trait imputation will provide a new avenue for ecological theory to be applied in community diversity and macroecology.

AUTHOR CONTRIBUTIONS

James Thorson, Romain Frelat and Aurore Maureaud conceived the ideas and designed methodology; Sarah Friedman, Maria L. D. Palomares, Samantha Price and Peter Wainwright collected the data; Jennifer Bigman merged the publicly available phylogenies for fishes; James Thorson and Aurore Maureaud analysed the data; and James Thorson led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

Support for SAP & PCW was provided by the National Science Foundation (grant DEB-1556953). This research is performed through discussions within the FISHGLOB group, 'Fish biodiversity under global change: a worldwide assessment from scientific trawl surveys', co-funded by the national synthesis centers for biodiversity, ecology and evolution CESAB of the French Foundation for Research on Biodiversity (FRB; www.fondationbiodiversite.fr), CIEE (www.ciee-icee.ca) and the French Embassy in Canada. Support for JSB was provided by a National Science Foundation Postdoctoral Research Fellow in Biology (PRFB) Fellowship to JSB (grant 2109411). We are grateful to the FISHGLOB members for providing discussion on this research, and in particular L. Pecuchet for numerous discussions regarding fish traits and the archetypes package. We also thank J. Cope, L. Barnett and three anonymous reviewers for comments on an earlier draft.

CONFLICT OF INTEREST STATEMENT

We have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

All data inputs, parameter estimates, trait predictions, and software are available publicly in R package FishLife (https://github.com/James-Thorson-NOAA/FishLife) using release number 3.0.1 (Thorson, 2023).

ORCID

James T. Thorson https://orcid.org/0000-0001-7415-1010
Aurore A. Maureaud https://orcid.org/0000-0003-4778-9443
Romain Frelat https://orcid.org/0000-0002-8631-4398
Bastien Mérigot https://orcid.org/0000-0001-5264-4324
Jennifer S. Bigman https://orcid.org/0000-0001-8070-3061

Sarah T. Friedman https://orcid.org/0000-0003-0192-5008
Maria Lourdes D. Palomares https://orcid.org/0000-0001-5905-1556
Malin L. Pinsky https://orcid.org/0000-0002-8523-8952
Samantha A. Price https://orcid.org/0000-0002-1389-8521
Peter Wainwright https://orcid.org/0000-0003-0498-4759

REFERENCES

Andersen, K. H. (2019). *Fish ecology, evolution, and exploitation*. Princeton University Press. https://press.princeton.edu/books/hardcover/9780691176550/fish-ecology-evolution-and-exploitation

Andersen, K. H., Marty, L., & Arlinghaus, R. (2018). Evolution of boldness and life history in response to selective harvesting. *Canadian Journal of Fisheries and Aquatic Sciences*, *75*(2), 271–281. https://doi.org/10.1139/cjfas-2016-0350

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research*, *20*(1), 40–49. https://doi.org/10.1002/mpr.329

Baker, J., Humphries, S., Ferguson-Gow, H., Meade, A., & Venditti, C. (2020). Rapid decreases in relative testes mass among monogamous birds but not in other vertebrates. *Ecology Letters*, *23*(2), 283–292. https://doi.org/10.1111/ele.13431

Barnett, L. A. K., Jacobsen, N. S., Thorson, J. T., & Cope, J. M. (2019). Realizing the potential of trait-based approaches to advance fisheries science. *Fish and Fisheries*, *20*(5), 1034–1050. https://doi.org/10.1111/faf.12395

Bruggeman, J., Heringa, J., & Brandt, B. W. (2009). PhyloPars: Estimation of missing parameter values using phylogeny. *Nucleic Acids Research*, *37*(suppl_2), W179–W184. https://doi.org/10.1093/nar/gkp370

Capdevila, P., Beger, M., Blomberg, S. P., Hereu, B., Linares, C., & Salguero-Gómez, R. (2020). Longevity, body dimension and reproductive mode drive differences in aquatic versus terrestrial life-history strategies. *Functional Ecology*, *34*(8), 1613–1625. https://doi.org/10.1111/1365-2435.13604

