POINT OF VIEW

Running head: DEEP COMPARATIVE METHODS

On the Promise (and Practice) of Deep Comparative Methods: Expressivity, Limitations, and Design

Caleb Charpentier[1,], Arthur Porto[2], Anuj Karpatne[3] and Josef C. Uyeda[1]

[1] Department of Biological Sciences, Virginia Tech, Blacksburg, VA 24061, U.S.A.
[2] Florida Museum of Natural History, University of Florida, Gainesville, FL 32605, U.S.A.
[3] Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL, U.S.A.
Corresponding author: calebcharpentier00@gmail.com

Keywords: Traits, Representation Learning, Phylogenetic Comparative Methods, Evolutionary Inference, Measurement Theory

Abstract

Comparative biologists explain the diversity of organismal form and function by deciding which traits to measure and which evolutionary processes to model. These decisions make inference possible, but they also confine it to traits, relationships, and processes that can be specified a priori. Neural networks can expand this space, learning characters directly from complex observations and supporting inference under models without tractable likelihoods. This expressivity comes with a characteristic cost. Neural networks learn from statistical regularities in their training data rather than from the entities and processes biologists use to formulate explanations. This difference reflects a deeper mismatch in reasoning: neural networks generally extend statistical relationships learned from existing data to new observations (induction), whereas comparative biologists often seek the evolutionary processes that best explain how observed patterns arose (abduction). As a result, learned representations can rely on incidental features, fail under distribution shift, or admit many statistically equivalent but biologically distinct interpretations, which post hoc explainability methods cannot by themselves resolve. We argue that this mismatch is a design problem rather than an intrinsic limitation. Architectures, loss functions, training data, and optimization procedures each offer points at which researchers can impose abductive constraints: biologically motivated restrictions on the invariances, scales, and structure of learned representations that make them commensurable with evolutionary theory. Simulations, benchmarks, and targeted validation then test whether those constraints succeeded. We call neural networks designed and evaluated in this way *Deep Comparative Methods*. Their promise lies not in eliminating the need to define biological meaning, but in expanding the range of candidate measurements and explanations that can be made explicit and empirically tested.

Introduction

> Rapid progress on any biological problem rests on the hope that there is at least one viewpoint to each problem that makes causation relatively simple. — Houle et al. (2010, p. 857)

> For every complex problem there is a solution which is clear, simple, and wrong. — H. L. Mencken, paraphrased[1]

To understand how organismal traits evolve, researchers must first decide how to measure them (Houle et al., 2011). These decisions depend on the biological questions the researcher seeks to answer. Systematists construct characters around hypotheses of homology (Freudenstein, 2005), while researchers interested in form, function, or heritability may use theoretical morphospaces (Evans et al., 2021; Raup, 1966; Raup and Michelson, 1965), developmental models (Cano-Fernández et al., 2024), or particularly significant axes of genetic variation, such as genetic lines of least resistance (Schluter, 1996). Whatever the context, measurement choices always emphasize some aspects of biological variation at the expense of others, constraining the kinds of relationships that comparative models can represent and limiting the evolutionary explanations they can support (Houle et al., 2011). Whether analyzing images and CT scans, examining museum specimens, collecting field observations, or assembling organismal trait data from the literature, comparative biologists repeatedly face the same challenge: transforming this complex and varied evidence into representations that retain the information needed for evolutionary inference (Houle et al., 2010).

Over the past decade, a growing number of researchers have framed machine learning and AI as a way to meet this informational challenge. Reviews have surveyed applications across phenomics, ecology, evolutionary morphology, and

[1] The commonly quoted formulation paraphrases Mencken's observation that "there is always a well-known solution to every human problem—neat, plausible, and wrong" (Mencken, 1921, p. 158).

phylogenetics (Borowiec et al., 2022; He et al., 2024; Lürig et al., 2021; Mo et al., 2024). More practically, researchers have developed tools that automate time-consuming tasks such as image segmentation, landmark placement, and trait measurement across large datasets (Di Martino et al., 2023; Feng et al., 2025; Lürig, 2022; Porto and Voje, 2020). With this technology at hand, it is easy to imagine that several major challenges in the biodiversity sciences—connecting phenotypes to genotypes (Houle et al., 2010), linking micro- and macroevolutionary processes (Machado et al., 2023), or understanding how intrinsic constraints shape species' responses to climate change (Goswami et al., 2015)—might yield to more data, or to more kinds of data (Houle, 2010; Liow et al., 2023).

In this Perspective, we argue that **overcoming the key bottlenecks to discovery in biodiversity science requires more than just faster measurement or larger datasets**. A deeper constraint is *expressivity*—the range of traits, relationships, and evolutionary processes that our measurements and models can *represent*. We examine this constraint through an emerging area of methodological research in biodiversity science that we call Deep Comparative Methods (DCMs). DCMs insert deep learning into the comparative inferential pipeline to construct biological trait representations (Boyko and Rabosky, 2026) or to infer evolutionary patterns and processes (Bokma, 2006; Nagel and Landis, 2026) with less direct human oversight than in traditional measurement procedures. We refer to these components, respectively, as *character construction* and *model construction*.

Whereas traditional methods generally require researchers to define traits from their observations and specify evolutionary models in advance, DCMs can learn trait representations directly from images, videos, or CT scans and infer evolutionary patterns without evaluating an explicit likelihood. An important mathematical basis for this flexibility is universal function approximation: sufficiently flexible neural networks can approximate a broad class of

relationships between numerical inputs and outputs (Cybenko, 1989; Hornik et al., 1989) (Fig. 1). Importantly, many rich, complex sources of data, such as images, videos, CT scans, text, and conventional trait measurements can all be encoded numerically. DCMs can learn representations from these sources without requiring researchers to define all traits in advance. Along with inference that does not require a tractable likelihood, this flexibility has the potential to reveal evolutionary patterns or relationships that traditional analyses would leave unexplored (Adaimé et al., 2025; Cuthill et al., 2019). Furthermore, DCMs can apply this expressivity to many of the same questions as traditional phylogenetic comparative methods (PCMs), including trait relationships, evolutionary history, dimensionality, and more (Caetano and Hearn, 2026; Collyer and Adams, 2021; Felsenstein, 1985; Martin and Weber, 2026; Uyeda and Harmon, 2014).

The expressive capacity of DCMs that opens new avenues for biological discovery also creates an interpretive problem. Neural networks do not *a priori* need to organize information in terms of concepts familiar to biologists (Pierucci et al., 2026). Their representations capture abstract statistical patterns among pixels, image features, or vertices in 3D meshes, but those patterns do not automatically tell us which biological structures or evolutionary processes they reflect. Translating them into claims about real organisms and causal mechanisms requires additional evidence and judgment, and leaves uncomfortable room for confirmation bias or uncritical "just-so" explanations (Gould and Lewontin, 1979; Hogg and Villar, 2024). Unfortunately, these difficulties are most acute where DCMs may be most useful, when high-dimensional data resist reduction to human-defined measurements (Houle, 2010) or complex evolutionary scenarios lack tractable likelihood functions (Nagel and Landis, 2026). Scientific inference depends on knowing what our measurements represent and whether they are appropriate for the analyses that follow (Houle et al., 2011). If neural networks—often criticized as opaque black boxes—have their greatest interpretive difficulties when the underlying biology

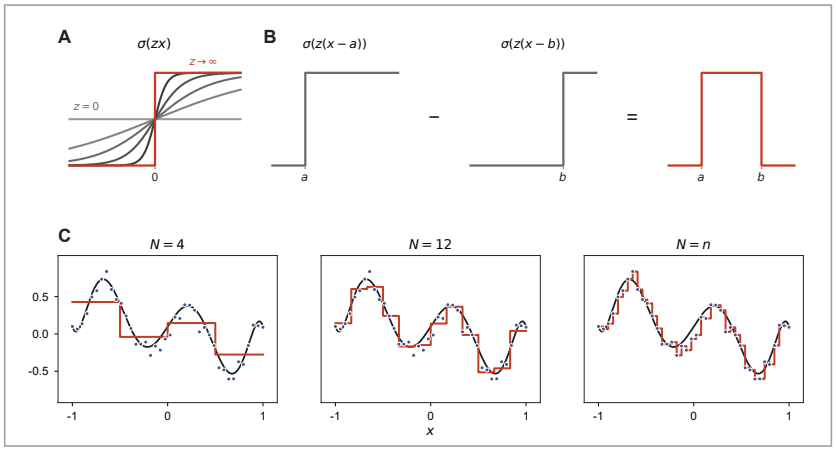

Figure 1: **An intuition for universal function approximation.** A neural network with one hidden layer can form a weighted sum of sigmoid functions applied to weighted and shifted inputs. With that in mind, this figure illustrates the intuition behind the universal approximation results of Cybenko (1989) and Hornik et al. (1989) in one input dimension. More generally, neural networks can approximate mappings between high-dimensional numerical inputs and outputs. Because images, videos, CT scans, text, and conventional trait measurements can all be encoded numerically, this capacity allows neural networks to map complex patterns in any of these data sources onto other numerical representations or predictions. **(A)** As the input weight z increases, a sigmoid becomes steeper and approaches a step function. **(B)** Subtracting two shifted steps produces a rectangular function over an interval. **(C)** Combining such functions produces arbitrarily fine approximations to an underlying curve (black). Blue points are noisy observations, and red lines are approximations using N intervals. With $N = n$, one interval per observation, the approximation follows the noise as well as the curve. Although networks with finite sigmoid weights remain continuous, sufficiently large networks can approximate increasingly complex functions and may fit noise along with the underlying relationship. Universal approximation therefore establishes what a network *can* represent, but not necessarily what it *will* learn or how well it will generalize.

would most benefit from their ability to represent complexity, how can we expect DCMs to possibly improve the situation?

Rather than treating these problems as isolated limitations of deep learning, we propose a novel research program centered on how DCMs can balance expressivity with biological meaning in character and model construction. This Perspective is organized as follows. First, we establish the practical scope of DCMs and show how researchers have used them to increase expressivity at various steps of the comparative inference pipeline. Next, we develop a

conceptual framework for DCMs, drawing on ideas from Representation Learning, Representational Measurement Theory, and abductive reasoning (key terms are defined in Table 1). Using these ideas, we then establish a taxonomy of the ways DCMs "think" differently from human biologists. Finally, we lay out a practical toolbox for addressing these differences and supporting meaningful inference with DCMs, which we hope will serve as the practical foundations for further work in DCMs as an emerging field of methodological research.

How DCMs Generalize the Inferential Pipeline in Practice

Although the term Deep Comparative Methods is new, the approaches it describes span several decades and a wide range of comparative problems. Representative studies are summarized in Table 2. One useful way to organize these approaches is by identifying the step in comparative inference that they automate or generalize within character or model construction. We describe six such steps below, including (1) character definition, (2) character association, (3) character interpretation, (4) evolutionary parameter estimation, (5) model selection, and (6) direct prediction.

Character definition[2] is an early step in the inferential pipeline in which researchers delimit phenotypic variation into traits, states, or dimensions and specify the scales or spaces used to represent them. This ordering is explicit in methods such as Phylogenetic Generalized Least Squares (PGLS) (Grafen, 1989), where traits must be measured and represented numerically before their associations can be estimated. Character definitions can nevertheless depend on the downstream analysis. In cladistic analysis, for example, characters and states delimited as hypotheses of primary homology may be revised through reciprocal illumination with the reconstructed phylogeny (de Pinna, 1991). More generally, characters defined with too many or too few states, or on inappropriate scales,

[2] We generalize the term from Pogue and Mickevich (1990), who applied it specifically to discrete morphological characters in phylogenetic inference.

can bias inference or make models unnecessarily complex or misleading (Huttegger and Mitteroecker, 2011; Khakurel et al., 2024; Machado et al., 2019, 2023; Mitteroecker and Huttegger, 2009; Simões et al., 2018). While researchers have traditionally defined characters manually using systematic expertise, DCMs can learn candidate character systems directly from data. Tsutsumi et al. (2023), Stevens et al. (2025), and Hartmann et al. (2026) learned morphological representations of mandibles, images, and footprints, while Adaïmé et al. (2024) and Hofmann et al. (2024) incorporated phylogenetic, taxonomic, or genetic information into character definition.

Character association concerns how phenotypic variation relates to other traits, environments, or evolutionary history. Comparative methods have long addressed such relationships among predefined characters, for example by accounting for shared ancestry or identifying multivariate directions associated with phylogenetic structure (Collyer and Adams, 2021; Felsenstein, 1985; Grafen, 1989; Hansen et al., 2008). DCMs extend this idea by learning phenotypic representations from complex observations and identifying the variation associated with other biological variables, or by allowing those variables to shape the representation itself. Lürig et al. (2025), for example, used high-dimensional butterfly wing-pattern representations and chemical-defense data to derive a visual axis of aposematic variation, whereas Dinnage and Kleineberg (2025) showed that learned representations of three-dimensional bird bills contained variation predictive of trophic niche beyond phylogenetic structure. Other approaches use taxonomic or phylogenetic information directly during representation learning, including hierarchical labels, ancestry levels, or evolutionary distances (Elhamod et al., 2023; Gu et al., 2025; Hunt et al., 2025; Khurana et al., 2025; Lau and Choi, 2026; Stevens et al., 2024).

Character interpretation involves determining what learned traits represent biologically, a step usually implicit in conventional comparative studies because researchers define characters from biological structures or properties they already

intuitively understand. A researcher measuring femur length, for example, must first recognize what a femur *is* before deciding how to measure its length. Neural networks, in contrast, can construct numerical descriptions of phenotypic variation directly from raw data without requiring a human to specify what physical variation those descriptions encode. A trait or dimension 'discovered' by a neural network might partly encode femur length while also responding to unrelated and potentially ineffable statistical artifacts in the image (Elhage et al., 2022). This black box adds a new layer of interpretive uncertainty about what learned characters mean, beyond the usual uncertainty in parameter estimates or predictions.