Cardillo, M., Mace, G. M., Gittleman, J. L., Jones, K. E., Bielby, J., & Purvis, A. (2008). The predictability of extinction: Biological and external correlates of decline in mammals. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1641), 1441–1448. https://doi.org/10.1098/rspb.2008.0179

Conn, P. B., Thorson, J. T., & Johnson, D. S. (2017). Confronting preferential sampling when analysing population distributions: Diagnosis and model-based triage. *Methods in Ecology and Evolution*, *8*(11), 1535–1546. https://doi.org/10.1111/2041-210X.12803

Cutler, A., & Breiman, L. (1994). Archetypal analysis. *Technometrics*, *36*(4), 338–347. https://doi.org/10.1080/00401706.1994.10485840

Cybis, G. B., Sinsheimer, J. S., Bedford, T., Mather, A. E., Lemey, P., & Suchard, M. A. (2015). Assessing phenotypic correlation through the multivariate phylogenetic latent liability model. *The Annals of Applied Statistics*, *9*(2), 969–991. https://doi.org/10.1214/15-AOAS821

Debastiani, V. J., Bastazini, V. A., & Pillar, V. D. (2021). Using phylogenetic information to impute missing functional trait values in ecological databases. *Ecological Informatics*, *63*, 101315.

Denéchère, R., van Denderen, P. D., & Andersen, K. H. (2022). Deriving population scaling rules from individual-level metabolism and life history traits. *The American Naturalist*, *199*(4), 564–575. https://doi.org/10.1086/718642

Díaz, S., Purvis, A., Cornelissen, J. H. C., Mace, G. M., Donoghue, M. J., Ewers, R. M., Jordano, P., & Pearse, W. D. (2013). Functional traits, the phylogeny of function, and ecosystem service vulnerability. *Ecology and Evolution*, *3*(9), 2958–2975. https://doi.org/10.1002/ece3.601

Diggle, P. J., Menezes, R., & Su, T. (2010). Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, *59*(2), 191–232.

Eugster, M. J. A., & Leisch, F. (2009). From Spider-Man to Hero - Archetypal Analysis inR. *Journal of Statistical Software*, *30*(8). https://doi.org/10.18637/jss.v030.i08

Felsenstein, J. (2012). A comparative method for both discrete and continuous characters using the threshold model. *The American Naturalist*, *179*(2), 145–156.

Fox, J., Nie, Z., & Byrnes, J. (2020). *Sem: Structural equation models*. R Package Version 3.1-11. https://CRAN.R-project.org/package=sem

Frauendorf, M., Allen, A. M., Verhulst, S., Jongejans, E., Ens, B. J., van der Kolk, H.-J., de Kroon, H., Nienhuis, J., & van de Pol, M. (2021). Conceptualizing and quantifying body condition using structural equation modelling: A user guide. *Journal of Animal Ecology*, *90*(11), 2478–2496. https://doi.org/10.1111/1365-2656.13578

Froese, R. (1990). FishBase: An information system to support fisheries and aquaculture research. *ICLARM Fishbyte*, *8*(3), 21–24.

Gallagher, R. V., Falster, D. S., Maitner, B. S., Salguero-Gómez, R., Vandvik, V., Pearse, W. D., Schneider, F. D., Kattge, J., Poelen, J. H., Madin, J. S., Ankenbrand, M. J., Penone, C., Feng, X., Adams, V. M., Alroy, J., Andrew, S. C., Balk, M. A., Bland, L. M., Boyle, B. L., … Enquist, B. J. (2020). Open Science principles for accelerating trait-based science across the tree of life. *Nature Ecology & Evolution*, *4*(3), 294–303. https://doi.org/10.1038/s41559-020-1109-6

Garrido, M., Hansen, S. K., Yaari, R., & Hawlena, H. (2022). A model selection approach to structural equation modelling: A critical evaluation and a road map for ecologists. *Methods in Ecology and Evolution*, *13*(1), 42–53. https://doi.org/10.1111/2041-210X.13742

Gislason, H., Daan, N., Rice, J. C., & Pope, J. G. (2010). Size, growth, temperature and the natural mortality of marine fish. *Fish and Fisheries*, *11*(2), 149–158.

Goolsby, E. W., Bruggeman, J., & Ané, C. (2017). Rphylopars: Fast multivariate phylogenetic comparative methods for missing data and within-species variation. *Methods in Ecology and Evolution*, *8*(1), 22–27. https://doi.org/10.1111/2041-210X.12612

Grace, J. B. (2006). *Structural equation modeling and natural systems*. Cambridge University Press. https://doi.org/10.1017/CBO9780511617799

Grace, J. B., & Irvine, K. M. (2020). Scientist's guide to developing explanatory statistical models using causal analysis principles. *Ecology*, *101*(4), e02962. https://doi.org/10.1002/ecy.2962

Gross, N., Le Bagousse-Pinguet, Y., Liancourt, P., Saiz, H., Violle, C., & Munoz, F. (2021). Unveiling ecological assembly rules from commonalities in trait distributions. *Ecology Letters*, *24*(8), 1668–1680. https://doi.org/10.1111/ele.13789

Guerrero-Ramírez, N. R., Mommer, L., Freschet, G. T., Iversen, C. M., McCormack, M. L., Kattge, J., Poorter, H., van der Plas, F., Bergmann, J., Kuyper, T. W., York, L. M., Bruelheide, H., Laughlin, D. C., Meier, I. C., Roumet, C., Semchenko, M., Sweeney, C. J., van Ruijven, J., Valverde-Barrantes, O. J., … Weigelt, A. (2021). Global root traits (GRooT) database. *Global Ecology and Biogeography*, *30*(1), 25–37. https://doi.org/10.1111/geb.13179

Hadfield, J. D., & Nakagawa, S. (2010). General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, *23*(3), 494–508. https://doi.org/10.1111/j.1420-9101.2009.01915.x

Hadj-Hammou, J., Mouillot, D., & Graham, N. A. J. (2021). Response and effect traits of coral reef fish. *Fronters in Marine Science*, *8*, 640619. https://doi.org/10.3389/fmars.2021.640619

Hassler, G. W., Gallone, B., Aristide, L., Allen, W. L., Tolkoff, M. R., Holbrook, A. J., Baele, G., Lemey, P., & Suchard, M. A. (2022).

Principled, practical, flexible, fast: A new approach to phylogenetic factor analysis. *Methods in Ecology and Evolution*, *13*(10), 2181–2197. https://doi.org/10.1111/2041-210X.13920

Hevia, V., Martín-López, B., Palomo, S., García-Llorente, M., de Bello, F., & González, J. A. (2017). Trait-based approaches to analyze links between the drivers of change and ecosystem services: Synthesizing existing evidence and future challenges. *Ecology and Evolution*, *7*(3), 831–844. https://doi.org/10.1002/ece3.2692

Holt, S. J. (1958). The evaluation of fisheries resources by the dynamic analysis of stocks, and notes on the time factors involved. *ICNAF Special Publication*, *1*, 77–95.

Johnson, T. F., Isaac, N. J. B., Paviolo, A., & González-Suárez, M. (2021). Handling missing values in trait data. *Global Ecology and Biogeography*, *30*(1), 51–62. https://doi.org/10.1111/geb.13185

Kaplan, D. (2001). Structural equation modeling. In N. J. Smelser & P. B. Baltes (Eds.), *International encyclopedia of the Social & Behavioral Sciences* (pp. 15215–15222). Pergamon. https://doi.org/10.1016/B0-08-043076-7/00776-2

Kattge, J., Díaz, S., Lavorel, S., Prentice, I. C., Leadley, P., Bönisch, G., Garnier, E., Westoby, M., Reich, P. B., Wright, I. J., Cornelissen, J. H. C., Violle, C., Harrison, S. P., Van BODEGOM, P. M., Reichstein, M., Enquist, B. J., Soudzilovskaia, N. A., Ackerly, D. D., Anand, M., … Wirth, C. (2011). TRY—A global database of plant traits. *Global Change Biology*, *17*(9), 2905–2935. https://doi.org/10.1111/j.1365-2486.2011.02451.x

King, J. R., & McFarlane, G. A. (2003). Marine fish life history strategies: Applications to fishery management. *Fisheries Management and Ecology*, *10*(4), 249–264.