Although a vast literature exists on interpreting neural-network representations and predictions, three approaches have been especially prominent in DCMs. The first, *feature attribution*, identifies input regions that influence a learned trait or prediction. In Zhang et al. (2025), for example, Finer-CAM highlighted that wing coloration distinguishes Blue Grosbeaks from Grandalas. The second, dimensionality reduction, uses tools such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), or Linear Discriminant Analysis (LDA) to reveal structure in learned representations, as in Lürig et al. (2024), who used PCA on BioEncoder to reveal separate axes of coloration and body shape in damselfly cross-sex mimicry. The final approach, generative sampling, generates synthetic phenotypes by sampling learned representations or conditioning the model on biological variables, as in Dinnage and Kleineberg (2025), who used a Variational Autoencoder to generate bird bills conditioned on trophic niche and identify niche-associated variation in bill shape.

Evolutionary Parameter Estimation is the inference of parameter values that describe an evolutionary process from observed data. In traditional comparative methods, these parameters are typically estimated under an explicit model with a tractable likelihood function—for example, a model of diversification (Morlon, 2014; Morlon et al., 2024) or morphological trait evolution (Blomberg et al., 2003;

Cornwallis and Griffin, 2024; Felsenstein, 1985; Hansen, 1997; Pagel, 1994). Some inference problems, however, do not easily lend themselves to explicit, tractable likelihoods. While likelihood-free methods such as Approximate Bayesian Computation (ABC) provide one solution in these cases (Csilléry et al., 2010; Xie et al., 2026), DCMs offer another approach. Using simulations, researchers can implement custom evolutionary processes that do not lend themselves to tractable likelihoods, then train DCMs on the resulting datasets to estimate their parameters from observed data. This approach has been applied across several areas, including diversification analyses (Lambert et al., 2023), phylodynamics and phylogeography (Thompson et al., 2024), molecular evolution and phylogenetic inference (Silvestro et al., 2024), and morphological trait evolution (Jones, 2026). More recently, Landis and Thompson (2025) developed phyddle, a Python-based deep-learning pipeline for evolutionary parameter estimation, model selection, and other phylogenetic inference tasks. To support inference under flexible phylogenetic models, including those without tractable likelihoods, phyddle accepts simulation scripts written in any language and automatically formats the resulting datasets for training DCMs (Landis and Thompson, 2025).

Model selection involves choosing among alternative evolutionary models that could explain an observed data distribution. Traditional comparative analyses usually compare candidate models through their likelihoods. The Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), and Bayes factors do this after fitting each model explicitly (Posada and Buckley, 2004; Sullivan and Joyce, 2005), while Reversible-jump Markov chain Monte Carlo (RJMCMC) estimates model identity and parameter values together (Huelsenbeck et al., 2004; Pagel and Meade, 2006). DCMs instead turn model selection into a classification problem, allowing researchers to choose among alternative model classes without first evaluating their likelihoods or fitting an evolutionary model. For example, Jones (2026) used simulated datasets to train a DCM to distinguish among continuous-character models such as Brownian motion,

Ornstein–Uhlenbeck, and Early Burst models of continuous character evolution. Similarly, Soewongsono and Landis (2026) used phyddle (Landis and Thompson, 2025) to distinguish among diversity-dependent biogeographic models in which local richness differentially affected speciation, extinction, and dispersal. In both cases, the DCM predicted which candidate evolutionary process generated the data rather than estimating parameters within a model chosen in advance.

Direct prediction concerns predicting some biological quantity or historical outcome without explicitly estimating an intermediate model and its parameters. For example, while traditional Ancestral State Reconstruction (ASR) methods require first choosing an evolutionary model and fitting its parameters, DCMs can be trained on simulated datasets to predict the trait values of ancestral nodes directly (Moreta et al., 2022; Nagel and Landis, 2026). In other cases, researchers have used similar approaches to do deep learning-based species distribution modeling with DCMs (Estopinan et al., 2024), or to estimate targets like the timing of mass extinction events (Du et al., 2026) or biodiversity trajectories across time (Cooper et al., 2024). These approaches bypass explicit model fitting without eliminating model assumptions, which remain encoded in the simulated or empirical data used for training.

In most of the inferential steps above, DCMs generalize character or model construction by skipping steps that are explicit in traditional comparative workflows. In character association, phenotypic data can be related to ecological or other biological factors without first defining a fixed set of characters. In model selection and direct prediction, DCMs can choose among alternative evolutionary models without estimating their parameters, or make predictions without choosing an explicit model at all. Skipping or combining earlier steps can increase expressivity by avoiding upstream decisions that otherwise constrain the inferential possibilities downstream. In ASR, for example, predictions are generally conditioned on an explicit, likelihood-based evolutionary model. DCM-based ASR can instead learn a direct mapping from

observed data to ancestral states, so the form of that relationship need not be specified as a single evolutionary model in advance (Nagel and Landis, 2026).

Character interpretation is the exception to this pattern because it adds an inferential step rather than removing one: DCMs can construct learned characters before researchers know which physical variation they encode. This interpretive gap is not necessarily a fundamental limitation, but addressing it requires a theory of biological measurement that accounts for learned representations.

Measurement and Meaning for DCMs

What would a theory of measurement for DCMs need to explain? We approach this question by bringing together two bodies of work that understand representation from different perspectives: Representation Learning and Representational Measurement Theory. Their contrast helps clarify how DCMs can gain expressivity from learned representations while remaining meaningful in biological terms.

In the Representation Learning literature, representations are often understood *operationally* as learned features that facilitate prediction or capture explanatory factors in the data (Bengio et al., 2013). In practice, models learn these representations from input data in complex formats such as images, videos, CT scans, three-dimensional meshes, and graphs, mapping each observation to a vector of real-valued numbers within an *embedding space* (Zhang et al., 2023, see also Box 1). Researchers then examine how embedded observations are organized within this space and interpret its structure *post hoc*. Methods such as PCA, LDA, t-SNE, and Sparse Autoencoders (SAEs) can identify directions, clusters, or features across these embeddings (Huben et al., 2024); for example, clusters may separate species (Gu et al., 2025), while particular directions or features may correspond to more abstract concepts such as gender or the presence of recognizable entities like the Golden Gate Bridge (Templeton et al.,

2024). Representations are thus learned by models with varying degrees of autonomy, and human researchers project meaning back onto them after the fact. Importantly, issues such as polysemanticity, superposition, and non-identifiability can leave these interpretations underdetermined, allowing the same learned structure to support multiple plausible meanings and inviting confirmation bias (Elhage et al., 2022; Gould and Lewontin, 1979; Hogg and Villar, 2024; Locatello et al., 2019). We will discuss these challenges in greater detail later in this Perspective; nonetheless, this post hoc approach has already proved enormously useful for discovering structure in complex data across many fields.

In contrast to Representation Learning, where neural embeddings are interpreted *post hoc*, Representational Measurement Theory (RMT) specifies observation-to-symbol mappings *a priori* through a *measurement procedure* tailored to the scientific question at hand (Hand, 2004; Houle et al., 2011). Houle et al. (2011), for example, consider whether body size influences male guppy attractiveness: this question motivates representing size as snout-to-tail length, measured by aligning fish at a common edge and comparing them against a ruler of repeated standard units. Because these lengths can be ordered and concatenated, their numerical representation forms a ratio scale, preserving relationships such as one fish being longer, or twice as long, as another (Houle et al., 2011). The resulting correspondence between the empirical and numerical systems is a *homomorphism*: a many-to-one mapping that preserves these specified length relationships while ignoring other physical differences, such as color or body shape, sharply restricting the physical meaning of the resulting numbers (Hand, 1996, 2004) and thereby reducing the *post hoc* ambiguity that comes with learned embeddings. Importantly, which relationships can be preserved depends on the researcher's *working ontology*: the provisional set of entities, attributes, and relations treated as "real", or at least relevant, to a particular scientific inquiry (Godfrey-Smith, 2003). These ontologies can themselves be goal-dependent, with different scientific questions motivating

Box 1. How do neural networks construct biological representations?

Deep Comparative Methods typically construct internal numerical representations of biological observations before making predictions. A useful example is the autoencoder. Rather than learning an arbitrary mapping, autoencoders approximate identity functions by compressing high-dimensional observations through a low-dimensional bottleneck. The resulting *latent embedding*—a vector of real-valued numbers—is used to reconstruct the original observation and can serve as a learned biological representation. For example, Adaimé et al. (2025) and Liu et al. (2026) fit Brownian motion models directly to learned embeddings to place difficult taxa on phylogenies and infer phylogenetic relationships.

Many DCMs instead employ the closely related *variational autoencoder* (VAE) (Kingma and Welling, 2019). Rather than predicting latent embeddings directly, VAEs (in their standard Gaussian formulation) learn amortized posterior distributions over the latent variables. Because the latent space is modeled probabilistically, researchers can sample new latent vectors and decode them into previously unseen observations, making VAEs one of the simplest forms of *generative AI*. Despite this probabilistic formulation, VAEs perform the same conceptual task as standard autoencoders: compressing complex biological observations into lower-dimensional latent representations.

This idea is not limited to autoencoders. Supervised classifiers, vision-language models, transformers, and many other DCM architectures likewise construct intermediate vector representations that preserve information relevant to their prediction tasks. Some workflows explicitly train an autoencoder before passing its learned embeddings to a second neural network. Others learn embeddings implicitly while mapping one biological observation to another, such as images to taxonomic labels or tree topologies and tip states to ancestral state reconstructions. Regardless of how they are learned, these latent representations constitute the internal representational language over which Deep Comparative Methods reason.

different ways of representing the same empirical system (Danks, 2015). RMT therefore gains interpretive precision by restricting representations to relationships specified in advance, although this does not guarantee that a measurement is *valid*: validity still depends on whether the measurement procedure actually captures the attribute it is intended to represent (Hand, 2004). In this sense, RMT can be less expressive than Representation Learning, as it can only represent concepts defined within a user's working ontology ahead of time, but more *commensurable* with scientific theory because measurement and theory operate in a shared language of defined entities, relationships, and scales.

In this Perspective, we treat representation learning in DCMs as a measurement process, with *learned representations* as its outputs. This hybrid account brings the expressivity of Representation Learning into tension with RMT's requirement that measurements be meaningful with respect to a scientific ontology. This tension is not unique to our framework. Hogg and Villar (2024), in asking whether machine learning is "good or bad" for the natural sciences, articulate the same basic problem: machine learning generally operates on statistical regularities in observed data, whereas scientists use those observations to reason about latent entities and processes such as traits, evolutionary rates, developmental processes, and selective pressures. We broadly agree with Hogg and Villar (2024)'s diagnosis, but do not view this mismatch as an intrinsic limitation of neural networks. Instead, we treat ontological alignment as a design problem: learned representations need not literally share a scientist's ontology, but they must be constrained such that their statistical structure can be interpreted in terms of the biological entities and relationships relevant to the scientific question. Seen this way, the tension is not simply between learned and manually constructed representations, but between the kinds of relationships machine-learning models readily discover, and the kinds of explanations scientists ultimately seek.

Building on Hogg and Villar (2024)'s epistemological argument, we refine this

distinction by considering the form of reasoning used to justify explanations. *Deductive reasoning* derives necessary consequences from specified premises, *inductive reasoning* generalizes observed relationships beyond the available data, and *abductive reasoning* infers the hypothesis, process, or underlying structure that best explains an observation (Godfrey-Smith, 2003). Many of the central questions in comparative evolutionary biology are fundamentally abductive: researchers seek the evolutionary processes or underlying structures that best explain observed patterns. Much of machine learning, including many approaches to explanation, instead emphasizes inductive generalization (Karpatne et al., 2017, 2024, 2026; Zahavy, 2026). Crucially, successful induction does not require a working ontology: a model can learn relationships that generalize to unseen data without specifying what entities, attributes, or processes those relationships represent. Learned representations can therefore remain scientifically underdetermined even when they generalize well, because multiple interpretations of the same statistical relationship may be compatible with the evidence (Danks, 2015). Abductive explanation, by contrast, requires candidate explanatory structures and therefore a working ontology in which those structures can be expressed and compared (Godfrey-Smith, 2003). We refer to biologically motivated constraints that restrict learned representations toward these candidate explanatory structures as *abductive constraints*.

The broader tension between expanding the space of possible explanations and constraining that space enough to support scientific inference is not unique to machine learning. Although *expressivity* is not common terminology in biology, the underlying problem is longstanding. In discussing genotype–phenotype maps, Houle et al. (2010) argued that discovering simple causal explanations may require first "expanding the universe of possible predictors". We interpret this as an expressivity constraint: if the available data or representations cannot express a plausible biological relationship, no downstream method can recover it. Yet greater dimensionality or apparent precision does not necessarily produce better

representations. In estimating rates of whole-organism evolution, Simpson (1984) used coarse taxonomic categories as proxies for phenotypic differences, arguing that an acknowledged vague approximation could be more useful than a seemingly exact but "spurious average". Houle similarly noted that expanding the number of possible predictors can create a *large-P, small-N* problem, in which models become difficult to estimate and generalize without prior knowledge or regularization (Houle et al., 2010). The same expressivity problem occurs at the level of model construction: models such as BiSSE can restrict the space of possible explanations by attributing diversification differences to an observed binary trait while excluding plausible alternatives such as unmeasured rate heterogeneity (Rabosky and Goldberg, 2015; Tarasov and Uyeda, 2026). In Bayesian terms, an explanation excluded from the model space is effectively given no prior support, so no amount of evidence can recover it (McElreath, 2020). DCMs do not eliminate these modeling assumptions; rather, their greater expressivity relocates many of them into choices about training data, model architecture, objectives, and regularization.

Returning to our primary thesis, we argue that many of the major bottlenecks to discovery in biodiversity science can be resolved, or at least mitigated, by the expressivity of DCMs: their capacity to represent biological entities, attributes, relationships, and evolutionary processes that may be excluded by more restrictive approaches to character and model construction. However, expressivity alone is not sufficient. For DCMs to support scientific explanation, their learned representations must also be constrained in ways that make them commensurable with the researcher's working ontology and the abductive questions being asked. The challenge for DCMs, then, is to combine the expressive flexibility of Representation Learning with the scientifically motivated constraints traditionally imposed through manual measurement and model construction. Meeting this challenge requires understanding the characteristic ways learned representations can diverge from biological reasoning. We turn

next to these explanatory and ontological limitations.