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*(5), 1–21. https://doi.org/10.18637/jss.v070.i05

Laubach, Z. M., Murray, E. J., Hoke, K. L., Safran, R. J., & Perng, W. (2021). A biologist's guide to model selection and causal inference. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1943), 20202815. https://doi.org/10.1098/rspb.2020.2815

Legras, G., Loiseau, N., Gaertner, J.-C., Poggiale, J.-C., Ienco, D., Mazouni, N., & Mérigot, B. (2019). Assessment of congruence between co-occurrence and functional networks: A new framework for revealing community assembly rules. *Scientific Reports*, *9*(1), 19996. https://doi.org/10.1038/s41598-019-56515-7

Lorenzen, K., Camp, E. V., & Garlock, T. M. (2022). Natural mortality and body size in fish populations. *Fisheries Research*, *252*, 106327. https://doi.org/10.1016/j.fishres.2022.106327

Mason, C. M., Goolsby, E. W., Humphreys, D. P., & Donovan, L. A. (2016). Phylogenetic structural equation modelling reveals no need for an 'origin' of the leaf economics spectrum. *Ecology Letters*, *19*(1), 54–61. https://doi.org/10.1111/ele.12542

Merckx, T., Nielsen, M. E., Heliölä, J., Kuussaari, M., Pettersson, L. B., Pöyry, J., Tiainen, J., Gotthard, K., & Kivelä, S. M. (2021). Urbanization extends flight phenology and leads to local adaptation of seasonal plasticity in lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(40), e2106006118. https://doi.org/10.1073/pnas.2106006118

Mims, M. C., Olden, J. D., Shattuck, Z. R., & Poff, N. L. (2010). Life history trait diversity of native freshwater fishes in North America. *Ecology of Freshwater Fish*, *19*(3), 390–400. https://doi.org/10.1111/j.1600-0633.2010.00422.x

Neubauer, P., & Andersen, K. H. (2019). Thermal performance of fish is explained by an interplay between physiology, behaviour and ecology. *Conservation Physiology*, *7*(1), coz025. https://doi.org/10.1093/conphys/coz025

Pacifici, M., Visconti, P., Butchart, S. H. M., Watson, J. E. M., Cassola, F. M., & Rondinini, C. (2017). Species' traits influenced their response to recent climate change. *Nature Climate Change*, *7*(3), 205–208. https://doi.org/10.1038/nclimate3223

Palomares, M. L. D., Parducho, V. A., Reyes, R., & Bailly, N. (2022). The interrelationship of temperature, growth parameters, and activity level in fishes. *Environmental Biology of Fishes*, *105*, 1475–1479. https://doi.org/10.1007/s10641-022-01261-5

Paradis, E. (2012). *Analysis of phylogenetics and evolution with R* (Vol. *2*). Springer.

Paradis, E., & Schliep, K. (2019). Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

Pauly, D. (1980). On the interrelationships between natural mortality, growth parameters, and mean environmental temperature in 175 fish stocks. *Journal du Conseil International pour l'Exploration de la Mer*, *39*(2), 175–192. https://doi.org/10.1093/icesjms/39.2.175

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*, 96–146. https://doi.org/10.1214/09-SS057

Pecuchet, L., Lindegren, M., Hidalgo, M., Delgado, M., Esteban, A., Fock, H. O., Gil de Sola, L., Punzón, A., Sólmundsson, J., & Payne, M. R. (2017). From traits to life-history strategies: Deconstructing fish community composition across European seas. *Global Ecology and Biogeography*, *26*(7), 812–822. https://doi.org/10.1111/geb.12587

Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Young, B. E., Graham, C. H., & Costa, G. C. (2014). Imputation of missing data in life-history trait datasets: Which approach performs the best? *Methods in Ecology and Evolution*, *5*(9), 961–970. https://doi.org/10.1111/2041-210X.12232