Consequences of the Ontological Mismatch in Learned Representations

Greater expressivity does not necessarily produce greater scientific understanding. A model may accurately predict evolutionary patterns while relying on statistical relationships that are difficult to reconcile with biological theory. The challenge, therefore, is not simply that neural networks are "black boxes," but that the representations supporting their predictions may organize biological variation in ways that differ from the entities, relationships, and processes researchers intend to study. We examine how this mismatch can lead to model failures, what existing explainability methods can reveal about them, and why interpretation still requires biological assumptions and scientific judgment.

Implicit Priors in Training Data Constrain What Models Can Learn

Because Deep Comparative Methods are trained on numerical datasets, they inherit the statistical regularities those datasets contain, whether biological or produced by historical, technological, or sampling artifacts (Mo and Siepel, 2023). Tessler and Little (2026), for example, warn that a model trained on herbarium images could identify plants by reading their specimen labels rather than using their phenotypic features. Similar shortcut learning (Geirhos et al., 2020) occurs in medical imaging: Hill et al. (2024) showed that convolutional neural networks could predict implausible outcomes such as refried-bean or beer consumption from knee radiographs using statistical regularities that researchers could not isolate or explain. Neural networks have no inherent basis for distinguishing such incidental structure from biologically meaningful signal, and can therefore achieve strong predictive performance while relying on relationships with little face validity. They may also absorb unintended but genuine structure: Alper and

Averbuch-Elor (2023), for example, found that vision-language models spontaneously reproduced the kiki–bouba effect, associating pseudowords such as "kitaki" with sharp shapes and "bodubo" with rounded ones without explicit supervision. Such results illustrate that learned representations can capture statistical relationships that are real within the training data but irrelevant to the scientific question at hand.

The same problem arises with simulated training data. Learned representations reflect not only the evolutionary models being approximated, but also choices about parameter ranges, model frequencies, and simulation procedures (Geirhos et al., 2020). Evolutionary models may be only weakly identifiable over parts of parameter space (Cornuault, 2022); for example, Brownian Motion, Ornstein–Uhlenbeck, and accelerating-decelerating change models should not be perfectly distinguishable across all parameter values (Grabowski et al., 2023; Uyeda et al., 2015). If simulations undersample these ambiguous regions or omit plausible alternatives, a neural network may appear to distinguish evolutionary processes because of the simulation design rather than information available in the observations (Landis and Thompson, 2025). Such assumptions are often unavoidable, but they nevertheless act as implicit priors on what the model can learn.

More fundamentally, there is no ontology-independent "correct" biological representation (Danks, 2015). The meaning of a measurement depends on the theoretical context in which it is constructed, so different questions can motivate different representations and working ontologies (Danks, 2015; Houle et al., 2011). Body size, for example, can be represented as a linear trait evolving under geometric Brownian motion or as a log-transformed trait evolving under Brownian motion (Ross, 2014). These representations are predictively equivalent but imply different relationships between the trait and evolutionary process, including whether scale-dependent rate heterogeneity is treated as expected. Unless these question-specific assumptions and desiderata are encoded into a

DCM, the network has little reason to recover the representation that is meaningful to the analyst. Although incorporating these assumptions into DCMs is difficult, researchers are developing ways to guide what the models learn and test whether the resulting representations support the intended inference, as we discuss in Building a Measurement-Theoretic Toolbox for Deep Comparative Methods.

Adversarial Vulnerability Reveals Mismatches Between Human and Machine Concepts of Biology

Adversarial examples are among the most influential discoveries in modern deep learning, revealing a fundamental disconnect between predictive performance and representational robustness (Ghorbani et al., 2019). Although deep neural networks often achieve remarkable accuracy under the independent and identically distributed (i.i.d.) conditions on which they are trained, carefully constructed out-of-distribution perturbations can produce dramatic and highly unintuitive prediction failures (Ghorbani et al., 2019). Zong et al. (2024), for example, showed that simply permuting the order of answer choices in multiple-choice questions—without changing the question, the available answers, or the correct response—can dramatically reduce the performance of both large language models and vision-language models, in some cases to below random chance. Humans immediately recognize these permutations as irrelevant, yet the models do not. Instead, they reveal a dependence on statistical regularities that have little to do with the underlying task. The opposite failure is also possible: a network may ignore changes that should alter its prediction, a problem Jacobsen et al. (2019) call excessive invariance. In their Figure 1, images recognizable as different ImageNet categories nevertheless produce the same prediction. The model treats images from different categories as equivalent, even though distinguishing those categories is the task it was trained to perform. In both cases, the model responds to distinctions that do not matter for the task while

overlooking ones that do.

Distribution Shift Breaks Learned Statistical Relationships

Distribution shift occurs when the statistical relationships present in a model's training data differ from those encountered when the model is applied elsewhere (Moreno-Torres et al., 2012). Machine learning researchers commonly distinguish three forms: *covariate shift*, in which the distribution of observations changes; *prior shift*, in which the relative frequencies of classes or groups change; and *concept shift*, in which the relationship between observations and their interpretation changes (Moreno-Torres et al., 2012). Because DCMs learn representations from these statistical relationships, each form of shift can change what a model learns or how well that representation transfers to a new biological setting.

Covariate shift can arise from biologically irrelevant differences in how specimens are collected or measured. For example, if one museum houses a disproportionate number of specimens from a particular taxonomic group and uses distinctive lighting, backgrounds, preservation methods, or imaging equipment, a network may associate those acquisition conditions with the taxon itself. Prior shift can arise when species occur at different frequencies in the training data and the population being studied. If museum images mostly depict one of two visually similar frog species, for example, a model may favor that species when classifying ambiguous images from a region where the other species is more common.

Concept shift occurs when the interpretation assigned to an observation changes. The same anatomical structure, for example, may be coded differently under alternative hypotheses of homology or different trait boundaries; species labels likewise depend on taxonomic decisions (de Pinna, 1991; Freudenstein, 2005; Pogue and Mickevich, 1990). A DCM trained under one set of definitions may produce consistent predictions that no longer answer the intended biological question under another. Evaluating such a model therefore requires examining

the biological definitions behind its training labels and testing whether its conclusions hold under relevant alternatives.

The Intrinsic Nonidentifiability of Latent Variable Models Prevents Unique Biological Interpretation

A more fundamental limitation of learned representations is that latent variables are generally not uniquely identifiable from observations alone (Locatello et al., 2019). Infinitely many invertible transformations of a latent space can produce the same distribution over observed data while corresponding to very different—and differently interpretable—latent variables (Locatello et al., 2019). A representation that appears cleanly disentangled can therefore be transformed into an equally valid but highly entangled one without changing the observations it generates. Recovering a particular biologically meaningful representation consequently requires assumptions beyond the data, such as supervision, architectural constraints, causal structure, or other inductive biases (Karpatne et al., 2026). This problem is not unique to deep learning: analogous forms of nonidentifiability occur throughout latent-variable modeling, including factor analysis and phylogenetic comparative models (Tarasov and Uyeda, 2026). The important point for DCMs is that biological meaning cannot generally be recovered from latent structure alone; additional constraints are required to privilege one interpretation over other statistically equivalent alternatives.

Existing Explainability Methods are Limited in Scope

Because latent representations are not uniquely determined by the data (Locatello et al., 2019), a natural response is to ask whether post hoc explainability methods can recover the biological meaning that representation learning leaves ambiguous (Marcinkevičs and Vogt, 2023; Rudin, 2019). Researchers working on explaining neural networks for scientific purposes have developed many approaches for making model behavior more interpretable,

including saliency maps (Selvaraju et al., 2020), feature-attribution methods such as SHAP (Lundberg and Lee, 2017), surrogate models such as LIME (Ribeiro et al., 2016), prototype-based explanations (Rudin, 2019), and sparse or disentangled representations (Stevens et al., 2025). Although these methods can be very useful, they answer comparatively narrow questions about model behavior: which inputs influence a prediction (Lundberg and Lee, 2017; Ribeiro et al., 2016), where discriminating information occurs (Prabhushankar et al., 2020), or which recurring features can be isolated within the model's internal representations (Stevens et al., 2025). Their outputs still require biological interpretation, and they do not remove the researcher's responsibility to determine which entities, relationships, scales, and competing explanations are relevant to the scientific question (Houle et al., 2011). The issue is therefore not whether XAI methods provide useful explanations, but what kinds of explanations they provide and how far those explanations can support biological inference. The following sections examine several widely used approaches from this perspective.

One common class of AI explanations is feature attribution, which estimates how particular inputs contribute to a prediction. SHAP (Lundberg and Lee, 2017), for example, uses cooperative game theory to assign feature contributions (Shapley, 1988), whereas LIME (Ribeiro et al., 2016) approximates a model locally with an interpretable linear model. Despite their popularity, SHAP and LIME can misestimate both how important each feature is and which features matter most when compared with exact, logic-based feature attributions (Yu et al., 2023). For image classifiers, attribution is often visualized with saliency maps: Grad-CAM (Selvaraju et al., 2020) highlights regions supporting a prediction, while Contrastive-CAM (Prabhushankar et al., 2020) and Finer-CAM (Zhang et al., 2025) highlight evidence favoring one class over another. Even accurate attributions remain limited because showing where a model looks does not explain what it does with that information (Rudin, 2019). More fundamentally,

biological traits need not correspond to spatially localized structures. Allometric relationships, for example, describe how a measurement of one trait scales with another, but do not strictly depend on the anatomical locations of either trait (Gould, 1974; Lindstedt and Hoppeler, 2023). A saliency map may highlight structures associated with both traits without revealing whether the model represents their absolute measurements, their ratios, or some other difficult-to-describe relationship between them, providing limited information about what the learned trait representations actually mean.

Part-prototype models developed by Rudin and colleagues reduce some of this ambiguity by making comparisons with learned image patches part of the prediction process itself (Rudin, 2019). In mammography, for example, a model compared regions along a breast lesion's margin with prototypes from previously diagnosed cases, allowing physicians to inspect the similarities supporting its prediction (Barnett et al., 2021). Prototype methods may likewise help identify localized biological characters (Manogaran et al., 2025). As with saliency mapping methods, however, interpretation becomes less straightforward when the relevant information concerns properties or relationships rather than the simple presence or absence of a structure. Manogaran et al. (2025), for example, found several prototypes corresponding to the necks of water birds. Because all birds have necks, these prototypes may reflect neck length, color, curvature, or some complex combination of different features. Determining which interpretation is appropriate—or whether any unique discriminating feature exists—requires expert judgment beyond inspecting the prototypes themselves.

Concept Disentanglement and Dictionary Learning Approaches

Some neural interpretability approaches move beyond input features or image regions by identifying patterns in hidden activations that may correspond to higher-level concepts. A classic analogy is the "Jennifer Aniston neuron": Quiroga et al. (2008) identified neurons that responded selectively to

photographs, drawings, and written names of the same person or object, such as references to the actress Jennifer Aniston. Sparse autoencoders similarly decompose dense neural activations into dictionaries of sparsely active features (Bricken et al., 2023), potentially revealing recurring patterns of biological variation that are easier to interpret than distributed representations in dense embeddings (Bricken et al., 2023; Pahuja et al., 2026; Quiroga et al., 2008; Stevens et al., 2025; Templeton et al., 2024). Yet sparsity does not guarantee biological identity. In an analysis of BERT (Devlin et al., 2019), Bolukbasi et al. (2021) found that the sentences maximally activating one neuron followed a coherent song-title pattern in the Quora Question Pairs dataset but a dates-and-historical-events pattern in a Wikipedia-derived question-answering dataset. Because the model and neuron were unchanged, the authors described these conflicting interpretations as an "interpretability illusion" (Bolukbasi et al., 2021). When applied to biodiversity data, then, a feature that appears coherent within one sample of organisms, clades, or environments may acquire a different interpretation when examined in a broader context.

Building a Measurement-Theoretic Toolbox for Deep Comparative Methods

Standard biological measurement involves assigning numbers to attributes of entities in ways that preserve meaningful relationships among those attributes (Houle et al., 2011). In practice, this process assumes that certain entities—organisms, anatomical regions, or developmental modules, for example—are ontologically 'real' and that they in some sense exist independently of the measurements we make of them (Hand, 2004). When we measure fin length, we take as given that there is such a thing as a fin, that it can be individuated from surrounding tissues, and that its geometry is constrained by both evolutionary history and by functional demands in an aquatic, three-dimensional environment. When we then decide how to measure the

properties of a fin for an evolutionary study, such as its color, size, or outline shape, our measurement procedures are conditioned in varying degrees by those predetermined ontological assumptions (Houle et al., 2011).

Neural networks, by contrast, are not inherently committed to any such ontology (Hogg and Villar, 2024). A deep neural network trained only on image data does not "know" what a fin is, nor does it distinguish an adipose fin from a dorsal fin unless such distinctions are statistically useful for its objective. If we attribute "understanding" to a neural network at all, that understanding consists of patterns of association among pixels, or other inputs, and outputs—not in recognition of biologically meaningful entities *per se*. The failure modes discussed above make this distinction consequential: adversarial examples show that networks can attend to irrelevant differences while ignoring changes that matter for a task (Ghorbani et al., 2019), and latent-variable non-identifiability means that equally successful representations can admit different biological interpretations (Locatello et al., 2019). Even when a DCM predicts well, performance alone cannot show whether its representation captures the traits and evolutionary relationships researchers intend to study, exploits artifacts of sampling or imaging, or organizes genuine biological variation in a way that does not support the intended interpretation.

For deep learning–derived traits to be scientifically meaningful—commensurable with evolutionary theory, and interpretable under specific scale types—they must therefore be grounded in abductive constraints. These constraints determine scale, dimensionality, invariance, and the kinds of processes under which the resulting representations can be interpreted (Houle et al., 2011). Without these constraints, interpreting learned representations risks amplifying confirmation bias: researchers may favor the biological explanation they expected to find even when other interpretations fit the same patterns equally well (Hogg and Villar, 2024).

How, then, can researchers build abductive constraints into DCMs and assess

whether the resulting representations are meaningful for evolutionary inference? We now turn to a practical toolbox for guiding what DCMs learn and testing whether their outputs support the biological interpretations researchers intend.