Price, S. A., Friedman, S. T., Corn, K. A., Larouche, O., Brockelsby, K., Lee, A. J., Nagaraj, M., Bertrand, N. G., Danao, M., Coyne, M. C., Estrada, J. R., Friedman, R., Hoeft, E., Iwan, M., Gross, D., Kao, J. H., Landry, B., Linares, M. J., McGlinn, C., … Wainwright, P. C. (2022). FishShapes v1: Functionally relevant measurements of teleost shape and size on three dimensions. *Ecology*, *103*(12), e3829. https://doi.org/10.1002/ecy.3829

Price, S. A., Friedman, S. T., Corn, K. A., Martinez, C. M., Larouche, O., & Wainwright, P. C. (2019). Building a body shape Morphospace of Teleostean fishes. *Integrative and Comparative Biology*, *59*(3), 716–730. https://doi.org/10.1093/icb/icz115

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rabosky, D. L., Chang, J., Title, P. O., Cowman, P. F., Sallan, L., Friedman, M., Kaschner, K., Garilao, C., Near, T. J., Coll, M., & Alfaro, M. E. (2018). An inverse latitudinal gradient in speciation rate for marine fishes. *Nature*, *559*(7714), 392–395. https://doi.org/10.1038/s41586-018-0273-1

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. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929. https://doi.org/10.1111/ecog.02881

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*, *12*(1), 77. https://doi.org/10.1186/1471-2105-12-77

Santos, J. C. (2012). Fast molecular evolution associated with high active metabolic rates in poison frogs. *Molecular Biology and Evolution*, *29*(8), 2001–2018. https://doi.org/10.1093/molbev/mss069

Schrodt, F., Kattge, J., Shan, H., Fazayeli, F., Joswig, J., Banerjee, A., Reichstein, M., Bönisch, G., Díaz, S., Dickie, J., Gillison, A., Karpatne, A., Lavorel, S., Leadley, P., Wirth, C. B., Wright, I. J., Wright, S. J., & Reich, P. B. (2015). BHPMF—A hierarchical Bayesian approach to gap-filling and trait prediction for macroecology and functional biogeography. *Global Ecology and Biogeography*, *24*(12), 1510–1521. https://doi.org/10.1111/geb.12335

Seth, S., & Eugster, M. J. A. (2014). Probabilistic archetypal analysis. *ArXiv:1312.7604 [Stat]*. http://arxiv.org/abs/1312.7604

Stein, R. W., Mull, C. G., Kuhn, T. S., Aschliman, N. C., Davidson, L. N. K., Joy, J. B., Smith, G. J., Dulvy, N. K., & Mooers, A. O. (2018). Global priorities for conserving the evolutionary history of sharks, rays and chimaeras. *Nature Ecology & Evolution*, *2*(2), 288–298. https://doi.org/10.1038/s41559-017-0448-4

Taugourdeau, S., Villerd, J., Plantureux, S., Huguenin-Elie, O., & Amiaud, B. (2014). Filling the gap in functional trait databases: Use of ecological hypotheses to replace missing data. *Ecology and Evolution*, *4*(7), 944–958. https://doi.org/10.1002/ece3.989

Thorson, J. T. (2020). Predicting recruitment density dependence and intrinsic growth rate for all fishes worldwide using a data-integrated life-history model. *Fish and Fisheries*, *21*(2), 237–251. https://doi.org/10.1111/faf.12427

Thorson, J. T. (2023). *FishLife: Predict life history parameters for any fish, release 3.0.1*. https://doi.org/10.5281/zenodo.7590994. https://github.com/James-Thorson-NOAA/FishLife

Thorson, J. T., Cope, J. M., & Patrick, W. S. (2014). Assessing the quality of life history information in publicly available databases. *Ecological Applications*, *24*(1), 217–226. https://doi.org/10.1890/12-1855.1