Abductive Biases in Neural Networks

Architecture. — One of the primary ways to introduce abductive knowledge into a neural network is through architecture. Architectural choices constrain the kinds of relationships a model can easily represent and the forms of variation it tends to ignore (Karpatne et al., 2024). In this sense, architecture encodes assumptions about *invariance*. These assumptions are also measurement-theoretic: a measurement scale is partly defined by the transformations under which its meaning is preserved (Houle et al., 2011), and biologically informed architectures can impose similar restrictions on learned characters. Architectural constraints therefore do more than improve generalization. They help determine whether the resulting representations are admissible for particular kinds of evolutionary inference.

Convolutional neural networks (CNNs) provide a simple example. A model trained mostly on photographs of airplanes in flight may come to expect them near the top of an image and struggle with one parked on the ground. By applying the same small filter, or "sliding window," across the image, a CNN can reduce this dependence on absolute position (Lecun et al., 1998). The same detector can recognize recurring edges, contours, and textures whether they appear near the top or bottom of the frame. Spatial invariance is not always desirable for biological images, where a structure's position relative to other structures can be part of what defines it. Adipose fins, for example, are identified partly by their position between the dorsal and caudal fins, despite variation in their anatomy and independent origins in different fish lineages (Stewart et al., 2014). Differences in appearance may allow a well-trained network to distinguish adipose fins from other median fins, but accurate classification alone would not

show that it specifically learned the spatial relationship relevant to their identification. Deciding which spatial relationships to preserve is therefore a modeling choice about what counts as the same character.

Other architectures impose different kinds of structure. Sequence models such as recurrent neural networks (RNNs) (Elman, 1990) and long short-term memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997) preserve ordered relationships through time or across a sequence. Axial transformers can respect the organization of multiple-sequence alignments (Rao et al., 2021), while DeepSDF represents morphology as a continuous function over three-dimensional space (Park et al., 2019). These choices make some biological relationships easier to learn than others: temporal order is built into RNNs and LSTMs (Elman, 1990; Hochreiter and Schmidhuber, 1997), long-range dependence into transformers (Rao et al., 2021), and geometric continuity into DeepSDF (Park et al., 2019). Architectures can also incorporate taxonomic, phylogenetic, or ontogenetic organization more explicitly. BioCLIP aligns images with hierarchical taxonomic language (Gu et al., 2025; Stevens et al., 2024); Elhamod et al. (2023) and Khurana et al. (2025) structure learned features across phylogenetic timescales; Manogaran et al. (2025) learn visual prototypes associated with internal nodes; and Ubbens et al. (2020) combined a CNN with an LSTM so that plant images were represented across ordered stages of development. Their shared point is architectural: each model narrows the representational problem by deciding in advance which relationships among observations should be preserved.

Architecture therefore determines which relationships a model can readily preserve and which differences it tends to erase. For Deep Comparative Methods, those choices help decide whether the learned representation matches the biological scale and structure needed for the intended inference.

Loss Function. — Another way biologists can incorporate abductive knowledge into neural networks is through their choice of loss function (Karpatne et al., 2024). The loss function is the objective function optimized during training: it determines what the model is rewarded or penalized for and therefore which of the representations allowed by the architecture are optimal (Zhang et al., 2023). In Deep Comparative Methods, loss functions can be used to shape the scale, dimensionality, geometry, and phylogenetic structure of learned traits.

Generic regularization terms provide the simplest examples. An L_1 penalty encourages many coefficients or latent activations to remain near zero and corresponds to a Laplace prior, while an L_2 penalty discourages large values and corresponds to a zero-centered Gaussian prior whose variance depends on the strength of the penalty (Murphy, 2022). These constraints can control how much information a learned representation uses. Pahuja et al. (2026), for example, used a sparsity penalty when training an autoencoder on insect-image features; changing the penalty's strength altered how many latent units were active and how well those units supported downstream trait descriptions. This is not just a technical concern: the number and scale of traits included in an analysis can substantially affect downstream phylogenetic and diversification inferences (Beaulieu and O'Meara, 2016; Khakurel et al., 2024; Simões et al., 2018).

Loss functions can also make learned representations commensurable with quantities that already have meaning in comparative biology. Several examples introduced earlier did this by encouraging distances among embeddings to reflect patristic distance (Adaïmé et al., 2024), genetic similarity (Hofmann et al., 2024), or taxonomic affinity (Gu et al., 2025; Stevens et al., 2024), while others imposed an expected phylogenetic covariance structure or a target level of phylogenetic signal (Hunt et al., 2025; Lau and Choi, 2026). The important point here is not the details of each model, but the shared role of the objective function: it turns an otherwise flexible embedding into a representation whose distances or covariances can be interpreted in relation to evolutionary theory, when properly

validated.

Loss functions are especially flexible because several objectives can be added, weighted, and optimized within the same general architecture (Karpatne et al., 2026; Zhang et al., 2023). A model can be trained at once to reconstruct its inputs, compress their representation, preserve phylogenetic structure, or separate biologically meaningful groups. Losses therefore do more than improve prediction. They help determine what kind of trait space the model constructs and which forms of evolutionary inference that space can reasonably support.

Training Data. — Training data can be used to control what a neural network treats as meaningful variation. Models often pick up on incidental regularities in their training sets or ignore distinctions that matter for the task (Geirhos et al., 2020; Hill et al., 2024; Mo and Siepel, 2023; Tessler and Little, 2026). As a result, characters and evolutionary processes may be constructed based on causally meaningless features in the data. Researchers can address this problem through dataset design. When using simulated data, they can sample broadly across plausible models and parameter values rather than relying on a narrow simulation setup (Landis and Thompson, 2025). For image data, simple augmentations—such as removing specimen labels, rotating or flipping images, or varying lighting and backgrounds—can break spurious correlations while preserving the biological features of interest (Zhang et al., 2023). These choices help ensure that the model focuses on the right signals.

Such nuisance variation can be hard to identify because confounding signals may be distributed across complex patterns. Domain-adaptation methods address this problem by encouraging models to retain information relevant to the inferential task while becoming less sensitive to differences between training and target domains (Farahani et al., 2021). In Mo and Siepel (2023), for example, researchers added domain adaptation to neural networks trained on population-genetic simulations to infer selection and recombination parameters.

An auxiliary classifier attempted to distinguish source-domain inputs from target-domain inputs, while a gradient-reversal layer trained the shared feature extractor to prevent that distinction. This improved inference under several simulated forms of distribution shift (Moreno-Torres et al., 2012) and allowed the selection model to account partly for differences between simulated and empirically inferred genealogies.

The same logic can be used not only to remove nuisance variation, but to direct a model toward the kinds of biological variation relevant to a particular question. Ezray et al. (2019), for example, did not analyze photographs of bumblebees directly. Instead, they transferred the color patterns reported in a field guide onto a standardized bee template. This removed variation in body size, outline, perspective, and lighting, leaving the network to compare the segmental color patterns relevant to Müllerian mimicry. Tsutsumi et al. (2023) made nearly the opposite choice because they were interested specifically in mandibular shape. They aligned primate mandibles, projected each specimen from three standardized directions, normalized size, and represented the projections as monochrome images, making variation in outline and overall form the dominant information available to the model. More generally, cropping organisms from museum images can prevent backgrounds, labels, rulers, and other collection artifacts from dominating learned representations (Tessler and Little, 2026). In each case, pre-processing narrows the model's otherwise broad expressivity by deciding in advance which forms of variation are available to be learned.

This use of training data parallels familiar practices in geometric morphometrics. Generalized Procrustes Analysis makes shapes invariant to translation, rotation, and scale so that subsequent analyses focus on variation in relative form (Mitteroecker and Schaefer, 2022). Dataset design and augmentation serve a similar role in Deep Comparative Methods, whether imposed directly or learned through domain adaptation, making models

invariant to causally meaningless features and focusing them more strongly on the features that matter for the researchers' purposes, such as shape or color patterns (Karpatne et al., 2024).

Optimization Process. — Optimization processes often overlap closely with loss functions because they determine when and how strongly each objective influences training. In PhyloSDF (Lau and Choi, 2026), for example, the phylogenetic consistency loss was applied only once every ten training batches; reconstruction and $L2$ regularization governed the intervening updates. The authors introduced this schedule partly to reduce computational cost, but it also meant that phylogenetic structure shaped the latent space without controlling every step of its formation.

One optimization strategy worth distinguishing from ordinary loss design is human-in-the-loop training (Mosqueira-Rey et al., 2023). Here, researchers repeatedly inspect model outputs and provide feedback that changes later rounds of annotation, fine-tuning, or prediction (Mosqueira-Rey et al., 2023). This allows them to guide a DCM toward distinctions that matter for their scientific purposes but that may be difficult to express as a fixed loss, architecture, or dataset assembled in advance. Van Dam and Štarhová Serbina (2026)'s Descriptron provides a broad example. Taxonomists choose the anatomical structures of interest, accept or reject masks proposed by SAM 2 (Ravi et al., 2025), correct masks and landmarks, assign biologically meaningful labels, and use those corrected annotations to fine-tune an instance-segmentation model. The resulting masks, measurements, and taxonomic labels are then supplied to a vision-language model, whose draft descriptions are again reviewed for omitted or incorrectly described characters (Van Dam and Štarhová Serbina, 2026). Rather than asking the model to discover an unrestricted vocabulary of visual differences, this iterative process makes it attend to structures and terminology already meaningful within the taxonomic system being studied.

Intermediate neural representations usually occupy high-dimensional spaces that are difficult for researchers to inspect or correct directly, limiting human-in-the-loop training mostly to annotation and review around the model rather than intervention within the embedding itself. Vision-language models offer one possible bridge by expressing learned differences in natural language. Vision-language approaches such as BioCAP (Zhang et al., 2026) and LLM-Mutate (Chiquier et al., 2024) attempt to describe visual traits directly from images, with the latter using an evolutionary search procedure to refine discrete natural-language attributes until they discriminate among image classes. At the same time, formal trait ontologies have increasingly been used to extract character data from free-text literature (Dahdul et al., 2018, 2010), link trait records back to specimen images and other primary evidence (Stack, 2025), and identify logical inconsistencies among character definitions (Dececchi et al., 2015). Related work in other fields has also begun using language models to curate ontologies from existing literature (Soares and Wassermann, 2025). Combining these approaches with human review could make practical what ontology-based systems have long aimed to achieve but struggled to scale because of the labor required to curate characters and relationships by hand (Deans et al., 2015): traits could be proposed from images and literature, checked against domain standards, and assembled into interpretable, reproducible representations for downstream evolutionary analysis.

Assessing and Interpreting Neural Network Behavior and Predictions

Much work on interpretability in machine learning focuses on explaining models after they have been trained (Rudin, 2019). Comparative biologists, however, use models for a much broader range of scientific purposes. The appropriate strategy for interpreting a Deep Comparative Method therefore depends not only on the model itself, but also on the scientific question being asked. As Houle et al. (2011) argue, the meaning of a representation depends on

its theoretical context. More broadly, different biological models are constructed to answer different kinds of scientific questions.

We distinguish two broad classes of explanatory goals for Deep Comparative Methods. First, researchers often wish to understand the behavior of a method itself. Under what conditions does it recover the correct evolutionary process, and how robust is it to hidden states, model mis-specification, measurement error, or distribution shift? These questions may be productively addressed through simulation studies, benchmark datasets, and controlled perturbation experiments. Second, researchers often seek to understand a specific fitted model. Here the goal is to evaluate the representations, predictions, and uncertainties of a particular analysis using approaches such as cross-validation, posterior predictive checks, generative validation, and related methods.

Simulation Studies and Benchmarks. — Simulation studies and benchmark datasets primarily address the first of these goals: understanding the behavior of a Deep Comparative Method itself. The central question is straightforward: if our assumptions about the generative process are correct, does the method behave as intended? For model-constructing Deep Comparative Methods, this question can often be addressed by generating synthetic datasets under known evolutionary models. More flexible simulation frameworks further allow researchers to investigate increasingly complex generative processes that may not admit tractable likelihood functions.

For tabular comparative data, packages such as diversitree (FitzJohn, 2012), RevBayes (Höhna et al., 2016), and treats (Guillerme, 2024) can simulate trait evolution and diversification under a range of evolutionary models. For example, diversitree implements a variety of state-dependent diversification models, while RevBayes provides a probabilistic programming framework for specifying complex hierarchical evolutionary models and integrates naturally into more complex R workflows through RevGadgets (Tribble et al., 2022) and Revticulate

(Charpentier and Wright, 2022). Likewise, treats provides a highly flexible framework for jointly simulating phylogenetic trees and interacting trait processes under user-specified diversification and evolutionary dynamics. More recently, Landis and Thompson (2025) developed phyddle to facilitate likelihood-free Deep Comparative Methods by decoupling the simulation engine from the learning algorithm, allowing arbitrary simulators written in R, RevBayes, Python, and other programming languages to be incorporated directly into a common deep learning workflow. At finer evolutionary scales, forward simulators such as SLiM (Haller and Messer, 2019) provide similarly flexible environments for population genetic simulations and have already been incorporated into Deep Comparative Method pipelines for tasks such as estimating population trees directly from multiple sequence alignments (Nesterenko et al., 2025) as well as combining forward genetic simulations with synthetic image generation of *Arabidopsis thaliana* to estimate quantitative trait loci directly from learned image embeddings (Ubbens et al., 2020).

For raw biological observations, simulation remains a substantially newer but rapidly growing area. As Deep Comparative Methods increasingly learn directly from images, videos, three-dimensional meshes, and other high-dimensional observations, benchmarking these methods requires synthetic datasets that faithfully capture the biological and physical processes generating those observations. TraitBlender (Charpentier et al., 2026) generates synthetic three-dimensional organisms and museum-specimen style images from both empirical and theoretical morphospaces, while replicAnt (Plum et al., 2023) uses the Unreal Engine to generate realistic images and videos of insect behavior under controlled environmental conditions. For plant morphology, a large ecosystem of procedural simulators based on Lindenmayer systems (L-systems) has been developed, with The Algorithmic Beauty of Plants (Prusinkiewicz and Lindenmayer, 1990) providing a comprehensive introduction to these methods, which have already been incorporated into Deep Comparative Method pipelines

for latent-space phenotyping of *Arabidopsis thaliana* (Ubbens et al., 2020, 2018). Similar procedural approaches have also been developed for other organismal groups, including algorithmic models of molluscan shell growth such as those described in (Fowler et al., 1992), which generate highly flexible theoretical morphospaces from biologically interpretable generative parameters.

As noted by Boyko and Rabosky (2026), morphospaces can be broadly divided into theory-informed and empirical approaches. Theory-informed morphospaces specify generative or developmental rules *a priori*, whereas empirical morphospaces infer geometric structure directly from observations using methods such as PCA or manifold learning. These approaches are naturally complementary within Deep Comparative Methods. Theory-informed morphospaces provide known generative processes against which automated character construction methods can be benchmarked, while Deep Comparative Methods can construct empirical morphospaces directly from high-dimensional observations. TraitBlender (Charpentier et al., 2026) currently incorporates two such theoretical morphospaces: a biologically motivated molluscan shell model extending the morphospaces of (Contreras-Figueroa and Aragón, 2023) and (Okabe and Yoshimura, 2017), and a simple synthetic morphospace consisting of evolving grids of circles whose diameters and opacities are treated as heritable characters. Likewise, Lindenmayer systems (L-systems) (Lindenmayer, 1968) provide a general mathematical formalism for specifying developmental programs and can themselves be viewed as theoretical morphospaces. Although most commonly applied to individual plants (Boudon et al., 2012; Prusinkiewicz and Lindenmayer, 1990), related approaches have been developed for mycorrhizal fungi (Schnepf et al., 2016), Ediacaran rangeomorph fronds (Hoyal Cuthill and Conway Morris, 2014), and other organismal groups (Prusinkiewicz, 2025).

More broadly, theoretical morphology has produced generative models for a wide range of biological systems which have the potential to be used in simulation pipelines for assessing DCMs. These include models of early

embryonic development (Cano-Fernández et al., 2024), reaction-diffusion systems for developmental pattern formation (Kondo, 2009; Meinhardt and Klingler, 1987), and structure-specific developmental models, such as the Voronoi-tessellation-based model of insect wing venation developed by Hoffmann et al. (2018). Similar developmental models have also been proposed for structures such as beaks, horns, and claws (Evans et al., 2021). Other tools, such as MorphoWeave (Porto, 2026), define empirical morphospaces from dense landmarks rather than theoretical models of development, but can still be used for sampling synthetic datasets for evaluating DCMs. More recently, Merchant et al. (2026) introduced Proto, a domain-specific programming language for representing protein structures and their assembly, illustrating how expressive specification languages can provide compact, generative descriptions of biological form. Together with the rapid development of phenotypic trait ontologies, these advances suggest the possibility of general-purpose morphospace specification languages combining the generative flexibility of probabilistic programming languages like RevBayes (Höhna et al., 2016) with the explicit biological semantics of ontology-based systems such as Phenoscape (Dahdul et al., 2010; Montanaro et al., 2024). Such languages could provide a common framework for describing, generating, and sampling raw biological observations across arbitrary modalities while integrating naturally with existing simulation environments.

While simulation studies provide control over the underlying generative process, synthetic datasets inevitably differ from empirical observations (Mo and Siepel, 2023). Carefully curated benchmark datasets can address this limitation by testing DCMs on standardized data acquired under real conditions, following long-standing practice in machine learning (Russakovsky et al., 2015), phylogenetics (Linder et al., 2010), and computational biology more broadly (Peters et al., 2018). Several existing datasets provide starting points for this kind of evaluation. Rove-Tree-11 (Hunt and Pedersen, 2022) benchmarks phylogenetic inference from museum specimen images, while Fish-Vista (Mehrab et al., 2025)

supports species classification, trait identification, and trait segmentation. TreeOfLife-200M (Gu et al., 2026) provides taxonomically curated images for training and evaluating biological foundation models such as BioCLIP 2 (Gu et al., 2025), and BIOSCAN-5M (Gharaee et al., 2024) pairs images with DNA barcodes from millions of insect specimens for multimodal benchmarking. As DCMs and other AI methods become more widely used in biodiversity science, shared benchmarks will become increasingly important for comparing what these models can recover from empirical data.

Cross-Validation Approaches. — Cross-validation is often used to assess how well machine learning models generalize to unseen data (Zhang et al., 2023). For Deep Comparative Methods, however, generalization is not necessarily the primary scientific objective. Rather, held-out datasets are most informative when they probe failure modes that are directly relevant to the biological interpretation of a model. In this sense, cross-validation is best viewed not simply as a measure of predictive performance, but as a tool for evaluating whether learned representations remain meaningful under biologically relevant distribution shifts.

For example, Adaïmé et al. (2024) evaluated their model for phylogenetic placement of fossil pollen by withholding extant taxa with known phylogenetic positions and testing whether the model correctly recovered their placement from image embeddings alone. Because the intended application is to infer the placement of taxa whose evolutionary relationships are unknown, this validation strategy closely mirrors the scientific problem the model was designed to solve. Likewise, Lürig et al. (2025) found that a direction in image embedding space associated with aposematic coloration and toxicity in butterflies also explained variation in moths, despite the model never being trained on moths. This transfer from butterflies to the sampled moth taxa provides evidence that the learned representation captures a broader biological pattern rather than a confounding signal restricted to butterflies. In both cases, the value of cross-validation lies not

simply in demonstrating better generalization, but in testing whether the learned representations fail in ways that would undermine their intended biological interpretation.

Generative Modeling. — Beyond simulation and cross-validation, generative models provide an additional mechanism for interrogating learned representations (Dinnage and Kleineberg, 2025). Comparative biology has long used posterior predictive simulations to evaluate whether fitted evolutionary models generate data resembling empirical observations and, more importantly, whether they support biologically plausible inferences (Brown, 2014; Mulvey et al., 2025). Because many Deep Comparative Methods, including variational autoencoders (Kingma and Welling, 2019) and related latent variable models (Dinnage and Kleineberg, 2025; Lau and Choi, 2026; Tsutsumi et al., 2023), learn amortized approximations to posterior distributions, these representations can be interrogated in analogous ways through conditional sampling (Kingma and Welling, 2019; Marcinkevičs and Vogt, 2023). Rather than evaluating only reconstruction quality or predictive accuracy, researchers can ask whether samples drawn from the learned latent space reproduce biologically meaningful structure, reveal implausible phenotypes, or preserve relationships expected under existing biological theory. Dinnage and Kleineberg (2025) provide one example of this strategy by conditioning a generative model of bird bill morphology on trophic niche and sampling synthetic bills to visualize the morphological distributions learned by the model. Conditional sampling could also expose anatomical violations, such as a generated fish with its dorsal fin positioned behind its caudal fin, or subtler departures from developmental and functional constraints that warrant closer examination of the learned representation.

Similar principles apply even when models are not explicitly generative. Learned embedding spaces can be compared directly against independently

measured biological traits or expert annotations to evaluate whether they capture biologically meaningful structure (Elhamod et al., 2023; Gu et al., 2025). Lürig et al. (2025) used LDA to identify a direction in image embedding space associated with aposematic coloration in butterflies and showed that it generalized to moths, while image statistics extracted independently from the learned embeddings were consistent with expert knowledge of warning coloration. Likewise, Elhamod et al. (2023) evaluated the learned representations in Phylo-NN by comparing distances in embedding space with both phylogenetic distances and distances derived from expert-measured morphological traits, demonstrating strong agreement between learned and independently measured patterns of organismal similarity. More generally, these approaches illustrate that the interpretability of Deep Comparative Methods can be assessed by comparing learned representations with independent biological knowledge, rather than relying solely on predictive performance or feature attribution methods.

Conclusions

Deep Comparative Methods offer comparative biologists a way to expand the space of candidate traits, relationships, and evolutionary processes beyond those that can be specified by hand in advance. This expressivity comes with a characteristic cost. Neural networks learn from statistical regularities in their training data rather than from the entities and processes biologists use to formulate explanations, so the representations they construct can rely on incidental features, fail under distribution shift, or admit many statistically equivalent but biologically distinct interpretations. Post hoc explainability methods can describe how a fitted model behaves, but they cannot by themselves establish which biological hypothesis best explains the observations.

We have argued that this mismatch is a design problem rather than an intrinsic limitation. Architectures, loss functions, training data, and optimization procedures each offer points at which researchers can impose abductive

constraints: biologically motivated restrictions on the invariances, scales, dimensionality, and structure of learned representations that make them commensurable with a working ontology. Simulation studies, benchmarks, biologically targeted cross-validation, and generative checks then provide ways to test whether those constraints succeeded. Making these choices explicit, and reporting them as carefully as we report the assumptions of conventional comparative models, will be essential if DCMs are to mature from a collection of promising applications into a coherent field of methodological research.

The promise of Deep Comparative Methods is not that neural networks eliminate the need to define biological meaning, but that they expand the range of candidate measurements and explanations whose meaning can be made explicit and tested.

Acknowledgments

C.C. and J.C.U. were supported by the National Science Foundation (OAC 2118240) Imageomics Institute award. C.C. was additionally supported by the National Science Foundation Graduate Research Fellowship Program. A.P. was supported by the National Science Foundation (DEB-2529932 and EAR-2422839). We wish to especially thank all members of the Imageomics Institute for their thoughtful feedback and collaborations. The conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

OpenAI's GPT-5.6 Pro and Anthropic's Claude Fable 5 were used for language editing. All suggested text was reviewed by the authors, and the corresponding author takes responsibility for the final content.

Table 1: Definitions used throughout this Perspective.

Term	Definition	Key references
Modes of inference		
Deductive reasoning	Inference in which the conclusion necessarily follows from the premises. If the premises are true and the argument is valid, the conclusion must also be true.	Godfrey-Smith (2003)
Inductive reasoning	Inference in which observations of particular cases support a broader generalization extending beyond the observed data. Unlike deduction, inductive conclusions are not logically guaranteed.	Godfrey-Smith (2003)
Abductive reasoning	Inference from observations to the hypothesis, process, or underlying structure that would best explain them. Also called *inference to the best explanation*.	Godfrey-Smith (2003)
Data, models, and machine learning		
Raw data	Observations in the form in which they were initially recorded, such as images, videos, or free text, before transformations or annotations used in analysis.	LeCun et al. (2015)
Phylogenetic comparative methods (PCMs)	Statistical methods that use information about shared evolutionary history to study trait evolution and relationships among biological variables across taxa.	Cornwallis and Griffin (2024); Felsenstein (1985)

Continued on next page

Table 1 continued

Term	Definition	Key references
Knowledge-guided machine learning (KGML)	Machine-learning approaches that use scientific knowledge and data as complementary sources of supervision to improve generalizability, scientific consistency, or explainability. Also called *theory-guided data science*.	Karpatne et al. (2017, 2024, 2026)
Representation and interpretation		
Representation learning	Machine-learning approaches in which a model learns transformations or features of the input data that are useful for one or more tasks, rather than relying entirely on predefined variables.	Bengio et al. (2013)
Embedding / embedding space	An *embedding* is a numerical representation of an observation as a point in a finite-dimensional space. The *embedding space* is the space containing these points, where their positions and distances may reflect relationships learned by the model.	Bengio et al. (2013); Zhang et al. (2023)
Monosemantic	Describes a learned component that consistently corresponds to a single interpretable feature or concept within the context in which it is evaluated.	Templeton et al. (2024)
Polysemantic	Describes a learned component that responds to multiple distinct features or concepts, which may be unrelated from the perspective of a human interpreter.	Templeton et al. (2024)

Continued on next page

Table 1 continued

Term	Definition	Key references
MEASUREMENT AND ONTOLOGY		
Representational measurement theory	A formal account of measurement in which an empirical system is mapped onto a numerical system while preserving the relationships relevant to the scientific question.	Hand (2004); Houle et al. (2011)
Ontology (philosophy)	An account of the entities, properties, and relationships taken to exist within a domain. A researcher's *working ontology*, as used in this Perspective, is the provisional ontology used to formulate and interpret a particular scientific investigation.	Danks (2015)
Ontology (computer science)	A formal, machine-readable specification of the entities, categories, properties, and relationships recognized within a domain.	Dahdul et al. (2018, 2010); Deans et al. (2015); Dececchi et al. (2015)
Validity (statistics)	The extent to which available evidence supports the interpretation and use of a measurement, model output, or inference for its intended purpose.	Hand (2004)
Validity (representational measurement theory)	The extent to which a numerical representation preserves the empirical relationships that give the represented quantities their scientific meaning.	Hand (2004); Houle et al. (2011)

Continued on next page

Table 1 continued

Term	Definition	Key references
Empirical relational system (ERS)	A set of empirical entities together with the relationships or operations among them that a measurement procedure is intended to represent.	Hand (2004); Houle et al. (2011)
Numerical relational system (NRS)	A set of numerical values together with the numerical relationships or operations used to represent the structure of an empirical relational system.	Hand (2004); Houle et al. (2011)
Invariance	The property that an interpretation or conclusion remains unchanged under transformations of a representation that preserve its empirical meaning.	Hand (2004); Karpatne et al. (2024)
Identifiability	Concerns whether the distribution of observable data uniquely determines the model parameters or latent structure of interest, given the model's assumptions.	Cornuault (2022); Locatello et al. (2019)

Table 2: Examples of Deep Comparative Methods (DCM) studies.