Thorson, J. T., Munch, S. B., Cope, J. M., & Gao, J. (2017). Predicting life history parameters for all fishes worldwide. *Ecological Applications*, *27*(8), 2262–2276. https://doi.org/10.1002/eap.1606

Tobias, J. A., Sheard, C., Pigot, A. L., Devenish, A. J. M., Yang, J., Sayol, F., Neate-Clegg, M. H. C., Alioravainen, N., Weeks, T. L., Barber, R. A., Walkden, P. A., MacGregor, H. E. A., Jones, S. E. I., Vincent, C., Phillips, A. G., Marples, N. M., Montaño-Centellas, F. A., Leandro-Silva, V., Claramunt, S., … Schleuning, M. (2022). AVONET: Morphological, ecological and geographical data for all birds. *Ecology Letters*, *25*(3), 581–597. https://doi.org/10.1111/ele.13898

Tolkoff, M. R., Alfaro, M. E., Baele, G., Lemey, P., & Suchard, M. A. (2018). Phylogenetic factor analysis. *Systematic Biology*, *67*(3), 384–399. https://doi.org/10.1093/sysbio/syx066

Tung Ho, L. si, & Ané, C. (2014). A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait Evolution Models. *Systematic Biology*, *63*(3), 397–408. https://doi.org/10.1093/sysbio/syu005

van der Bijl, W. (2018). Phylopath: Easy phylogenetic path analysis in R. *PeerJ*, *6*, e4718.

van Wijk, S. J., Taylor, M. I., Creer, S., Dreyer, C., Rodrigues, F. M., Ramnarine, I. W., van Oosterhout, C., & Carvalho, G. R. (2013). Experimental harvesting of fish populations drives genetically based shifts in body size and maturation. *Frontiers in Ecology and the Environment*, *11*(4), 181–187.

Ver Hoef, J. M., Hanks, E. M., & Hooten, M. B. (2018). On the relationship between conditional (CAR) and simultaneous (SAR) autoregressive models. *Spatial Statistics*, *25*, 68–85. https://doi.org/10.1016/j.spasta.2018.04.006

Violle, C., Navas, M.-L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., & Garnier, E. (2007). Let the concept of trait be functional! *Oikos*, *116*(5), 882–892. https://doi.org/10.1111/j.0030-1299.2007.15559.x

von Hardenberg, A., & Gonzalez-Voyer, A. (2013). Disentangling evolutionary cause-effect relationships with phylogenetic confirmatory path analysis. *Evolution; International Journal of Organic Evolution*, *67*(2), 378–387. https://doi.org/10.1111/j.1558-5646.2012.01790.x

Waller, J. T., Willink, B., Tschol, M., & Svensson, E. I. (2019). The odonate phenotypic database, a new open data resource for comparative studies of an old insect order. *Scientific Data*, *6*(1), 316. https://doi.org/10.1038/s41597-019-0318-9

Wilman, H., Belmaker, J., Simpson, J., de la Rosa, C., Rivadeneira, M. M., & Jetz, W. (2014). EltonTraits 1.0: Species-level foraging attributes of the world's birds and mammals. *Ecology*, *95*(7), 2027. https://doi.org/10.1890/13-1917.1

Winemiller, K. O., & Rose, K. A. (1992). Patterns of life-history diversification in north American fishes: Implications for population regulation. *Canadian Journal of Fisheries and Aquatic Sciences*, *49*, 2196–2218.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Supplementary Materials A: List of all notation.

Supplementary Materials B: Constructing the correlation matrix for an additive tree

Supplementary Materials C: Illustrating missing data.

Supplementary Materials D: Comparison with phylogenetic comparative methods.

Supplementary Materials E: Detailed results.

How to cite this article: Thorson, J. T., Maureaud, A. A., Frelat, R., Mérigot, B., Bigman, J. S., Friedman, S. T., Palomares, M. L. D., Pinsky, M. L., Price, S. A., & Wainwright, P. (2023). Identifying direct and indirect associations among traits by merging phylogenetic comparative methods and structural equation models. *Methods in Ecology and Evolution*, *14*, 1259–1275. https://doi.org/10.1111/2041-210X.14076