Application	Representative studies
CHARACTER CONSTRUCTION	
Argue for data-driven approaches to characterizing the structure of phenotypic spaces.	Boyko and Rabosky (2026)

Continued on next page

Table 2 continued

Application	Representative studies
Use phylogenetic or taxonomic knowledge to constrain learned phenotypic representations toward taxonomically or phylogenetically structured variation.	Elhamod et al. (2023); Gu et al. (2025); Hunt et al. (2025); Khurana et al. (2025); Lau and Choi (2026); Manogaran et al. (2025); Stevens et al. (2024)
Use learned phenotypic representations or similarity measures to test ecological or evolutionary hypotheses with traits that are difficult to define *a priori*.	Cuthill et al. (2019); de Solan et al. (2020); Dinnage and Kleineberg (2025); Fan et al. (2022); Lürig et al. (2025)
Learn alternative phenotypic character systems directly from complex data rather than restricting analyses to predefined measurements.	Feldmann et al. (2021); Hartmann et al. (2026); Lucny (2020); Stevens et al. (2025); Tsutsumi et al. (2023); Ubbens et al. (2020); Zhao et al. (2023)
Bridge phenotypic and molecular data by using learned morphological characters in phylogenetic and comparative inference.	Adaïmé et al. (2024); Adaimé et al. (2025); Hofmann et al. (2024)
Use generative models to reveal, visualize, and manipulate complex biological variation that is difficult to represent with predefined traits.	Bourou et al. (2024); Lamiable et al. (2023); Ziegler et al. (2023)
Model construction	
Provide general frameworks for simulation-trained inference under user-defined phylogenetic models, including models without tractable likelihoods.	Landis and Thompson (2025)

Continued on next page

Table 2 continued

Application	Representative studies
Learn mappings from phylogenetic data to evolutionary parameters without requiring conventional parameter-estimation procedures at inference.	Bokma (2006); Thompson et al. (2024)
Treat ancestral-state reconstruction as a learned prediction problem, including under complex evolutionary models.	Nagel and Landis (2026)
Use learned classifiers to choose among alternative evolutionary models directly from data.	Jones (2026); Perez and Gascuel (2026)
Learn flexible components or relationships within evolutionary and ecological models that are difficult to specify parametrically.	Bourhis et al. (2023); Silvestro et al. (2024)
Use deep learning to replace or augment conventional components of phylogenetic reconstruction, including topology, branch-length, and evolutionary-distance estimation.	Nesterenko et al. (2025); Smith and Hahn (2023); Suvorov et al. (2020); Suvorov and Schrider (2024)

References

Adaïmé, M.-É., S. Kong, and S. W. Punyasena. 2024. Deep learning approaches to the phylogenetic placement of extinct pollen morphotypes. PNAS nexus 3:pgad419.

Adaimé, M., M. A. Urban, S. Kong, C. Jaramillo, and S. W. Punyasena. 2025. Pollen morphology, deep learning, phylogenetics, and the evolution of environmental adaptations in *podocarpus*. New Phytologist 247:1460–1473.

Alper, M., and H. Averbuch-Elor. 2023. Kiki or bouba? sound symbolism in vision-and-language models. *In* Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23. Curran Associates Inc., Red Hook, NY, USA.

Barnett, A. J., F. R. Schwartz, C. Tao, C. Chen, Y. Ren, J. Y. Lo, and C. Rudin. 2021. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. Nature Machine Intelligence 3:1061–1070.

Beaulieu, J. M., and B. C. O'Meara. 2016. Detecting hidden diversification shifts in models of trait-dependent speciation and extinction. Systematic Biology 65:583–601.

Bengio, Y., A. Courville, and P. Vincent. 2013. Representation learning: a review and new perspectives. IEEE transactions on pattern analysis and machine intelligence 35:1798–1828.

Blomberg, S. P., T. Garland JR., and A. R. Ives. 2003. Testing for phylogenetic signal in comparative data: Behavioral traits are more labile. Evolution 57:717–745.

Bokma, F. 2006. Artificial neural networks can learn to estimate extinction rates from molecular phylogenies. Journal of Theoretical Biology 243:449–454.

Bolukbasi, T., A. Pearce, A. Yuan, A. Coenen, E. Reif, F. Viégas, and M. Wattenberg. 2021. An interpretability illusion for bert.

Borowiec, M. L., R. B. Dikow, P. B. Frandsen, A. McKeeken, G. Valentini, and A. E. White. 2022. Deep learning as a tool for ecology and evolution. Methods in Ecology and Evolution 13:1640–1660.

Boudon, F., C. Pradal, T. Cokelaer, P. Prusinkiewicz, and C. Godin. 2012. L-py: An l-system simulation framework for modeling plant architecture development based on a dynamic language. Frontiers in Plant Science Volume 3 - 2012.

Bourhis, Y., J. R. Bell, C. R. Shortall, W. E. Kunin, and A. E. Milne. 2023. Explainable neural networks for trait-based multispecies distribution modelling—a case study with butterflies and moths. Methods in Ecology and Evolution 14:1531–1542.

Bourou, A., T. Boyer, M. Gheisari, K. Daupin, V. Dubreuil, A. De Thonel, V. Mezger, and A. Genovesio. 2024. PhenDiff: Revealing Subtle Phenotypes with Diffusion Models in Real Images . *In* proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. Vol. lncs 15003. Springer Nature Switzerland.

Boyko, J. D., and D. L. Rabosky. 2026. The geometry of macroevolution: Phenotypic evolution on non-euclidean manifolds. The American Naturalist 207:751–762.

Bricken, T., A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread Https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Brown, J. M. 2014. Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. Systematic Biology 63:334–348.

Caetano, D. S., and D. J. Hearn. 2026. Probabilistic principal component analysis for phylogenetic comparative studies. Evolution 80:902–911.

Cano-Fernández, H., M. Brun-Usan, T. Tissot, and I. Salazar-Ciudad. 2024. A morphospace exploration using a general model of development reveals a basic set of morphologies for early animal development and evolution. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution 344:45–58.

Charpentier, C., T. M. Linscott, J. Bradley, E. Campolongo, M. Thompson, A. Porto, and J. C. Uyeda. 2026. Traitblender.

Charpentier, C. P., and A. M. Wright. 2022. Revticulate: An r framework for interaction with revbayes. Methods in Ecology and Evolution 13:1177–1184.

Chiquier, M., U. Mall, and C. Vondrick. 2024. Evolving interpretable visual classifiers with large language models. Pages 183–201 *in* Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXIV. Vol. 15122 of *Lecture Notes in Computer Science*. Springer.

Collyer, M. L., and D. C. Adams. 2021. Phylogenetically aligned component analysis. Methods in Ecology and Evolution 12:359–372.

Contreras-Figueroa, G., and J. L. Aragón. 2023. A mathematical model for mollusc shells based on parametric surfaces and the construction of theoretical morphospaces. Diversity 15.

Cooper, R. B., J. T. Flannery-Sutherland, and D. Silvestro. 2024. Deepdive: estimating global biodiversity patterns through time using deep learning. Nature Communications 15:4199.

Cornuault, J. 2022. Bayesian analyses of comparative data with the ornstein-uhlenbeck model: Potential pitfalls. Systematic biology 71:1524–1540.

Cornwallis, C. K., and A. S. Griffin. 2024. A guided tour of phylogenetic comparative methods for studying trait evolution. Annual Review of Ecology, Evolution, and Systematics 55:181–204.

Csilléry, K., M. G. Blum, O. E. Gaggiotti, and O. François. 2010. Approximate bayesian computation (abc) in practice. Trends in Ecology & Evolution 25:410–418.

Cuthill, J. F. H., N. Guttenberg, S. Ledger, R. Crowther, and B. Huertas. 2019. Deep learning on butterfly phenotypes tests evolution's oldest mathematical model. Science Advances 5:eaaw4967.

Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems 2:303–314.

Dahdul, W., P. Manda, H. Cui, J. P. Balhoff, T. A. Dececchi, N. Ibrahim, H. Lapp, T. Vision, and P. M. Mabee. 2018. Annotation of phenotypes using ontologies: a gold standard for the training and evaluation of natural language processing systems. Database : the journal of biological databases and curation 2018.

Dahdul, W. M., J. P. Balhoff, J. Engeman, T. Grande, E. J. Hilton, C. Kothari, H. Lapp, J. G. Lundberg, P. E. Midford, T. J. Vision, M. Westerfield, and P. M. Mabee. 2010. Evolutionary characters, phenotypes and ontologies: curating data from the systematic biology literature. PloS one 5:e10708.

Danks, D. 2015. Goal-dependence in (scientific) ontology. Synthese 192:3601–3616.

de Pinna, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics 7:367–394.

de Solan, T., J. P. Renoult, P. Geniez, P. David, and P.-A. Crochet. 2020. Looking for mimicry in a snake assemblage using deep learning. The American naturalist 196:74–86.

Deans, A. R., S. E. Lewis, E. Huala, S. S. Anzaldo, M. Ashburner, J. P. Balhoff, D. C. Blackburn, J. A. Blake, J. G. Burleigh, B. Chanet, L. D. Cooper, M. Courtot, S. Csösz, H. Cui, W. Dahdul, S. Das, T. A. Dececchi, A. Dettai, R. Diogo, R. E. Druzinsky, M. Dumontier, N. M. Franz, F. Friedrich, G. V. Gkoutos, M. Haendel, L. J. Harmon, T. F. Hayamizu, Y. He, H. M. Hines, N. Ibrahim, L. M. Jackson, P. Jaiswal, C. James-Zorn, S. Köhler, G. Lecointre, H. Lapp, C. J. Lawrence, N. Le Novère, J. G. Lundberg, J. Macklin, A. R. Mast, P. E. Midford, I. Mikó[3], C. J. Mungall, A. Oellrich, D. Osumi-Sutherland, H. Parkinson, M. J. Ramírez, S. Richter, P. N. Robinson, A. Ruttenberg, K. S. Schulz, E. Segerdell, K. C. Seltmann, M. J. Sharkey, A. D. Smith, B. Smith, C. D.

Specht, R. B. Squires, R. W. Thacker, A. Thessen, J. Fernandez-Triana, M. Vihinen, P. D. Vize, L. Vogt, C. E. Wall, R. L. Walls, M. Westerfeld, R. A. Wharton, C. S. Wirkner, J. B. Woolley, M. J. Yoder, A. M. Zorn, and P. Mabee. 2015. Finding our way through phenotypes. PLOS Biology 13:1–9.

Dececchi, T. A., J. P. Balhoff, H. Lapp, and P. M. Mabee. 2015. Toward synthesizing our knowledge of morphology: Using ontologies and machine reasoning to extract presence/absence evolutionary phenotypes across studies. Systematic Biology 64:936–952.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Pages 4171–4186 *in* J. Burstein, C. Doran, and T. Solorio, eds. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota.

Di Martino, E., B. Berning, D. P. Gordon, P. Kuklinski, L. H. Liow, M. H. Ramsfjell, H. L. Ribeiro, A. M. Smith, P. D. Taylor, K. L. Voje, A. Waeschenbach, and A. Porto. 2023. Deepbryo: A web app for ai-assisted morphometric characterization of cheilostome bryozoans. Limnology and Oceanography: Methods 21:542–551.

Dinnage, R., and M. Kleineberg. 2025. Generative AI extracts ecological meaning from the complex three dimensional shapes of bird bills. PLoS computational biology 21:e1012887.

Du, M., W. Wang, J. Tan, and J. Barido-Sottani. 2026. Deep learning for mass extinction detection on fossilized phylogenies: power, limitations and lessons for simulation-based birth–death inference. Royal Society Open Science 13:252096.

Elhage, N., T. Hume, C. Olsson, N. Schiefer, T. Henighan, S. Kravec,

Z. Hatfield-Dodds, R. Lasenby, D. Drain, C. Chen, R. Grosse, S. McCandlish, J. Kaplan, D. Amodei, M. Wattenberg, and C. Olah. 2022. Toy models of superposition.

Elhamod, M., M. Khurana, H. B. Manogaran, J. C. Uyeda, M. A. Balk, W. Dahdul, Y. Bakis, H. L. Bart, P. M. Mabee, H. Lapp, J. P. Balhoff, C. Charpentier, D. Carlyn, W.-L. Chao, C. V. Stewart, D. I. Rubenstein, T. Berger-Wolf, and A. Karpatne. 2023. Discovering novel biological traits from images using phylogeny-guided neural networks. Page 3966–3978 *in* Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. KDD '23. Association for Computing Machinery, New York, NY, USA.

Elman, J. L. 1990. Finding structure in time. Cognitive Science 14:179–211.

Estopinan, J., P. Bonnet, M. Servajean, F. Munoz, and A. Joly. 2024. Modelling species distributions with deep learning to predict plant extinction risk and assess climate change impacts.

Evans, A. R., T. I. Pollock, S. G. C. Cleuren, W. M. G. Parker, H. L. Richards, K. L. S. Garland, E. M. G. Fitzgerald, T. E. Wilson, D. P. Hocking, and J. W. Adams. 2021. A universal power law for modelling the growth and form of teeth, claws, horns, thorns, beaks, and shells. BMC biology 19:58.

Ezray, B. D., D. C. Wham, C. E. Hill, and H. M. Hines. 2019. Unsupervised machine learning reveals mimicry complexes in bumblebees occur along a perceptual continuum. Proceedings. Biological sciences 286:20191501.

Fan, L., C. Xu, E. A. Jarzembowski, and X. Cui. 2022. Quantifying plant mimesis in fossil insects using deep learning. Historical Biology 34:907–916.

Farahani, A., S. Voghoei, K. Rasheed, and H. R. Arabnia. 2021. A brief review of domain adaptation. Pages 877–894 *in* R. Stahlbock, G. M. Weiss, M. Abou-Nasr, C.-Y. Yang, H. R. Arabnia, and L. Deligiannidis, eds. Advances in Data Science and Information Engineering. Springer International Publishing, Cham.

Feldmann, M. J., J. L. Gage, S. D. Turner-Hissong, and J. R. Ubbens. 2021. Images carried before the fire: The power, promise, and responsibility of latent phenotyping in plants. The Plant Phenome Journal 4:e20023.

Felsenstein, J. 1985. Phylogenies and the comparative method. The American Naturalist 125:1–15.

Feng, Z., Z. Wang, S. I. Bueno, T. Frelek, A. Ramesh, J. Bai, L. Wang, Z. Huang, J. Gu, J. Yoo, T.-Y. Pan, A. Chowdhury, M. Ramirez, E. G. Campolongo, M. J. Thompson, C. G. Lawrence, S. Record, N. Rosser, A. Karpatne, D. Rubenstein, H. Lapp, C. V. Stewart, T. Berger-Wolf, Y. Su, and W.-L. Chao. 2025. Static segmentation by tracking: A frustratingly label-efficient approach to fine-grained segmentation.

FitzJohn, R. G. 2012. Diversitree: comparative phylogenetic analyses of diversification in r. Methods in Ecology and Evolution 3:1084–1092.

Fowler, D. R., H. Meinhardt, and P. Prusinkiewicz. 1992. Modeling seashells. SIGGRAPH Comput. Graph. 26:379–387.

Freudenstein, J. V. 2005. Characters, states, and homology. Systematic Biology 54:965–973.

Geirhos, R., J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. 2020. Shortcut learning in deep neural networks. Nature Machine Intelligence 2:665–673.

Gharaee, Z., S. C. Lowe, Z. Gong, P. M. Arias, N. Pellegrino, A. T. Wang, J. B. Haurum, I. Zarubiieva, L. Kari, D. Steinke, G. W. Taylor, P. Fieguth, and A. X. Chang. 2024. Bioscan-5m: A multimodal dataset for insect biodiversity. Pages 36285–36313 *in* A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, eds. Advances in Neural Information Processing Systems. Vol. 37. Curran Associates, Inc.

Ghorbani, A., A. Abid, and J. Zou. 2019. Interpretation of neural networks is fragile. *In* Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Godfrey-Smith, P. 2003. Theory and Reality: An Introduction to the Philosophy of Science. University of Chicago Press, Chicago.

Goswami, A., W. J. Binder, J. Meachen, and F. R. O'Keefe. 2015. The fossil record of phenotypic integration and modularity: A deep-time perspective on developmental and evolutionary dynamics. Proceedings of the National Academy of Sciences 112:4891–4896.

Gould, S. J. 1974. The origin and function of "bizarre" structures: Antler size and skull size in the "irish elk," megaloceros giganteus. Evolution 28:191–220.

Gould, S. J., and R. C. Lewontin. 1979. The spandrels of san marco and the panglossian paradigm: a critique of the adaptationist programme. Proceedings of the Royal Society of London. B. Biological Sciences 205:581–598.

Grabowski, M., J. Pienaar, K. L. Voje, S. Andersson, J. Fuentes-González, B. T. Kopperud, D. S. Moen, M. Tsuboi, J. Uyeda, and T. F. Hansen. 2023. A cautionary note on "a cautionary note on the use of ornstein uhlenbeck models in macroevolutionary studies". Systematic Biology 72:955–963.

Grafen, A. 1989. The phylogenetic regression. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 326:119–157.

Gu, J., S. Stevens, E. Campolongo, M. Thompson, N. Zhang, J. Wu, A. Kopanev, Z. Mai, A. White, J. Balhoff, W. Dahdul, D. Rubenstein, H. Lapp, T. Berger-Wolf, W.-L. H. Chao, and Y. Su. 2025. Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. Pages 102778–102811 *in* D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and

N. Chen, eds. Advances in Neural Information Processing Systems. Vol. 38, main conference. Curran Associates, Inc.

Gu, J., S. Stevens, E. G. Campolongo, M. J. Thompson, N. Zhang, J. Wu, A. Kopanev, Z. Mai, A. E. White, J. Balhoff, W. M. Dahdul, D. Rubenstein, H. Lapp, T. Berger-Wolf, W.-L. Chao, and Y. Su. 2026. TreeOfLife-200M (revision 94bbcob).

Guillerme, T. 2024. treats: A modular r package for simulating trees and traits. Methods in Ecology and Evolution 15:647–656.

Haller, B. C., and P. W. Messer. 2019. Slim 3: Forward genetic simulations beyond the wright–fisher model. Molecular Biology and Evolution 36:632–637.

Hand, D. J. 1996. Statistics and the theory of measurement. Journal of the Royal Statistical Society Series A: Statistics in Society 159:445–473.

———. 2004. Measurement Theory and Practice: The World Through Quantification. John Wiley & Sons, Chichester, UK.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51:1341–1351.

Hansen, T. F., J. Pienaar, and S. H. Orzack. 2008. A comparative method for studying adaptation to a randomly evolving environment. Evolution; international journal of organic evolution 62:1965–1977.

Hartmann, G., T. Blakesley, P. E. dePolo, and S. L. Brusatte. 2026. Identifying variation in dinosaur footprints and classifying problematic specimens via unbiased unsupervised machine learning. Proceedings of the National Academy of Sciences 123:e2527222122.

He, Y., J. M. Mulqueeney, E. C. Watt, A. Salili-James, N. S. Barber, M. Camaiti, E. S. E. Hunt, O. Kippax-Chui, A. Knapp, A. Lanzetti, G. Rangel-de Lázaro, J. K. McMinn, J. Minus, A. V. Mohan, L. E. Roberts, D. Adhami, E. Grisan,

Q. Gu, V. Herridge, S. T. S. Poon, T. West, and A. Goswami. 2024. Opportunities and challenges in applying ai to evolutionary morphology. Integrative Organismal Biology 6:obae036.

Hill, B. G., F. L. Koback, and P. L. Schilling. 2024. The risk of shortcutting in deep learning algorithms for medical imaging research. Scientific Reports 14:29224.

Hochreiter, S., and J. Schmidhuber. 1997. Long short-term memory. Neural Computation 9:1735–1780.

Hoffmann, J., S. Donoughe, K. Li, M. K. Salcedo, and C. H. Rycroft. 2018. A simple developmental model recapitulates complex insect wing venation patterns. Proceedings of the National Academy of Sciences 115:9905–9910.

Hofmann, M., S. Kiel, L. M. Kösters, J. Wäldchen, and P. Mäder. 2024. Inferring taxonomic affinities and genetic distances using morphological features extracted from specimen images: A case study with a bivalve data set. Systematic biology 73:920–940.

Hogg, D. W., and S. Villar. 2024. Is machine learning good or bad for the natural sciences?

Höhna, S., M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology 65:726–736.

Hornik, K., M. Stinchcombe, and H. White. 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2:359–366.

Houle, D. 2010. Numbering the hairs on our heads: The shared challenge and promise of phenomics. Proceedings of the National Academy of Sciences 107:1793–1799.

Houle, D., D. R. Govindaraju, and S. Omholt. 2010. Phenomics: the next challenge. Nature Reviews Genetics 11:855–866.

Houle, D., C. Pélabon, G. Wagner, and T. Hansen. 2011. Measurement and meaning in biology. The Quarterly Review of Biology 86:3–34.

Hoyal Cuthill, J. F., and S. Conway Morris. 2014. Fractal branching organizations of ediacaran rangeomorph fronds reveal a lost proterozoic body plan. Proceedings of the National Academy of Sciences of the United States of America 111:13122–13126.

Huben, R., H. Cunningham, L. Smith, A. Ewart, and L. Sharkey. 2024. Sparse autoencoders find highly interpretable features in language models. Pages 7827–7845 *in* B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, eds. International Conference on Learning Representations. Vol. 2024.

Huelsenbeck, J. P., B. Larget, and M. E. Alfaro. 2004. Bayesian phylogenetic model selection using reversible jump markov chain monte carlo. Molecular Biology and Evolution 21:1123–1133.

Hunt, R., and K. S. Pedersen. 2022. Rove-tree-11: The not-so-wild rover, a hierarchically structured image dataset for deep metric learning research. Pages 2967–2983 *in* Proceedings of the Asian Conference on Computer Vision (ACCV).

Hunt, R., J. L. Reyes-Hernández, J. J. Shaw, A. Solodovnikov, and K. S. Pedersen. 2025. Integrating deep learning derived morphological traits and molecular data for total-evidence phylogenetics: Lessons from digitized collections. Systematic Biology 74:453–468.

Huttegger, S. M., and P. Mitteroecker. 2011. Invariance and meaningfulness in phenotype spaces. Evolutionary Biology 38:335–351.

Jacobsen, J.-H., J. Behrmann, R. Zemel, and M. Bethge. 2019. Excessive invariance

causes adversarial vulnerability. *In* International Conference on Learning Representations.

Jones, M. 2026. Classifying patterns in comparative data using a neural network pipeline. Oral presentation at the Society for the Study of Evolution Annual Meeting.

Karpatne, A., G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar. 2017. Theory-guided data science: A new paradigm for scientific discovery from data. IEEE Trans. on Knowl. and Data Eng. 29:2318–2331.

Karpatne, A., X. Jia, and V. Kumar. 2024. Knowledge-guided machine learning: Current trends and future prospects.

———. 2026. Knowledge-guided machine learning: A paradigm shift in AI for science. Pages 39715–39717 *in* Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 40.

Khakurel, B., C. Grigsby, T. D. Tran, J. Zariwala, S. Höhna, and A. M. Wright. 2024. The fundamental role of character coding in bayesian morphological phylogenetics. Systematic biology 73:861–871.

Khurana, M., A. Daw, M. Maruf, J. C. Uyeda, W. Dahdul, C. Charpentier, Y. Bakış, H. L. Bart, P. M. Mabee, H. Lapp, J. P. Balhoff, W.-L. Chao, C. Stewart, T. Berger-Wolf, and A. Karpatne. 2025. Hierarchical conditioning of diffusion models using tree-of-life for studying species evolution. Pages 137–153 *in* A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, eds. Computer Vision – ECCV 2024. Springer Nature Switzerland, Cham.

Kingma, D. P., and M. Welling. 2019. An introduction to variational autoencoders. Foundations and Trends® in Machine Learning 12:307–392.

Kondo, S. 2009. How animals get their skin patterns: Fish pigment pattern as a

live turing wave. Pages 37–46 *in* S. Nakanishi, R. Kageyama, and D. Watanabe, eds. Systems Biology. Springer Japan, Tokyo.

Lambert, S., J. Voznica, and H. Morlon. 2023. Deep learning from phylogenies for diversification analyses. Systematic Biology 72:1262–1279.

Lamiable, A., T. Champetier, F. Leonardi, E. Cohen, P. Sommer, D. Hardy, N. Argy, A. Massougbodji, E. Del Nery, G. Cottrell, Y.-J. Kwon, and A. Genovesio. 2023. Revealing invisible cell phenotypes with conditional generative modeling. Nature Communications 14:6386.

Landis, M. J., and A. Thompson. 2025. phyddle: Software for exploring phylogenetic models with deep learning. Systematic biology 74:1007–1019.

Lau, K., and G. P. T. Choi. 2026. Phylosdf: Phylogenetically-conditioned neural generation of 3d skull morphology via residual flow matching.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. Nature 521:436–444.

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86:2278–2324.

Lindenmayer, A. 1968. Mathematical models for cellular interactions in development i. filaments with one-sided inputs. Journal of Theoretical Biology 18:280–299.

Linder, C. R., R. Suri, K. Liu, and T. Warnow. 2010. Benchmark datasets and software for developing and testing methods for large-scale multiple sequence alignment and phylogenetic inference. PLoS currents 2:RRN1195.

Lindstedt, S. L., and H. Hoppeler. 2023. Allometry: revealing evolution's engineering principles. The Journal of experimental biology 226.

Liow, L. H., J. Uyeda, and G. Hunt. 2023. Cross-disciplinary information for understanding macroevolution. Trends in Ecology & Evolution 38:250–260.

Liu, Z., K. Yu, C. He, X. Cai, X. Ye, H. Wang, G. Ye, and J. Bu. 2026. Morphology-aware multimodal representation learning for insect phylogenetic reconstruction.

Locatello, F., S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. Pages 4114–4124 *in* K. Chaudhuri and R. Salakhutdinov, eds. Proceedings of the 36th International Conference on Machine Learning. Vol. 97 of *Proceedings of Machine Learning Research*. PMLR.

Lucny, A. 2020. On lindenmayer systems and autoencoders. Pages 217–222 *in* Proceedings of the 20th Conference Information Technologies – Applications and Theory (ITAT 2020). Vol. 2718 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Lundberg, S. M., and S.-I. Lee. 2017. A unified approach to interpreting model predictions. Page 4768–4777 *in* Proceedings of the 31st International Conference on Neural Information Processing Systems. NIPS'17. Curran Associates Inc., Red Hook, NY, USA.

Lürig, M. D., L. T. Shirai, L. L. Mota, K. Willmott, A. V. L. Freitas, and A. Porto. 2025. Aposematic color patterns are the dominant axis of phenotypic diversification in nymphalid butterflies. bioRxiv .

Lürig, M. D. 2022. phenopype: A phenotyping pipeline for python. Methods in Ecology and Evolution 13:569–576.

Lürig, M. D., E. Di Martino, and A. Porto. 2024. Bioencoder: A metric learning toolkit for comparative organismal biology. Ecology Letters 27:e14495. E14495 ELE-00472-2024.R1.

Lürig, M. D., S. Donoughe, E. I. Svensson, A. Porto, and M. Tsuboi. 2021. Computer vision, machine learning, and the promise of phenomics in ecology and evolutionary biology. Frontiers in Ecology and Evolution Volume 9 - 2021.

Machado, F. A., A. Hubbe, D. Melo, A. Porto, and G. Marroig. 2019. Measuring the magnitude of morphological integration: The effect of differences in morphometric representations and the inclusion of size. Evolution; international journal of organic evolution 73:2518–2528.

Machado, F. A., C. S. Mongle, G. Slater, A. Penna, A. Wisniewski, A. Soffin, V. Dutra, and J. C. Uyeda. 2023. Rules of teeth development align microevolution with macroevolution in extant and extinct primates. Nature ecology & evolution 7:1729–1739.

Manogaran, H. B., M. Maruf, A. Daw, K. S. Mehrab, C. Charpentier, J. Uyeda, W. Dahdul, M. Thompson, E. Campolongo, K. Provost, W.-L. Chao, T. Berger-Wolf, P. Mabee, H. Lapp, and A. Karpatne. 2025. What do you see in common? learning hierarchical prototypes over tree-of-life to discover evolutionary traits. Pages 45722–45758 *in* Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, eds. International Conference on Learning Representations. Vol. 2025.

Marcinkevičs, R., and J. E. Vogt. 2023. Interpretable and explainable machine learning: A methods-centric overview with concrete examples. WIREs Data Mining and Knowledge Discovery 13:e1493.

Martin, B. S., and M. G. Weber. 2026. Stochastic character mapping of continuous traits on phylogenies. Systematic Biology page syag031.

McElreath, R. 2020. Statistical Rethinking: A Bayesian Course with Examples in R and Stan. 2nd ed. CRC Press, Boca Raton, FL.

Mehrab, K. S., M. Maruf, A. Daw, A. Neog, H. B. Manogaran, M. Khurana, Z. Feng, B. Altintas, Y. Bakis, E. G. Campolongo, M. J. Thompson, X. Wang, H. Lapp, T. Berger-Wolf, P. Mabee, H. Bart, W.-L. Chao, W. M. Dahdul, and A. Karpatne. 2025. Fish-vista: A multi-purpose dataset for understanding & identification of traits from images. Pages 24275–24285 *in* 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Meinhardt, H., and M. Klingler. 1987. A model for pattern formation on the shells of molluscs. Journal of Theoretical Biology 126:63–89.

Mencken, H. L. 1921. Prejudices: Second Series. Jonathan Cape, London.

Merchant, A. T., D. Guo, B. Viggiano, L. Brennan-Almaraz, E. Hur, T. Mai, P. Yin, S. H. King, E. A. Ashley, and B. L. Hie. 2026. A high-level programming language for generative biology with proto. bioRxiv .

Mitteroecker, P., and S. M. Huttegger. 2009. The concept of morphospaces in evolutionary and developmental biology: Mathematics and metaphors. Biological Theory 4:54–67.

Mitteroecker, P., and K. Schaefer. 2022. Thirty years of geometric morphometrics: Achievements, challenges, and the ongoing quest for biological meaningfulness. American journal of biological anthropology 178 Suppl 74:181–210.

Mo, Y. K., M. W. Hahn, and M. L. Smith. 2024. Applications of machine learning in phylogenetics. Molecular Phylogenetics and Evolution 196:108066.

Mo, Z., and A. Siepel. 2023. Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. PLOS Genetics 19:1–22.

Montanaro, G., J. P. Balhoff, J. C. Girón, M. Söderholm, and S. Tarasov. 2024. Computable species descriptions and nanopublications: applying ontology-based technologies to dung beetles (coleoptera, scarabaeinae). Biodiversity data journal 12:e121562.

Moreno-Torres, J. G., T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera. 2012. A unifying view on dataset shift in classification. Pattern Recognition 45:521–530.

Moreta, L. S., J. Hein, T. Hamelryck, O. Rønning, A. S. Al-Sibahi, and D. Theobald. 2022. Ancestral protein sequence reconstruction using a tree-structured ornstein-uhlenbeck variational autoencoder. ... International Conference on Learning Representations 2022:9556–9577.

Morlon, H. 2014. Phylogenetic approaches for studying diversification. Ecology Letters 17:508–525.

Morlon, H., J. Andréoletti, J. Barido-Sottani, S. Lambert, B. Perez-Lamarque, I. Quintero, V. Senderov, and P. Veron. 2024. Phylogenetic insights into diversification. Annual Review of Ecology, Evolution, and Systematics 55:1–21.

Mosqueira-Rey, E., E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal. 2023. Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review 56:3005–3054.

Mulvey, L. P. A., M. R. May, J. M. Brown, S. Höhna, A. M. Wright, and R. C. M. Warnock. 2025. Assessing the adequacy of morphological models using posterior predictive simulations. Systematic Biology 74:34–52.

Murphy, K. P. 2022. Probabilistic Machine Learning: An introduction. MIT Press.

Nagel, A. A., and M. J. Landis. 2026. Ancestral state reconstruction with discrete characters using deep learning. bioRxiv : the preprint server for biology .

Nesterenko, L., L. Blassel, P. Veber, B. Boussau, and L. Jacob. 2025. Phyloformer: Fast, accurate, and versatile phylogenetic reconstruction with deep neural networks. Molecular Biology and Evolution 42:msaf051.

Okabe, T., and J. Yoshimura. 2017. Optimal designs of mollusk shells from bivalves to snails. Scientific Reports 7:42445.

Pagel, M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. Proceedings of the Royal Society B: Biological Sciences 255:37–45.

Pagel, M., and A. Meade. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. The American naturalist 167:808–825.

Pahuja, V., S. Stevens, A. East, S. Record, and Y. Su. 2026. Automatic image-level morphological trait annotation for organismal images. *In* The Fourteenth International Conference on Learning Representations.

Park, J. J., P. Florence, J. Straub, R. Newcombe, and S. Lovegrove. 2019. Deepsdf: Learning continuous signed distance functions for shape representation. *In* Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Perez, M. F., and O. Gascuel. 2026. Phylocnn: Improving tree representation and neural network architecture for deep learning from trees in phylodynamics and diversification studies. Systematic Biology 75:776–795.

Peters, B., S. E. Brenner, E. Wang, D. Slonim, and M. G. Kann. 2018. Putting benchmarks in their rightful place: The heart of computational biology. PLOS Computational Biology 14:1–3.

Pierucci, F., M. B. Syrnikov, M. Prandi, M. Galisai, F. Giarrusso, and P. Bisconti. 2026. Xeno-interpretability: Investigating the alien minds of llms.

Plum, F., R. Bulla, H. K. Beck, N. Imirzian, and D. Labonte. 2023. replicant: a pipeline for generating annotated images of animals in complex environments using unreal engine. Nature Communications 14:7195.

Pogue, M. G., and M. F. Mickevich. 1990. Character definitions and character state delineation: The bete noire of phylogenetic inference. Cladistics 6:319–361.

Porto, A. 2026. Morphoweave: Atlas-based 3d landmark transfer and statistical shape modeling.

Porto, A., and K. L. Voje. 2020. Ml-morph: A fast, accurate and general approach for automated detection and landmarking of biological structures in images. Methods in Ecology and Evolution 11:500–512.

Posada, D., and T. R. Buckley. 2004. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. Systematic Biology 53:793–808.

Prabhushankar, M., G. Kwon, D. Temel, and G. AlRegib. 2020. Contrastive explanations in neural networks. Pages 3289–3293 *in* 2020 IEEE International Conference on Image Processing (ICIP).

Prusinkiewicz, P. 2025. Pillars of theoretical biology: 'mathematical models for cellular interaction in development, i and ii'. Journal of Theoretical Biology 609:112142.

Prusinkiewicz, P., and A. Lindenmayer. 1990. The Algorithmic Beauty of Plants. The Virtual Laboratory. Springer-Verlag, New York, NY.

Quiroga, R. Q., G. Kreiman, C. Koch, and Fried. 2008. Sparse but not 'grandmother-cell' coding in the medial temporal lobe. Trends in cognitive sciences 12:87–91.

Rabosky, D. L., and E. E. Goldberg. 2015. Model inadequacy and mistaken inferences of trait-dependent speciation. Systematic Biology 64:340–355.

Rao, R. M., J. Liu, R. Verkuil, J. Meier, J. Canny, P. Abbeel, T. Sercu, and A. Rives. 2021. Msa transformer. Pages 8844–8856 *in* M. Meila and T. Zhang, eds. Proceedings of the 38th International Conference on Machine Learning. Vol. 139 of *Proceedings of Machine Learning Research*. PMLR.

Raup, D. M. 1966. Geometric analysis of shell coiling: General problems. Journal of Paleontology 40:1178–1190.

Raup, D. M., and A. Michelson. 1965. Theoretical morphology of the coiled shell. Science 147:1294–1295.

Ravi, N., V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollar, and C. Feichtenhofer. 2025. Sam 2: Segment anything in images and videos. Pages 28085–28128 *in* Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, eds. International Conference on Learning Representations. Vol. 2025.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. Page 1135–1144 *in* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16. Association for Computing Machinery, New York, NY, USA.

Ross, S. 2014. - brownian motion and stationary processes. Pages 607–644 *in* Introduction to Probability Models (Eleventh Edition), eleventh edition ed. Academic Press, Boston.

Rudin, C. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1:206–215.

Russakovsky, O., J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. 2015. Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115:211–252.

Schluter, D. 1996. Adaptive radiation along genetic lines of least resistance. Evolution 50:1766–1774.

Schnepf, A., D. Leitner, P. F. Schweiger, P. Scholl, and J. Jansa. 2016. L-system model for the growth of arbuscular mycorrhizal fungi, both within and outside of their host roots. Journal of the Royal Society, Interface 13.

Selvaraju, R. R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2020. Grad-cam: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision 128:336–359.

Shapley, L. S. 1988. A value for n-person games. Pages 31–40 *in* A. E. Roth, ed. The Shapley Value: Essays in Honor of Lloyd S. Shapley. Cambridge University Press, Cambridge.

Silvestro, D., T. Latrille, and N. Salamin. 2024. Toward a semi-supervised learning approach to phylogenetic estimation. Systematic Biology 73:789–806.

Simpson, G. G. 1984. Tempo and Mode in Evolution. Columbia Classics in Evolution. Columbia University Press, New York. Reprint of the 1944 edition, with a new introduction.

Simões, T. R., M. W. Caldwell, A. Palci, and R. L. Nydam. 2018. Giant taxon-character matrices ii: a response to laing et al. (2017). Cladistics 34:702–707.

Smith, M. L., and M. W. Hahn. 2023. Phylogenetic inference using generative adversarial networks. Bioinformatics 39:btad543.

Soares, V. M. A., and R. Wassermann. 2025. Specific domain ontology construction using large language models. *In* Proceedings of the Second International Workshop on Next-Generation Language Models for Knowledge Representation and Reasoning (NeLaMKRR 2025).

Soewongsono, A. C., and M. J. Landis. 2026. Phylogenetic estimation of diversity-dependent biogeographic rates using deep learning. bioRxiv : the preprint server for biology .

Stack, J. 2025. An ontological morphological phylogenetic framework for living and extinct ray-finned fishes (actinopterygii). The Anatomical Record n/a.

Stevens, S., W.-L. Chao, T. Berger-Wolf, and Y. Su. 2025. Interpretable and testable vision features via sparse autoencoders.

Stevens, S., J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. Berger-Wolf, W.-L. Chao, and Y. Su. 2024. Bioclip: A vision foundation model for the tree of life. Pages 19412–19424 *in* 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Stewart, T. A., W. L. Smith, and M. I. Coates. 2014. The origins of adipose fins: an analysis of homoplasy and the serial homology of vertebrate appendages. Proceedings. Biological sciences 281:20133120.

Sullivan, J., and P. Joyce. 2005. Model selection in phylogenetics. Annual Review of Ecology, Evolution, and Systematics 36:445–466.

Suvorov, A., J. Hochuli, and D. R. Schrider. 2020. Accurate inference of tree topologies from multiple sequence alignments using deep learning. Systematic biology 69:221–233.

Suvorov, A., and D. R. Schrider. 2024. Reliable estimation of tree branch lengths using deep neural networks. PLOS Computational Biology 20:1–25.

Tarasov, S., and J. Uyeda. 2026. Unidentifiability and false-positive inference in state-dependent diversification models. Nature Communications 17:9076.

Templeton, A., T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread .

Tessler, M., and D. P. Little. 2026. On herbarium specimen images and artificial intelligence. New Phytologist 251:666–676.

Thompson, A., B. J. Liebeskind, E. J. Scully, and M. J. Landis. 2024. Deep learning and likelihood approaches for viral phylogeography converge on the same answers whether the inference model is right or wrong. Systematic Biology 73:183–206.

Tribble, C. M., W. A. Freyman, M. J. Landis, J. Y. Lim, J. Barido-Sottani, B. T. Kopperud, S. Höhna, and M. R. May. 2022. Revgadgets: An r package for visualizing bayesian phylogenetic analyses from revbayes. Methods in Ecology and Evolution 13:314–323.

Tsutsumi, M., N. Saito, D. Koyabu, and C. Furusawa. 2023. A deep learning approach for morphological feature extraction based on variational auto-encoder: an application to mandible shape. npj Systems Biology and Applications 9:30.

Ubbens, J., M. Cieslak, P. Prusinkiewicz, I. Parkin, J. Ebersbach, and I. Stavness. 2020. Latent space phenotyping: Automatic image-based phenotyping for treatment studies. Plant phenomics (Washington, D.C.) 2020:5801869.

Ubbens, J., M. Cieslak, P. Prusinkiewicz, and I. Stavness. 2018. The use of plant models in deep learning: an application to leaf counting in rosette plants. Plant methods 14:6.

Uyeda, J. C., D. S. Caetano, and M. W. Pennell. 2015. Comparative analysis of principal components can be misleading. Systematic Biology 64:677–689.

Uyeda, J. C., and L. J. Harmon. 2014. A novel bayesian method for inferring and interpreting the dynamics of adaptive landscapes from phylogenetic comparative data. Systematic Biology 63:902–918.

Van Dam, A. R., and L. Štarhová Serbina. 2026. Descriptron: Artificial intelligence for automating taxonomic species descriptions with a user-friendly software package. Systematic Entomology 51:e70005.

Xie, S., L. Valente, and R. S. Etienne. 2026. Inferring state-dependent diversification rates using approximate bayesian computation. Methods in Ecology and Evolution 17:805–820.

Yu, J., A. Ignatiev, and P. J. Stuckey. 2023. On formal feature attribution and its approximation.

Zahavy, T. 2026. Position: LLMs can't jump. *In* Forty-third International Conference on Machine Learning Position Paper Track.

Zhang, A., Z. C. Lipton, M. Li, and A. J. Smola. 2023. Dive into Deep Learning. Cambridge University Press. https://D2L.ai.

Zhang, Z., J. Gu, A. Chowdhury, Z. Mai, D. Carlyn, T. Berger-Wolf, Y. Su, and W.-L. Chao. 2025. Finer-cam: Spotting the difference reveals finer details for visual explanation. Pages 9611–9620 *in* 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Zhang, Z., X. Ma, A. Chowdhury, E. G. Campolongo, M. J. Thompson, N. Zhang, S. Stevens, H. Lapp, T. Berger-Wolf, Y. Su, W.-L. Chao, and J. Gu. 2026. BioCAP: Exploiting synthetic captions beyond labels in biological foundation models. *In* The Fourteenth International Conference on Learning Representations.

Zhao, Z., Y. Lu, Y. Tong, X. Chen, and M. Bai. 2023. Penet: A phenotype encoding network for automatic extraction and representation of morphological discriminative features. Methods in Ecology and Evolution 14:3035–3046.

Ziegler, C., J. Martin, C. Sinner, and F. Morcos. 2023. Latent generative landscapes as maps of functional diversity in protein sequence space. Nature Communications 14:2222.

Zong, Y., T. Yu, R. Chavhan, B. Zhao, and T. Hospedales. 2024. Fool your (Vision and) language model with embarrassingly simple permutations. Pages 62892–62913 *in* R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver,

J. Scarlett, and F. Berkenkamp, eds. Proceedings of the 41st International Conference on Machine Learning. Vol. 235 of *Proceedings of Machine Learning Research*. PMLR.